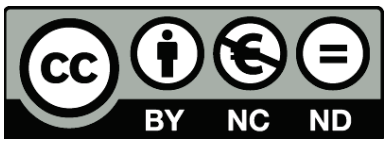
UNIVERSITAT DE BARCELONA

Towards a large-scale human inflammation atlas at single-cell resolution

Laura Jiménez Gracia



Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative*** **Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License***.

Towards a large-scale human inflammation atlas at single-cell resolution

Laura Jiménez Gracia

DOCTORAL THESIS
Barcelona, 2024

Doctoral Programme in Biomedicine - Bioinformatics

Faculty of Biology, Universitat de Barcelona

Centre Nacional d'Anàlisis Genòmica (CNAG) and Centre de Regulació Genòmica (CRG)

Towards a large-scale human inflammation atlas at single-cell resolution

Thesis submitted by
Laura Jiménez Gracia

to qualify for the
Doctorate Degree by the Universitat de Barcelona

Doctorate Candidate
Laura Jiménez Gracia

Dr. Ivo G. Gut	**Dr. Holger Heyn**	**Dr. Josep F. Abril Ferrando**
Co-Director	Director	Tutor

Barcelona, Octubre 2024

The most important things in life
are almost always impossible to predict

- Joe Meno

Acknowledgments

Potser el final d'aquesta etapa havia de ser així, per guardar mimetisme amb el seu inici… Al març de 2020, enmig d'un mar d'incerteses i després d'enviar un correu al que es convertiria en el meu director de tesi, vaig aplicar a la FPU sense saber en quin projecte m'estava endinsant, just a les portes del que seria una pandemia mundial. Ara, al setembre de 2024, puc afirmar que va ser una gran decisió, tot i que fer realitat aquesta tesi ha estat una cursa a contrarellotge que mai hauria imaginat culminant amb un final tan satisfactori. I ara, amb l'escenografia a punt, s'obre el teló.

I want to start by thanking **Holger**, the director of my thesis, who, out of nowhere, gave me the opportunity to become part of the Single Cell Dream Team. I admire the passion with which you conduct research, the energy you transmit, and your determination to take opportunities beyond the limits of a lab. But, I value your human side more - the one that takes care of your personal well-being at work and outside of it. Thank you for the trust you've placed in me since the very first day I walked through that door, for believing in me blindly (sometimes even too much), and for opening doors for me now that the path is widening. Thanks for making sure I was satisfied along the way and helping me achieve my dreams. And now that this journey is coming to an end, promise me that your door will always remain open, with a cup of tea ready whenever it's needed.

Next, I want to thank **Ivo**, the co-director of my thesis, who, without even knowing me, accepted me into the CNAG family. Thank you for all your advice on how to navigate a scientific career and life itself. També vull agrair al meu tutor de la UB, en **Josep Abril**, amb qui he compartit totes les etapes universitàries i que va ser la primera espurna del canvi de la poiata al teclat. Thanks to **Anna** and **Damjana** from the CRG Training and Academic Office (TAO) for guiding me from the other side of Barcelona. Lastly, I want to thank the members of my Thesis Advisory Committee (TAC), **Thomas Graf** and **Azucena Salas**, for their feedback year after year to improve this thesis.

I once heard at a single-cell conference, "Sometimes the sum of the parts is greater than the whole", and that couldn't be truer for our team. Gracias **Paula**, por ser compañera en esta nuestra llorería, pero amiga fuera de ella, por los cafés, las cenas, los conciertos, las cerves y nuestros kalimotxos. Has sido un pilar indispensable para esta tesis en lo computacional pero sobre todo en lo personal, dándome apoyo cada día, y acompañándome cuando más lo he necesitado. Admiro tu bondad, compañerismo y predisposición para todo y siempre ¡Qué placer el mío! Esta etapa se acaba, pero nosotras no, no hasta llamarte Dra. A ti **Juan**, gracias por ser el hermano mayor que se asegura de que todo salga bien y por compartir tu conocimiento inmunológico, que no es poco. Gracias por ser hombro durante esta tesis, por esos cafés que siempre acaban en cerveza, y por esas cervezas que acaban en presentaciones y nuevos proyectos! **Davide**, I admire your determination and professionalism at work, and I couldn't be happier to see how much you've blossomed during this time. Thank you for all the good moments in the office, and outside; those beers, the hikes, and the museums (with dinosaurs). Thank you for looking out for me, grazie mille per tutto! Qué decirte a ti, **Sergio**, que consigues sacar lo mejor y peor de mi a la vez. Gracias por empezar esta aventura a mi lado, por los almuerzos y paseos, por cuidarme y haber sido hogar cuando más lo necesité. ¡Será que esos músculos también saben abrazar! A vosotros dos, Davide y Sergio (y también Francesco, Juan y Holger) gracias por no inflamaros a mi lado este año, por compartir y hacer posible esta locura; creo que nunca podré olvidar esas jornadas eternas bebiendo café a litros.

Thanks to everyone from the 10th fioor, now the 6th! Gràcies **Inés** pels mesos compartits a la 2a, per ser una pijipi en tota regla, i pels moments de bucle eterns compensats amb riures i bones imitacions (me debes una!). **Irepan**, gracias por todo el feedback y apoyo cuando estuvimos codo a codo (perdón por exiliarte), por ser el DJ de las ffestas, por quemar las pistas de Valencia a mi lado, y por enchilarme (por chula) en esa barbacoa. Thank you **Ema**, for going from calling me "Madam Gracia" in our ffirst email to closing discos during your ffirst week in Barcelona, and for separating the meticulous supervision you received in the ofl ce from me as a colleague and friend. You will always be my ffirst TFM, but never my last shoot!

Thank you to those who are in the ffield, receiving and processing samples. **Giulia**, gracias por ser guardián en CC de tantos correos, y por tu forma tan dulce de ayudar. **Sara**, gracias (o no) por ser el principal suministro de librerías (y nightmares) durante meses, y por compartir todas tus vivencias (me debes una de twerking y pole dance). **Patri**, gracias por ser esos "buenos días" al llegar, y los buenos momentos de marujas compartidos. **Gin**, gracias por venir a vernos, asegurarte que estamos bien, y cuidarnos a base de dulces italianos. I, a tu **Elena**, gràcies per aconseguir que el temps i la distància no ens afecti, i que tot continuï igual desde que vam conèixe'ns al LAB4, quan tot just començava la carrera i era un bebé inexpert; és un plaer poder tornar a treballar amb tu, i em pregunto si ens hi trobarem un altre cop més endavant. Gràcies **Anna**, per ser pura energia, vitalitat i bon rotllo; gràcies per ser guia a Berlín però tornar a Barcelona i quedar-t'hi. I would like to extend my thanks to the entire group: **Helena**, **Sonal**, **Gerard**, **Elena**, **Irene, Sam,** and **Lorena**, with whom I have shared some coffees, beers, and meals to break the routine of the days.

Gràcies als que ja no estan al grup, però sempre seran familia. **Marc**, gràcies per ser bondat i professionalitat, i per compartir tantes converses emocionals (sí, laborals i personals), "bogeria" a base de clares i rondes de chupitos traicioneros. Et mereixes el millor, i espero seguir al teu costat per a veure-ho i celebrar-ho. **Ramon**, gràcies pel coneixement injectat en vena que em vas donar a l'entrar al grup, ets sabiduría i innocència. Gràcies sobretot per les abraçades càlides i escoltar-me quan més ho necessitava… a l'altre banda de l'Atlàntic estarem esperant que tornis aviat a petar-ho fort. Thanks to all the past members of the group: **Mimma**, thank you for sharing the FixNCut journey. **Will**, thanks for being a box of surprises and a Doraemon's pocket; the one to turn to if I get locked out of my house or to share an afternoon making limoncello. **Ireen**, thank you for the ¡Lekker! pizza dinners on your terrace. **Pia**, thank you for bringing (twice) your contagious smile and good vibes. **Mo**, thank you for all the sweets that brighten everyone's day.

I want to extend my thanks to all my colleagues at **CNAG**, with whom I have shared the wait at the microwaves, the line for coffee, a lunch at TupperPark, an afternoon in the bowling, or the Christmas dinner, and to my **CNAG-CRG PhD mates** with thom I started online, but then I shared beer sessions, symposiums, night clubs, and the PhD retreats; thank you **Ivo**, for our shared trips around Europe, i **Júlia,** gràcies per ser tan dolça. I would also like to thank all the researchers with whom I have shared projects, reports, and deadlines over the years, especially the **Batlle Lab** at IRB and the **IBD Lab** at IDIBAPS. Thanks **Lu**, for involving me in FixNCut. Also, thanks to the partners of the major projects, **DOCTIS**, **3TR** and **EF-CLIF**, for the collaborative environment and all the learning that has offered me, the discussions during the meetings and the networking afterwards. També m'agradaria agraïr al **PCB** l'oportunitat de participar en diferents activitats de divulgació científfica i poder despertar inquietuds cap al món bio i la recerca a l'alumnat més jove.

I, en lo personal, vull donar les gràcies als de sempre. Començant per tu, **Ire**; la persona amb la que porto més camí recorregut, viatges compartits, summers, dies de platja i nits eternes; gràcies pel teu suport, ànims, carinyo, i amor durant tots aquests anys, ets familia. **Laura**, gràcies per ser la "popu" amb personalitat del cole, compartir casals, disfresses, per separar-nos i tornar a compartir juntes. Gràcies per atrevir-te a viure fa uns anys i fer cas a la boja de la teva amiga… et miro ara, i no saps lo contenta que em fa veure't així de feliç. **Marta** (o mejor dicho, Pitufa!), gracias por hacer amenas clases eternas, por ser compi de trabajos y de paseos sin rumbo, pero ser rumba constante. Gracias por tantas tardes de verano compartidas, las llamadas eternas, escapadas y locuras. Pero, sin duda, gracias por haber sido y ser mi confidente, por compartir vida, por entenderme y dejarme hacerlo a mi contigo, no me faltes nunca. Gracias a las dos por hacer que existan los botánicos; **Diego** y **Carlos**, gracias por ser amigas cuando las mías no se dignan. **Ainoa**, gracias por ser una cabezona de remate y una luchadora innata, por todo lo vivido (y robado), por la distancia que nos separó, y volver más fuertes que nunca. Gracias por ofrecerme las llaves de tu casa de manera tan altruista y sincera, fue un regalo que costará poder compensar. **Arnau**, gràcies per les tardes per SantAndreu, per les converses eternes de vida i filosofia, per ser el desastre personificat que més estimo, i aconseguir ser-hi sempre sense estar-hi. **Marta** y **Alejandro**, a vosotros, gracias por las aventuras de los 4 fantásticos, y por permanecer.

De l'etapa a la universitat vull agrair als **Biomed**, amb qui he compartit 4 anys d'apunts, examens, no-biblio, i hores i hores de pràctiques al laboratori. Gràcies per compensar això amb eterns sopars, aniversaris, sorpreses, festes (i LA biofesta) i, si-us-plau, que Montgat i Vila mai ens faltin. Gracias **Alex** por tu manera de cuidar y querer, por tu deseo de libertad eterna y por perseguir tus sueños hasta la saciedad. Gracias por enseñarme a volar sin miedo a caer, por infundir locura y darme paz. Gracias por encargarte de que te sienta cerca estando lejos (pero vuelve, te echo de menos!). Gràcies **Elena** per salvar-me Endo i BioDes, per compartir poiata i tots els cafés de catchup; ets una treballadora innata i Oslo té molta sort de tenir-te allà. **Claudia**, gràcies per ser tan teva però deixar-me conèixer el gran cor que tens, i per ser l'espurna que ens manté units. **Anna B**, ets la gent de la que t'envoltes, i tu tens un cercle preciós, gràcies per compartir-lo sempre amb els braços oberts. **Marc**, gràcies per la teva serenitat i calma, i per fer aportacions essencials en els moments adequats. **Diego**, gràcies per voler el millor i estar sempre predisposat a cuidar i fer feliços als altres. **Anna M**, gràcies per la teva essència i transparència. **Estela**, gràcies per haver estat un referent acadèmic i professional, i per tota l'ajuda brindada aquests anys, també en lo personal; és curiós com, malgrat la intermitència del nostre contacte, res canvia quan estem juntes. I, d'aquesta etapa també vull agrair a la **Gemma Marfany**, qui em va transmetre la seva passió per la ciència, la recerca i la docència. Gràcies per l'exigència i rigorositat en la feina durant el TFG al LAB4 i, sobretot, per ser referent científic i consellera en la meva carrera i cada pas que he fet.

Dels mesos a Leiden, ho agraeixo tot; la ciutat per la seva màgia, la gent per la seva acollida, i els companys del lab, per la confiança i l'aprenentatge. Gracias **Carmen**, por ser y ofrecerme hogar antes incluso de llegar a la ciudad, por compartir piso, habitación y cama en menos de un mes, por tus desayunos al despertar y nuestras noches por ahí. Thank you **Ale**, for being family all those months, for joining me on the adventure like a sister, and for taking care of me like a mum. Thank you for everything we shared, and for always coming back. Also, many thanks to **Alex**, **Nikolas, Stuart, Thomas, Wilson, Ilaria** and '**the Italians'!**

Del màster sempre recordaré les tardes eternes a la biblioteca, els treballs de Perl i les pràctiques d'estructural que tant ens van unir entre nosaltres. Gràcies **Olfat** per confiar, per no llençar la tovallola, per acompanyar-me, i quedar-te a prop malgrat les nostres diferències. Gracias **Nerea** por regalarnos otra oportunidad, por ser ternura y hacer de la distancia la excusa perfecta para vernos siempre a tu vuelta. Gracias a mis mayordomas; **Alda,** por tu determinación en lo profesional y en lo personal, y **Bea**, por ser bondad en estado puro, y una currante innata. Ah, y a ambas por nuestra inesperada aventura con **ColabStock!** Gràcies als altres **Bioinfo** amb els que retrobar-nos a conferències sempre ens farà recordar els vells temps i compartir experiències d'aquí i d'allà.

También quiero agradecer mi etapa por el **BSC**, dónde empecé a programar, y tanto aprendí usando el MareNostrum (a veces, colapsando también). Gracias **Fransua**, por la oportunidad y la confianza, y por ser el caos que siempre estaba ready para ayudar. Y a ti **Jon**, por ser el alma social, por los ánimos y las cerves de estos años, y las que nos esperan como buenos vecinos!

Bea, gràcies per fer-me florir de nou; per desendreçar-me completament però ajudar-me a (re)construir-me en la persona i amb els valors que m'identifico per a sentir-me complerta i lliure. **Silvia**, gracias por guiarme en las sesiones de yoga y contribuir en la alineación de mis chakras.

Gràcies a tothom que ha aparegut (i m'ha acompanyat) durant aquesta etapa. A mis liadas, ¡brindo por las casualidades de la vida, y la vida que me dáis! **Miren**, gracias por aparecer en verano pero ser primavera; admiro la fidelidad de tus ideales, tu bondad y tu astucia, tus ganas de vivir y desgastar suela. Gracias por tu confianza, por hacerme partícipe en tu círculo y unirte al mío sin pensarlo, por cuidarme y decirme aquello que no quiero oír pero tanto necesito escuchar. Gracias por todos los vermús al sol, las tardes a birras, y noches a roncola; ah, y por descubrirme Plataforma (que ya nos echa de menos!). A ti **Olga**, nunca imaginé que un trayecto de BusNit me regalara una persona así. Gracias por ser compi del 4latas, Quilombo y Mojito, por volar a Marruecos con 5 jerséis o bajar a Málaga para cerrar chiringuitos. Gracias por hacer que esperar siempre valga la pena, por ser casa y apuntarte siempre a un último baile. **Edu**, a tu, gràcies per la teva manera tan especial de ser i cuidar-me, i per ser el culpable de la meva addicció a l'adrenalina fent les corbes del Tibidabo. **Sergi**, gràcies per ser un racó de calma, per ensenyar-me que cal molt poc per ser feliços, i que el més important és gaudir amb els nostres. **Mariona**, gràcies per trobar la manera de connectar juntes, sigui amb una cervesa en mà o fotent-li al sac. Tens una manera molt bonica de somriure a la vida, no la perdis. **Ferran**, ets el pupas amb qui més vola el temps, i això és massa perillós. **Anna**, gràcies per aparèixer a Mykonos i tornar a Cantàbria, per les hores de cotxe i sol, per confiar i compartir, per lo viscut i pel que surfejarem d'ara endavant. **David**, gràcies per tornar i posar-li melodia a moltes de les hores invertides en aquesta tesi. **Elena**, gracias por esos ratitos improvisados, de risas y lágrimas, entre tus guardias. **Sergio**, gràcies per acollir-me amb carinyo i integrar-me al món farma com una més. Per vosaltres, faig un brindis, per seguir fent camí i gaudir de tot el que està per venir!

Gràcies també a tots i a totes els que van començar aquest camí de la meva mà, però l'han deixat abans d'arribar al final. Gràcies pels passos compartits, pels moments viscuts i gaudits, pels aprenentatges plegats, i pels bons records que heu deixat en mi.

Gràcies a la meva família.

A mis abuelos, **Antonio, Pepita, Rosendo** y **María**, gracias porqué sin saber bien qué es esto del Doctorado ni a qué me dedico, me apoyáis incondicionalmente y celebráis como vuestro cada paso que doy y cada éxito que consigo. Gracias por nunca olvidar de dónde venís pese a lo conseguido aquí, y por no perder el cariño a vuestra tierra y su esencia. Gracias por cuidarme todos estos años, y por todo el cariño que dais sin saberlo.

A les meves cosines, gràcies. **Laia**, gràcies per fer que l'edat mai sigui un impediment entre nosaltres, per compartir moments, i demanar consell (o donar-me'l) en tots els aspectes de la vida. Vull que sàpigues que has sigut, ets i seràs sempre la meva princesa. **Marta**, gràcies per convertir-te en la germana que mai he tingut, tan diferents en molts aspectes, de caràcter fort però d'una tendresa innegable. Gràcies per derrapar allà on anem, i fem el que fem; pels brunchitos, viatges, vinets, i nits sin freno (per ser DJ i GPS, i capturar-ho tot, per recordar-ho sempre). Gràcies per cuidar-me quan més ho he necessitat, i encarregar-te de mantenir el somriure permanent. Gràcies per saber que sota la pell de responsable, mai diré que no a propostes dignes de liadas. **Jordi**, gràcies per tornar trepitjant fort i demostrar-me que sempre podré comptar amb tu.

Per últim, gràcies **papa** i **mama**. Gràcies per haver-me acompanyat sempre en totes les etapes, per haver escoltat, acceptat i recolzat totes les meves decisions en lo personal i professional (a vegades no compartides) i mai intentar influenciar-les. Gràcies per cuidar-me tant (en quantitat) i tan bé (en qualitat), per haver sigut pilar sempre, i recordar-me que casa vostra sempre serà casa meva per moltes vegades que marxi. Gràcies per animar-me a seguir endavant en aquesta aventura en els moments d'incertesa, i per confiar en mi com ningú,… definitivament, ha valgut la pena. Però, si de debò estic agraïda per alguna cosa, és pels valors amb els que m'heu criat dia rere dia, on l'esforç, la superació, l'honestedat, la lleialtat i la bondat sempre han anat per davant, i sense dubte, m'han fet arribar on sóc i ser qui sóc. Mai podré retornar-vos tot el que m'heu donat… Gràcies.

Abstract

Inflammation is a physiological response essential for maintaining homeostasis, but when the immune system becomes dysregulated, it can drive a range of pathological conditions and become a key contributor to many diseases. Understanding these complex, disease-specific inflammatory mechanisms is crucial for developing effective treatments. Single-cell sequencing technologies offer a powerful approach to study the immune system at high resolution, providing detailed insights into cellular heterogeneity and immune dynamics. To fully exploit their potential in large-scale studies, standardized protocols for sample collection and processing, coupled with advanced computational methods for robust data analysis, are needed. Since immune-driven diseases often target specific tissues and exhibit distinct inflammatory profiles, a comprehensive approach is required. Blood, as a minimally invasive sample, offers a valuable opportunity to monitor inflammation by tracking cytokine levels and cellular activity, making it ideal for elucidating disease mechanisms and therapeutic responses.

In this thesis, experimental protocols and computational approaches were developed for large-scale studies focused on analyzing human inflammatory profiles using single-cell RNA sequencing (scRNA-seq). First, we implemented FixNCut, a protocol that preserves transcriptional profiles and tissue composition by fixing samples at the time of collection, enabling subsequent tissue dissociation with minimal stress-induced artifacts. We then generated a comprehensive reference of the inflammatory spectrum in circulating immune cells using single-cell transcriptomic profiles from over 1,000 patients suffering from 20 immune-driven diseases and healthy individuals. By analyzing the immune cell composition and inflammatory molecule expression, we characterized the immune response across diseases, identifying both commonalities and disease-specific differences. Using interpretable machine learning algorithms, we captured disease-specific inflammatory genes at the cell type level that could serve as clinical biomarkers. Furthermore, we proposed a framework for classifying patients based on their blood-derived inflammatory profiles. Finally, we studied the immunomodulatory effects of albumin treatment in patients with acute decompensated cirrhosis, which resulted in immune transcriptional reprogramming that primarily affected B lymphocytes and enhanced neutrophil antimicrobial activity, boosting the patients' immune response against infections.

In summary, this thesis integrates scRNA-seq experimental and computational methods to facilitate large-scale, decentralized research on immune-driven diseases with the goal of applying these approaches for disease diagnosis, monitoring progression, and assessing therapy responses in clinical settings.

Sinopsi

La inflamació és una resposta fisiològica essencial per mantenir l'homeòstasi, però quan el sistema immunitari es desregula, pot resultar patològica i desencadenar malalties. Entendre la complexitat dels mecanismes inflamatoris de cada malaltia és clau per desenvolupar tractaments efectius. Les tecnologies de seqüenciació de cèl·lula única permeten estudiar el sistema immunitari a alta resolució, proporcionant informació sobre la heterogeneïtat cel·lular i la dinàmica de la resposta immunitària. Per maximitzar-ne el potencial, calen protocols estandarditzats per a la recollida i processament de mostres així com mètodes computacionals que garanteixin un anàlisi robust. Les malalties amb component immunitari sovint afecten teixits específics i presenten perfils inflamatoris diferents pel que cal un enfoc global. La sang permet monitoritzar la inflamació mesurant els nivells de citocines i l'activitat cel·lular, convertint-la en una mostra ideal per entendre els mecanismes de la malaltia i les respostes terapèutiques.

En aquesta tesi, s'han desenvolupat protocols experimentals i eines computacionals per a l'estudi de perfils inflamatoris humans amb seqüenciació d'ARN de cèl·lula única (scRNA-seq). Per començar, hem implementat el FixNCut, un protocol que fixa les mostres en el moment d'obtenció, preservant el perfil transcripcional i la composició cel·lular i minimitzant els artefactes tècnics. També hem creat una referència exhaustiva de l'espectre inflamatori de cèl·lules immunitaries de la sang amb més de 1,000 pacients, incloent 20 malalties i individus sans. Analitzant la composició cel·lular i l'expressió de molècules inflamatòries, hem caracteritzat la resposta immunitària d'aquestes malalties. Mitjançant l'ús d'algoritmes d'aprenentatge automàtic, hem identificat gens inflamatoris especifics de malaltia i tipus cel·lular que podrien servir com a biomarcadors clínics, i hem proposat una estratègia per classificar pacients segons els seus perfils inflamatoris derivats de la sang. Finalment, hem analitzat els efectes immunomoduladors del tractament amb albúmina en pacients amb cirrosi descompensada aguda, observant reprogramació transcripcional en els limfòcits B i un augment de l'activitat antimicrobiana dels neutròfils, millorant la resposta immunitària envers infeccions.

Per concloure, aquesta tesi integra mètodes experimentals i computacionals de scRNA-seq per la investigació de malalties originades per la disfunció del sistema immunitari amb l'objectiu d'aplicar aquests coneixements al diagnòstic, monitoratge i respostes terapèutiques en un context clínic.

Table of contents

Acknowledgments ..... **V**
Abstract ..... **XI**
Sinopsi ..... **XIII**
Table of contents ..... **XV**
Abbreviations ..... **XVII**

INTRODUCTION **1**
1. Immune system: key player in health and disease 3
1.1. Innate immunity 3
1.1.1. Cellular components of innate response 4
1.1.2. Soluble mediators of innate response 5
1.1.3. Signaling pathways of innate responses 6
1.2. Adaptive immunity 7
1.2.1. Humoral responses: B cells 8
1.2.2. Cell-mediated responses: T cells 8
1.2.3. Soluble mediators of adaptive response 10
1.3. Orchestrating the inflammatory response 11
1.3.1. Resolution of inflammation 12
1.4. Immune-driven diseases 13
1.4.1. Immune-mediated Inflammatory Diseases (IMIDs) 13
1.4.2. Acute inflammation 16
1.4.3. Chronic inflammation 17
1.4.3.1. Acute-on-chronic liver failure in acute decompensated cirrhosis 18
1.4.4. Infection 19
1.4.5. Solid tumors 21
1.5. Using inflammation as a diagnostic tool 21
2. Single-cell technology 23
2.1. Origin and evolution 23
2.2. Experimental design 24
2.2.1. Single-cell RNA-sequencing (scRNA-seq) 25
2.2.2. Sample input 26
2.3. Raw data pre-processing 29
2.3.1. From sequencing reads to count matrices 29
2.3.2. Sample demultiplexing: sample-of-origin cell identity 30
2.4. scRNA-seq data processing 32
2.4.1. Data quality control 32
2.4.2. Normalization 34
2.4.3. Dimensionality reduction 34
2.5. Identifying cellular structures 36
2.5.1. KNN graph representation 36
2.5.2. Unsupervised clustering: Leiden algorithm 36
2.5.3. Cell identity annotation 37

2.6. Batch-effect removal and data Integration....39
2.7. scRNA-seq data analysis....40
2.7.1. Comparative analysis of experimental conditions.... 40
2.7.2. Machine Learning (ML) for single-cell genomics....42

OBJECTIVES.... 47

PUBLICATIONS.... 51
THESIS DIRECTOR'S REPORT.... 53
Publication 1. FixNCut: Single-cell genomics through reversible tissue fixation and dissociation.... 55
Publication 2. Interpretable Inflammation Landscape of Circulating Immune cells....101
Publication 3. Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis....155

DISCUSSION....205
1. Large-scale research projects: overcoming technical challenges....207
1.1. Sample preservation for decentralized studies....207
1.2. Hybrid sample demultiplexing: combining cell-tagging and SNP-based deconvolution methods.... 210
1.3. Impact of data processing....212
2. Characterizing disease inflammatory profile....218
2.1. Inflammation signatures across diseases and cell types....218
2.2. Biomarker discovery through interpretable modeling.... 219
2.3. Patient stratification based on inflammatory profiles.... 220
3. The immunomodulatory effects of albumin in acute decompensated cirrhosis.... 221
4. Core scientific principles.... 223
4.1. Fostering Collaboration....223
4.2. FAIR data and Open Code.... 223
4.3. Widespread Science Dissemination.... 224
5. Shaping the future....225

CONCLUSIONS....227

REFERENCES....231
APPENDICES....249
Appendix A. FixNCut: a Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assay.... 251
Appendix B. Metastatic recurrence in colorectal cancer arises from residual EMP1+ cells. 267
Appendix C. Mex3a marks drug-tolerant persistent colorectal cancer cells that mediate relapse after chemotherapy.... 287

Declarations....309

Abbreviations

Ab: Antibody
ACLF: Acute-on-Chronic Liver Failure
ACPA: Anti-Citrullinated Protein Antibody
AD: Acute Decompensation
AE: Autoencoder
Ag: Antigen
AIDS: Acquired ImmunoDeficiency Syndrome
ANA: Antinuclear Antibody
APC: Antigen Presenting Cell
ARI: Adjusted Rand Index
BAM: Binary Alignment Map
BCL: Binary Base Call
BCR: B Cell Receptor
BRCA: BReast CAncer
CD: Crohn's Disease
CGD: Chronic Granulomatous Disease
CLR: Centered Log Ratio
CMO: Cholesterol-Modified Oligonucleotide
CNS: Central Nervous System
COPD: Chronic Obstructive Pulmonary Disease
CPM: Count-per-million
CRC: ColoRectal Cancer
CVAE: Conditional Variational Autoencoder
DAMP: Damage-Associated Molecular Pattern
DC: Dendritic Cell
DEA: Differential Expression Analysis
DEG: Differentially Expressed Gene
DL: Deep Learning
DNN: Deep Neural Network
DSP: Dithiobis-Succinimidyl Propionate
FAIR: Findable, Accessible, Interoperable and Reusable
EGA: European Genome-phenome Archive
FC: Fold-change
FFPE: Formalin-Fixed, Paraffin Embedded
GBDT: Gradient Boosted Decision Tree
GEM: Gel Bead-in-Emulsions
GEO: Gene Expression Omnibus
GME: Generalized Mixed Effect
GSEA: Gene Set Enrichment Analysis
GSAI: Gene Set Activity Inference
HBV: Hepatitis B Virus
HC: Healthy Control
HCA: Human Cell Atlas
HEG: Highly Expressed Genes
HIV: Human Immunodeficiency Virus
HLA: Human Leukocyte Antigen
HNSCC: Head and Neck Squamous Cell Carcinoma
HSA: Human Serum Albumin
HT: High-Throughput
HTO: Hashtag Oligonucleotides
HVG: Highly Variable Genes
PCA: Principal Component Analysis
IBD: Inflammatory Bowel's Disease
IFN: Interferon
Ig: Immunoglobulin
IL: Interleukin
ILC: Innate Lymphoid Cell
IMID: Immune-driven Inflammatory Disease
IRF: Interferon Regulatory Factor
JAK: JAnus Kinases
KNN: K-Nearest Neighbors
LISI: Local Inverse Simpson's Index
LMO: Lipid-Modified Oligonucleotide
MAIT: Mucosal-Associated Invariant T
MAPK: Mitogen-Activated Protein Kinase
mDC: Myeloid Dendritic Cells
MDSC: Myeloid-Derived Suppressor Cells
MHC: Major Histocompatibility Complex

ML: Machine Learning
MLM: Multivariate Linear Model
MNN: Mutual nearest Neighbors
MS: Multiple Sclerosis
NF-kβ: Nuclear factor kappa-light-chain enhancer of activated B cells
NK: Natural Killer
NLR: NOD-Like Receptors
NMI: Normalized Mutual Information
NO: Nitric Oxygen
NPC: NasoPharyngeal Cancer
PAMP: Pathogen-Associated Molecular Pattern
PBMC: Peripheral Blood Mononuclear Cell
PC: Principal Component
pDC: Plasmacytoid Dendritic Cell
PMN: PolyMorphoNuclear Leukocytes
polyA: Polyadenylated
PRR: Pattern Recognition Receptor
PS: PSoriasis
PsA: Psoriatic Arthritis
QC: Quality Control
RA: Rheumatoid Arthritis
RF: Rheumatoid Factor
RLR: RIG-I-like Receptor
ROS: Reactive Oxygen Species
RT: Reverse Transcription
scRNA-seq: single-cell RNA sequencing
SHAP: SHapley Additive exPlanations
SLE: Systemic Lupus Erythematosus
SNP: Single Nucleotide Polymorphisms
SRA: Sequence Read Archive
STAT: Signal Transducers and Activators of Transcription
STEM: Science, Technology, Engineering and Mathematics
TCR: T Cell Receptor
TF: Transcription Factor
TGF: Transforming growth factor
Th: T Helper cell
TLR: Toll-Like Receptors
TME: Tumor MicroEnvironment
TNF: Tumor necrosis factor
Treg: T Regulatory cell
t-SNE: t-distributed Stochastic Neighbor Embedding
TSO: Template Switch Oligo
UC: Ulcerative Colitis
ULM: Univariate Linear Model
UMAP: Uniform Manifold Approximation and Projection
UMI: Unique Molecular Identifiers

Introduction

1. Immune system: key player in health and disease

The immune system is orchestrated by a combination of molecules, cells, tissues, and organs that work together to maintain homeostasis. To ensure a comprehensive defense mechanism against a wide range of potential challenges, it comprises two main components: the innate immune response and the adaptive immune response. While the innate immune system provides an immediate but non-specific response to threats, the adaptive immune system targets specific invaders and offers long-term protection. The interplay between both components provides strong protection against pathogens but requires careful regulation to prevent persistent inflammation, autoimmunity, immune deficiencies, or the development of tumors. Understanding immune response and its dysregulation in disease is crucial for exploring therapeutic options[1,2].

1.1. Innate immunity

Innate immunity serves as the body's first line of defense against pathogens and is characterized by its rapid, non-specific reaction to various harmful agents including bacteria, viruses, fungi, and parasites. Upon interaction between the host and the invading pathogen, a signaling cascade is triggered by the release of pre-existing molecules, leading to immune cells recruitment to the site of infection and the induction of an inflammatory tissue response[3]. This mechanism boosts the immune response and connects the innate and adaptive branches of the immune system. This immune response involves various components, including physical barriers (skin and mucosa), cellular components, antimicrobial peptides, soluble mediators such as cytokines and serum proteins, and cellular receptors (Figure 1)[4,5].

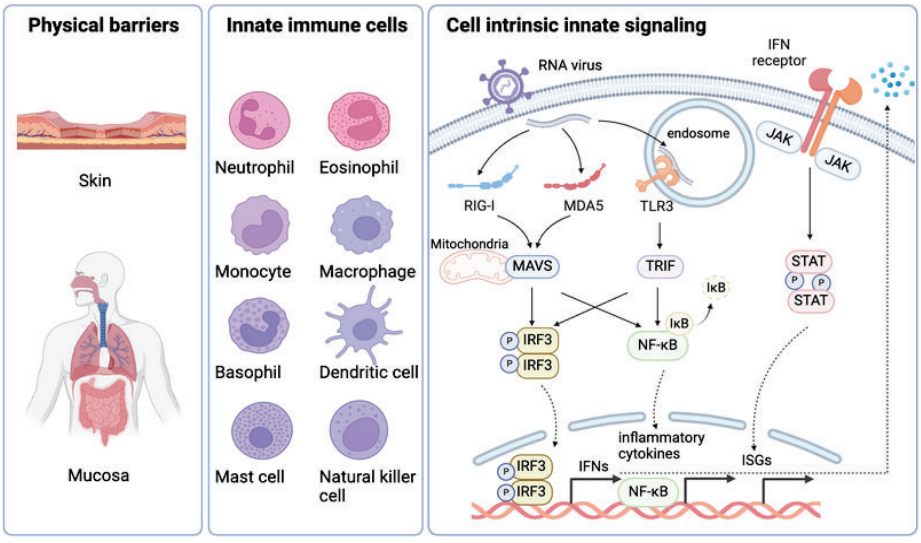


Figure 1. Innate immune system. The figure illustrates the major components of the innate immune system, including physical barriers (**left**), cellular components (**middle**), and an example of signaling pathways (**right**), encompassing cellular receptors, intermediate inflammatory mediators, and transcription factors. *Extracted from BioRender [https://www.biorender.com/]*.

1.1.1. Cellular components of innate response

The innate immune system's cellular components include phagocytic cells, such as granulocytes (neutrophils, eosinophils, basophils, and mast cells), monocytes/macrophages, dendritic cells, innate lymphoid cells, natural killer cells, endothelial/epithelial cells, and platelets[4].

- **Granulocytes: Neutrophils** are the first responders recruited to the inflammation site, where they release their granule contents[6]. **Eosinophils**, found in the respiratory, gastrointestinal, and urinary tracts, combat parasites and play a role in allergic responses[7]. In contrast, **basophils** and **mast cells**, located in mucosal and connective tissues, are involved in allergic reactions and viral defense[8].

- **Monocytes** circulate in the bloodstream and migrate to inflammatory sites in response to specific stimuli, where they differentiate into tissue macrophages. **Macrophages** play a key role in antimicrobial and anti-tumoral activities by promoting cytokine release and modulating the immune response, including antigen presentation when the adaptive response is required. They also aid in resolving inflammation and tissue repair[9].

- **Dendritic cells (DC)**, the main antigen-presenting cells (APC), are primarily located in the skin and mucosal tissues. They capture and process antigens, release soluble mediators, and migrate to primary lymphatic nodes, where they perform antigen presentation to active T cells, facilitating cell-mediated immunity against pathogens. Additionally, DCs play a role in inducing peripheral immunological tolerance[10].

- **Innate lymphoid cells (ILC)** regulate tissue homeostasis and inflammation without responding to antigens in a specific manner. There are different types of ILCs based on their functions. Notably, Group I, which includes ILC1 and Natural Killer cells, produce pro-inflammatory and type 1 cytokines and induce cytotoxicity through the release of perforin and granzymes. Specifically, **Natural Killer (NK) cells** help to regulate the immune response by detecting and destroying virus-infected or malignant cells while maintaining the immunological homeostasis and controlling tumor growth[11].

- **Epithelial and endothelial cells**, while not classified as immune cells, act as physical barriers and detect pathogens through pattern recognition receptors (PRRs) that recognize pathogen-associated molecular patterns (PAMPs) on microbial surfaces. They also secrete antimicrobial molecules and cytokines to eliminate pathogens and amplify the immune response[12].

- **Platelets** release chemotactic molecules to attract leukocytes to the inflammatory sites[13].

1.1.2. Soluble mediators of innate response

Cytokines, including TNF-α, interferons, interleukins, chemokines, and TGF-β, are small, soluble, cell-signaling protein molecules that play a crucial role in inflammation. They mediate interactions among cells involved in the innate response and the development of acquired immunity[4]. On one hand, pro-inflammatory cytokines (e.g., TNF-α, IL-1β, and IL-6) are primarily produced by activated innate cells and are involved in the up-regulation of the inflammatory reaction through cellular activation, differentiation and proliferation. In contrast, anti-inflammatory cytokines (e.g., IL-4, IL-10, IL-11, and IL-13) regulate the pro-inflammatory cytokine response, preventing an excessive reaction that could lead to tissue damage, while also playing a key role in maintaining tissue homeostasis and ensuring proper function[14].

- **Tumor Necrosis Factor α (TNF-α)** is a pro-inflammatory cytokine that binds to TNFR1/2 receptors on endothelial cells and neutrophils, activating various signaling pathways, including NF-kB and MAPK, that upregulate other pro-inflammatory mediators. Additionally, it orchestrates apoptosis, fever and production of stress-related proteins[15].

- **Interferons (IFNs) Type I**, including **IFN-α** and **IFN-β,** interact with the IFNR receptor complex to mediate antiviral defense and immunoregulation. They enhance immune responses by activating NK cells and macrophages and by inducing the expression of class I major histocompatibility complex (MHC) molecules on somatic cells, thereby improving the recognition and destruction of infected cells[14].

- **Interleukins (ILs)** mediate communication between cells, influencing the growth and differentiation of immune cells to coordinate the immune response. Dysregulation of interleukin production or signaling contributes to various diseases, and their pathways are targeted by drugs in treatments for these conditions. Depending on the context, they can have either pro-inflammatory or anti-inflammatory effects[4,16].

  Pro-inflammatory ILs: IL-1 promotes inflammation, supports the differentiation of Th17 cells crucial for chronic inflammation and autoimmune responses, and contributes to the acute phase response[17]. **IL-6** induces the synthesis of acute phase response in the liver, supports B cell proliferation, and enhances antibody production[18]. **IL-12** drives T cell differentiation towards Th1, stimulates the production of IFN-γ, and enhances the cytotoxic activity of NK cells[19]. **IL-15** supports the proliferation and survival of NK cells and T cells[20]. **IL-18** activates NK, promotes the differentiation of T cells towards Th1, and stimulates the production of IFN-γ[21].

  Anti-inflammatory ILs: IL-10 exhibits immunosuppressive effects on specific T cell subsets through APC, helping to regulate and dampen the immune response. It also promotes B cell proliferation, and T cell differentiation into Tregs[22].

- **Chemokines** are a type of chemoattractant cytokines, also secreted by endothelial and T cells, which promote the activation and migration of primarily leukocytes to sites of inflammation, as well as inducing cell arrest or adhesion. They can be classified into four groups based on their cysteine residues: CCL (RANTES, MCP-1 and MIP-1α), CXCL (IL-8), XCL (lymphotactin) or CX3CL (fractalkine). These molecules bind to conventional (cCKRs) or atypical (ACKRs) receptors on the surface of leukocytes[23,24].

  Examples: **RANTES**, also known as **CCL5**, recruits and activates T cells, eosinophils, and monocytes through CCR1/3/5 receptors, and also plays a role in coordinating inflammatory regulation, tissue repair, and wound healing[25]. **MCP-1**, also known as **CCL2**, recruits monocytes, memory T cells, and dendritic cells regulating the migration and infiltration of immune cells by binding to CCR2 receptors[26]. **IL-8**, also known as **CXCL8**, acts as chemoattractant for neutrophils, promoting their activation and guiding them to sites of infection or injury, where they carry out phagocytic activity. It binds to CXCR1/2 receptors. Additionally, it contributes to angiogenesis[27].

- **Transforming Growth Factor β (TGF-β)** is a multifunctional cytokine that exhibits both pro- and anti-inflammatory effects through binding to TGFBR1/2 receptors. It regulates the inflammatory process by suppressing T cell activation and promoting the development of regulatory T cells, thereby maintaining immune tolerance. Additionally, TGF-β also plays a significant role in fibrosis, tissue repair, and cell growth and differentiation. In cancer, TGF-β acts as tumor suppressor in early stages but can promote progression and metastasis in later stages[28].

1.1.3. Signaling pathways of innate responses

Different signaling pathways orchestrate the innate immune response, crucial for coordinating the body's immediate defense mechanisms. Some examples are:

- The **NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) pathway** initiates and regulates the inflammatory response. Upon activation by receptors such as Toll-like receptors (TLRs) and TNF receptors, it triggers a signaling cascade that results in the activation of NF-κB. This transcription factor then induces the expression of pro-inflammatory cytokines[29].

- The **JAK/STAT (Janus kinases / Signal transducers and activators of transcription) pathway** starts when cytokines and growth factors bind to cell receptors, activating JAKs. JAK molecules then activate STAT ones, which on the nuclei control the expression of genes involved in inflammation, cell growth, cancer, and immune regulation[30].

- The **MAPK (mitogen-activated protein kinase) pathway** regulates cellular responses to stress, growth, and inflammation through a series of kinases, including MAPKs such as ERK, p38, and JNK, which are activated by upstream signals[31].

- The **IRF (interferon regulatory factor) pathway** is pivotal in the antiviral response and the production of Type I IFNs. This process is mediated by IRFs such as IRF3/7, which are activated through receptors like RIG-I-like (RLRs) and TLRs[32].

- The **inflammasome pathway** is critical for activating inflammatory responses triggered by cellular stress or pathogens. This pathway involves the assembly of a multiprotein complex, including NLRP3, ASC, and caspase-1, which leads to the maturation and secretion of pro-inflammatory cytokines[33].

- The **complement pathway** is an integral part of the innate immune system that helps identify and eliminate pathogens through opsonization and recruitment of immune cells. This system is activated by external factors via classical (complex antigen-antibody), lectin, or alternative (linked to the C3 tickover event) pathways. It also plays an important role in the homeostasis of adaptive immunity[34].

In brief, the innate immune system is fundamentally designed to prevent infection, eliminate invading pathogens, and stimulate the adaptive immune response, utilizing a variety of cells, molecules and signaling pathways to provide comprehensive defense against harmful agents.

1.2. Adaptive immunity

Adaptive immunity is a sophisticated defense mechanism that provides a specific and long-lasting response to pathogens. Unlike the innate response, the adaptive immune system tailors its responses to specific antigens by recognizing unique molecular structures on pathogens, working in conjunction with the innate immune system[35]. A key feature of adaptive immunity is immunological memory, which allows the immune system to remember previous encounters with pathogens. Upon initial exposure, it generates a primary response that is slower but generates memory cells that persist long after the infection has resolved. These memory cells enable a quicker and effective secondary response upon subsequent exposures to the same pathogen[36]. Despite its benefits, the adaptive immune response has limitations. It requires time to develop and may be less effective against certain pathogens that can evade or suppress immune detection[35].

The adaptive system is mediated by two types of lymphocytes: B cells, which drive humoral responses, and T cells, which are responsible for cell-mediated responses[37].

1.2.1. Humoral responses: B cells

B cells are lymphocytes that play a central role in the humoral immune response by producing antibodies (Ab), which are proteins also known as immunoglobulins (Ig). B cells are activated when their B cell receptors (BCRs), which are antibodies bound to the cellular membrane, recognize and bind to a specific antigens (Ag). Co-stimulation is also crucial for B cell activation and differentiation, with key signals provided through the interaction of CD40:CD40L and CD28:CD80/CD86 (also known as B7) between B and T cells, respectively (Figure 2)[37,38].

Once activated, B cells differentiate into **plasma cells**, which secrete large quantities of antibodies. The interaction between CD40 and CD40L is crucial for immunoglobulin isotype class switching, enabling the production of different antibody types such as IgG, IgA, IgE, IgD, or IgM, depending on the antigen (Figure 2). These antibodies circulate in the bloodstream until they encounter a pathogen, to which they specifically bind by recognizing its antigen. This forms Ab-Ag complexes that neutralize pathogens by blocking their ability to bind to receptors on host cells and marking them for destruction by phagocytic immune cells, such as macrophages and NK cells[37,39]. In addition, some activated B cells differentiate into **memory B cells**. These memory cells persist in the body for long periods and provide a rapid and robust response if the same antigen is encountered again in the future[37,38,40].

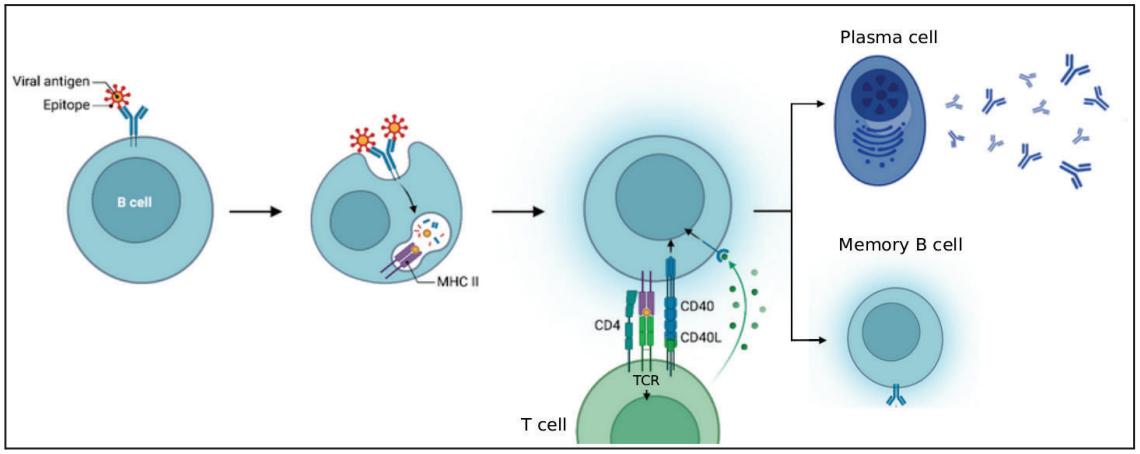


Figure 2. B cell activation and differentiation. The figure illustrates B cell activation after viral exposure. The viral particle binds specifically to the BCR, is internalized, degraded, and presented on MHC class II. Interaction with T cells through TCR, along with CD40:CD40L co-stimulation, promotes B cell activation, leading to differentiation into plasma cells (with immunoglobulin class switching) and Memory B cells. *Adapted from BioRender [https://www.biorender.com/]*.

1.2.2. Cell-mediated responses: T cells

T cells are lymphocytes that play a central role in cell-mediated immune response. T cell activation and subsequent differentiation require three simultaneous signals: interaction of the T cell receptors (TCRs) with a specific antigen presented on MHC class II molecules by APCs, signaling from co-stimulatory molecules, and cytokine production to promote clonal expansion[37].

T cell activation through antigen presentation

Naïve T cells migrate to secondary lymphoid organs to encounter specific antigens presented by an APC. In the lymph nodes, T cells make temporary contacts with various APC, but they only bind to those presenting antigens specific to their receptors. TCRs recognize and bind to specific antigens presented by class II MHC molecules (peptide-MHC complex) on the APC surface[41]. Simultaneously, the T cell co-receptor CD28 interacts with CD80/CD86 on the APC surface, providing necessary co-stimulatory signals[42]. When the TCR binds to the antigen, it triggers the reorganization of membrane molecules and signaling proteins at the contact site with the DC, forming an immunological synapse[41] (Figure 3).

Following antigen recognition and the provision of co-stimulatory signals, a complex intracellular signaling process begins, leading to the activation of three key transcription factors (TFs): NFAT, AP-1, and NF-κB. This activation is mediated by CD45, which activates the tyrosine kinases Fyn and Lck, associated with the ε chains of the CD3 and the co-receptors CD4/CD8, respectively. Once activated, these TF enter the nucleus and initiate the transcription of genes that lead to IL-2 secretion, expression of its receptor (IL-2Rα), integrins for cellular adhesion, co-stimulatory molecules like CD40L, and anti-apoptotic proteins[43–45]. IL-2, along with IL-15 and IL-21, plays a pivotal role in driving clonal expansion, resulting in the proliferation of large numbers of T cells with receptors specific to the original antigen[37,46] (Figure 3).

Differentiation of T cells into CD4 and CD8 T cell subtypes

Once T cells are activated, the binding of polarizing cytokines to their respective receptor on the cell surface drives their differentiation into distinct effector CD4+ and CD8+ cell subtypes.

- **CD8+ Cytotoxic T cells** circulate throughout the body, identifying and binding to infected or abnormal cells presenting antigens on MHC class I molecules. Upon recognition, they release cytotoxic molecules such as perforin and granzymes, which induce apoptosis in the infected cells, effectively eliminating the threat. Additionally, they release cytokines like IFN-γ and TNF-α, which help combat viral infections and regulate the proliferation of tumor cells (Figure 3)[47].

- **CD4+ Helper T cells (Th)** recognize antigens presented on MHC class II molecules and play a crucial role in activating and directing other immune cells, such as macrophages, B cells, and NK cells, to combat infections. They coordinate this response by secreting specific cytokines. There are several Th cell phenotypes, each with a distinct cytokine profile and functions: Th1, Th2, Th17, regulatory T cells (Treg), as well as the recently identified Th9, Th22, and follicular helper T cells (Figure 3)[48,49].

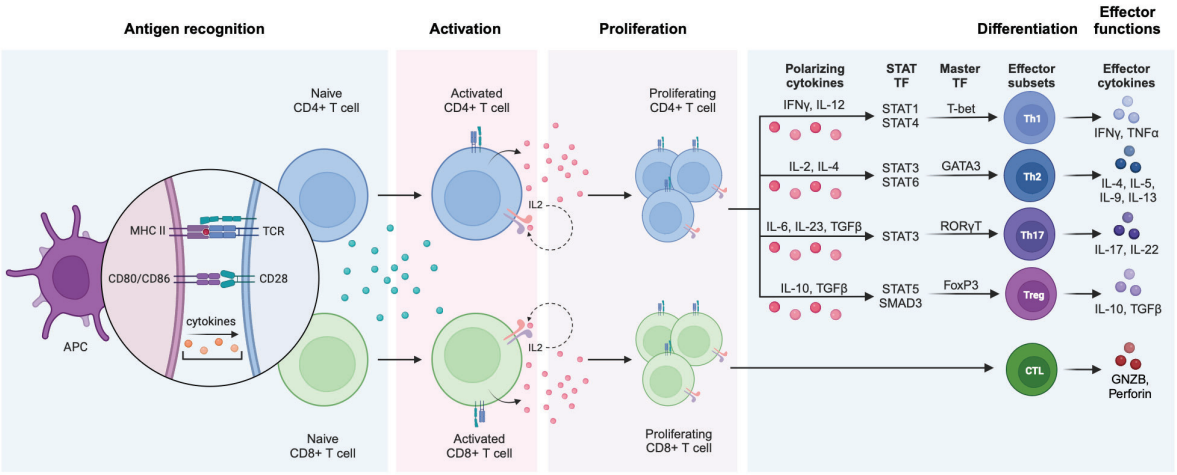


Figure 3. T cell activation, proliferation and differentiation. The figure illustrates T cell activation upon interaction with APCs presenting antigens on MHC class molecules. This interaction, coupled with CD80/86:CD28 co-stimulation and cytokine signaling, leads to T cell activation and proliferation. Subsequently, polarizing cytokines induce gene expression changes, resulting in differentiation into various CD4+ and CD8+ subtypes. *Extracted from BioRender [https://www.biorender.com/].*

In parallel, a subset of the activated T cells undergoes further maturation to form **long-lived memory T cells**, which persist in the body to provide a rapid and robust immune response upon subsequent encounters with the same antigen[50]. However, under certain disease conditions, such as chronic infections or cancer, this process is altered, and the T cells evolve into an exhausted or dysfunctional state. **Exhausted T cells** decrease their proliferative potential, reduce their effector functions, and promote anti-inflammatory response mechanisms through the upregulation of inhibitory receptors, such as PD1, CTLA4, TIGIT, TIM3, and LAG3, which ultimately contribute to the impaired immune response[51].

1.2.3. Soluble mediators of adaptive response

In addition to the cytokines directly involved in innate immunity, others play a role in the adaptive immune response, with either pro-inflammatory or anti-inflammatory functions.

- **Pro-inflammatory interleukins**: **IL-2** stimulates the proliferation of T and B cells and promotes the development of regulatory T cells to help maintain immune tolerance. It drives clonal expansion through autocrine and paracrine signaling[46]. **IL-5**, secreted by Th2, supports the growth and activation of eosinophils and B cell maturation, while **IL-9** promotes the survival of T and mast cells, both involved in allergic responses and asthma[52]. **IL-15** and **IL-21** also participate in the clonal expansion, enhancing T and NK cell activity[53]. **IL-17** promotes neutrophil recruitment and is involved in autoimmune diseases[54]. **IL-22** and **IL-23** play roles in barrier defense and chronic inflammation, especially in skin and mucosal tissues. **IL-23** drives differentiation of Th17 cells[55].

- **Anti-inflammatory interleukins: IL-4** and **IL-13** promote Th2 cell differentiation, induce immunoglobulin class switching (IgE), and B cell proliferation[56]. They are also involved in allergic responses. **IL-11** regulates chronic inflammation and promotes tissue repair[57].

- **Chemokines**. Examples: **XCL1** recruits and activates T cells and DC through its receptor XCR1[58]. **CXCL13** primarily recruits B cells by binding to its receptor CXCR5, driving lymphoid tissue organization and the development of follicles[59].

- **IFN-γ**, a **Type II IFN**, interacts with the IFNGR1/2 receptor complex to activate the JAK-STAT signaling pathway. It enhances the antimicrobial and antiviral activity of macrophages, neutrophils, and NK cells, and promotes cell-mediated immunity. IFN-γ upregulates MHC-I expression on somatic cells and MHC-II on APCs, facilitating antigen presentation. It also drives differentiation of T cells into Th1[14,60].

- **TNF-β,** also called **lymphotoxin-α,** is a pro-inflammatory cytokine mainly produced by activated lymphocytes, that activates innate immune cells, such as macrophages and neutrophils, enhances the production of other pro-inflammatory cytokines, and contributes to the development of lymphoid organs. Also, due to its cytotoxic activity, it can directly induce cell apoptosis[61].

1.3. Orchestrating the inflammatory response

The interplay between the innate and adaptive immune systems is essential for effective pathogen defense. The innate response acts as the first line of defense, rapidly responding to threats in a nonspecific manner. It also sets the stage for the adaptive immunity by shaping it through antigen presentation and cytokine secretion, which activate B and T cells. This activation leads to a targeted defense and the establishment of long-term immunity. Proper regulation of the immune response is critical to ensure that both innate and adaptive immunity work together to eliminate pathogens while preventing excessive inflammation or self-damage[2,62] (Figure 4).

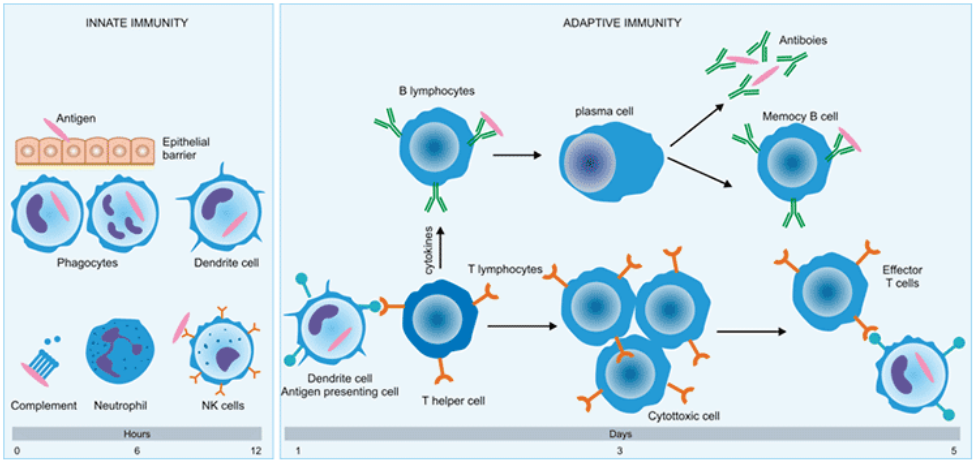


Figure 4. The basic components and interplay of the innate and adaptive immune responses. *Extracted from Cusabio [https://www.cusabio.com/].*

1.3.1. Resolution of inflammation

Inflammation is a pathophysiological reaction central to the immune response; while it is beneficial and vital for defending against harmful agents, it must be controlled to prevent excessive tissue damage, chronic inflammation or disease[63].

The **resolution phase**, which marks the return to normal inflammatory conditions, involves more than just breaking down inflammatory mediators and halting inflammatory processes. It is an active and coordinated anti-inflammatory and pro-resolving program aimed at restoring tissue homeostasis, integrity, and function. This process includes several key steps: depletion in cytokine signaling, cease of inflammatory cell recruitment (e.g., neutrophils), induction of effector cell death and removal of apoptotic cells, reprogramming of macrophages, removal of immune cells (e.g., APCs), positive modulation of suppressive immune cells, and the induction of tissue repair to restore normal tissue function and return to body homeostasis[64,65].

Cellular cleanup

Once neutrophils migrate to sites of inflammation, their lifespan and fate are influenced by the local inflammatory environment. The balance between neutrophil survival and death is regulated by a network of signaling factors, such as TNF-α concentrations and iron saturation state[66]. Apoptotic neutrophils help attenuate inflammation by secreting mediators that inhibit further recruitment of additional neutrophils. Through the expression of “find me” and “eat me” molecules such LPC, S1P, ATP and UTP, they facilitate their own clearance by macrophage phagocytosis in a process called **efferocytosis**[67]. Following efferocytosis, macrophages cease production of pro-inflammatory cytokines and switch to a resolution-phase program, which involves releasing IL-10 and TGF-β, as well as the secretion of chemoattractants like XCL1, CCL5, and CXCL13 to recruit lymphocytes. Those repopulate sites of resolved inflammation and contribute to protective effects during the post-resolution phase. Additionally, they attenuate TNF-α-mediated NF-κB activation and express TIM4 and TGF-β, which are crucial for reinforcing the anti-inflammatory signaling pathway, clearing cells, and restoring tissue homeostasis[64,68].

Suppressive immune cells

The cellular components of the resolution phase are largely driven by suppressive immune cells, including regulatory T cells and myeloid-derived suppressor cells.

- **Regulatory T cells (Tregs),** a subset of CD4+ T cells, are crucial for maintaining immune balance by suppressing the activation and effector functions of various immune cells, including autoreactive CD4+ and CD8+ T cells (which evade thymic negative selection), NK cells, B cells, and APCs. They promote neutrophil apoptosis, enhance macrophage efferocytosis, and secrete pro-resolving and anti-inflammatory cytokines such as IL-10

and TGF-β. While Tregs are essential for controlling immune tolerance, their suppressive effects can hinder antitumor responses, facilitating tumor progression[69].

- **Myeloid-derived suppressor cells (MDSCs)** modulate the immune response and arise from abnormal differentiation of immature myeloid in pathological contexts such as infections, sepsis, trauma, or tumors. They clear apoptotic neutrophils through the release of nitric oxide (NO) and reactive oxygen species (ROS), which contribute to programmed cell death. Additionally, MDSCs release anti-inflammatory cytokines like IL-10 and enhance the expansion of Tregs by increasing *Foxp3* expression in a manner dependent on IL-10, TGF-β, and arginase[70].

1.4. Immune-driven diseases

Disruptions in the balance between the innate and the adaptive immune system, as well as in the resolution of inflammation, can lead to various immune-driven diseases, due to either excessive or insufficient immune responses. Autoimmunity, for example, occurs when the immune system mistakenly targets self-tissues. Chronic inflammation, on the other hand, can create an environment favorable to tumor growth, facilitating cancer development. Inadequate immune responses can also result in immunodeficiencies, increasing vulnerability to infections[71–73].

Given the extensive range of diseases, I will focus on the classification grouping and diseases presented in **Publication 2** of this thesis.

1.4.1. Immune-mediated Inflammatory Diseases (IMIDs)

Immune-mediated inflammatory diseases (IMIDs) are a heterogeneous, clinically diverse group of disorders characterized by an aberrant immune response directed to self-antigens, leading to prolonged, chronic inflammation that results in tissue damage and disruption of normal physiological functions. This inappropriate immune activation occurs in genetically predisposed individuals triggered by specific environmental or immunological factors[74].

Although IMIDs share common immune mechanisms, they differ in clinical phenotype, tissue location, and sex or age distribution. The spectrum of IMIDs include: rheumatoid arthritis, psoriasis, psoriatic arthritis, inflammatory bowel diseases, multiple sclerosis, systemic lupus erythematosus, among others.

- **Rheumatoid arthritis (RA)** is an autoimmune disease primarily affecting joints in a symmetrical pattern, leading to pain, swelling, stiffness, and loss of function. Women are 2-3 times more likely to develop RA than men. The risk of RA is significantly associated with specific MHC-II, also known as human leukocyte antigen (HLA) loci. Clinically, RA is marked by the presence of autoantibodies targeting IgG (rheumatoid factor, RF) and citrullinated proteins, referred to as anti-citrullinated protein antibodies (ACPAs), which

are generated against modified mucosal peptides recognized by APCs. This initiates synovial inflammation characterized by mononuclear cell infiltration, predominantly by CD4+ T cells and macrophages. The formation of immune complexes and complement activation induces cytokine production (including CCL19, CCL21, TNF, IL-1/2/6/8/17, RANKL, MMPs), leading to synovial vascular leakage and chronic inflammation, eventually causing structural joint destruction. Currently, successful therapeutic strategies target IL-6 and TNF cytokines, as well as the JAK-STAT pathway, with limited success in targeting IL-1 and IL-17[75–77].

- **Psoriasis (PS)** is an immune-mediated skin disease that manifests in various phenotypes, characterized by cycles of skin cell replication, forming thick, red patches. The genetics of PS is linked to variations in the *HLA-C* gene. In PS, the self-nucleotides form complexes detected by plasmacytoid DCs (pDCs), which release IFN-α/β and stimulate myeloid DCs (mDCs) to secrete IL-12/23 and TNF. This immune cascade leads to the activation of Th1, Th17, and Th22 cells, which migrate into the epidermis. There, they trigger the release of local factors such as TNF, IL-17/22, and CCL20, promoting further immune infiltration and stimulating keratinocyte proliferation. PS is also marked by the activation of cAMP, as well as the JAK-STAT and NF-κB signaling pathway. Current treatments for PS focus on inhibiting TNF, IL-12, IL-23, and IL-17[78–80].

- **Psoriatic arthritis (PsA)** is an autoimmune disease primarily affecting joints, commonly developing as comorbidity in patients that already have psoriasis. Diagnosis is based on clinical symptoms, with negative tests for RF and ACPAs, which distinguish PsA from RA. Genetic susceptibility to PsA involves specific HLA alleles and SNPs in genes related to the IL-17/IL-23 pathway, as well as interferon and NF-κB signaling. The immune activation phase in PsA is driven by the infiltration of Th cells, particularly Th17, and the overexpression of cytokines such as IL-23, IL-17A, IL-22, and TNF. These factors contribute to inflammation, excessive bone remodeling, and joint damage. Current treatments include TNF, IL-12, IL-23, and IL-17 inhibitors. Alternative therapies involve blocking co-stimulatory signals (CD80/CD86:CD28) and using JAK inhibitors[81–83].

Inflammatory bowel diseases (IBD) include disorders like Crohn's disease and ulcerative colitis. Both are characterized by chronic inflammation but differ in their pathophysiology, the parts of the gastrointestinal tract affected, disease progression, and treatment approaches[84].

- **Crohn's disease (CD)** is an IBD that can affect any part of the gastrointestinal tract in a transmural manner, affecting all layers of the bowel wall and leading to fibrosis. Prevalence varies by sex and age, with a higher risk in males during childhood but becoming more predominant in females in adulthood. Genetic susceptibility, including variations in *NOD2*, *ATG16L1* and *IL23R* genes, as well as factors such as gut dysbiosis

and diet, play significant roles in its development. Epithelial barrier dysfunction allows luminal contents to penetrate the lamina propria, triggering DCs to activate naive T helper (Th0) cells, as well as Th1, Th17 and Th2 cells. These T cells produce pro-inflammatory cytokines, including IFN-γ, TNF and IL-17. In response, macrophages release IL-12 and IL-23, which activate NK cells and further perpetuate the intestinal inflammation through the release of additional cytokines such as TNF, IL-1, IL-4, IL-6, IL-8, IL-12, IL-21, IL-18, and IL-22. Current anti-inflammatory treatments target TNF, as well as IL-12 and IL-23 signaling pathways[85–87].

- **Ulcerative colitis (UC)** is an IBD that affects the rectum and extends into the colon, involving only the mucosal layer and causing superficial inflammation. UC occurs across all ages, with no sex-based differences, and is associated with modest loci risk factors at specific genetic loci. In UC, disruption of the intestinal epithelium allows increased microbiota to cross the barrier, which activates macrophages and APCs, resulting in chemokine expression that attracts neutrophils and promotes immune cell extravasation. Infiltrating monocytes mature to macrophages and produce TNF, IL-12, IL-23 and IL-6, which polarize Th1 cells. Additionally, IL-36γ derived from the epithelium triggers fibrogenesis, inhibits Tregs, and induces Th9 cell polarization, while IL-13 from NK cells contributes to barrier dysfunction. Current treatment strategies focus on symptom management with 5-ASA, corticosteroids and thiopurines, along with therapies targeting TNF, IL-12, and IL23p40. Anti-adhesion therapies and JAK inhibitors are also used[88,89].

- **Multiple sclerosis (MS)** is a neurodegenerative disease of the central nervous system (CNS) that primarily affects young adults, with higher prevalence in women. It is characterized by the accumulation of demyelinating lesions in both the white and grey matter of the brain and spinal cord. MS arises from complex gene-environment interactions, with Epstein–Barr virus infection being a significant risk factor. Genetic susceptibility is associated with SNPs in genes encoding HLA, IL2, IL7R, and TNF. In MS, resident CNS cells such as microglia and astrocytes release mediators that recruit immune cells, including B, T, and myeloid cells, further driving inflammation. These cells create a pro-inflammatory environment marked by reactive oxygen species (ROS), mitochondrial dysfunction, iron accumulation, hypoxia, and altered glutamate homeostasis, all contributing to neuronal demyelination. Astrocytes also release CCL2 and GM-CSF, which promotes microglia recruitment and activation while inhibiting oligodendrocyte remyelination. Additionally, immune dysregulation is exacerbated by Th17 secreting IL-17, Th1 releasing IFN-γ and insufficient Tregs. Novel therapies, including S1P inhibitors and anti-CD20, are in clinical trials to treat MS patients[90–92].

- **Systemic lupus erythematosus (SLE)** is a potentially fatal, multisystemic autoimmune disease primarily affecting women between puberty and menopause due to hormonal

and endocrine factors. While SLE is mostly inherited in a polygenic manner, it is often triggered by environmental factors. Nucleic acid-containing cellular particles activate TLRs on pDCs, releasing type I IFN that triggers the innate immune system and forms neutrophil extracellular traps. IFNs activate the JAK-STAT pathway in APCs, enhancing the presentation of antigens (including self-antigens) to T cells, which amplifies cytokine production, inflammation, and the self-directed immune response. CD40:CD40L interactions and TNF from DCs drive B cells to produce autoantibodies, such as antinuclear antibodies (ANAs). The loss of immune tolerance allows self-reactive T and B cells to persist. Immune complexes containing nucleic acids and ANAs trigger inflammation, and when deposited in target tissues, they cause organ damage. Current treatments include corticosteroids and immunosuppressive agents, while new therapies in clinical development focus on B cell modulation, IFN antagonists, and IL-6/IL-10 blockers[93–95].

1.4.2. Acute inflammation

Acute inflammatory diseases are characterized by a rapid onset of severe inflammation in response to triggers, such as infections, injuries, or irritants. This inflammation represents the body's quick response to address and resolve the initial damage. Inflammatory conditions such as acute appendicitis, bronchitis, and severe infections, if excessive or uncontrolled, can cause extensive tissue damage and lead to serious systemic complications, including sepsis[63].

- **Sepsis** is a severe and life-threatening condition resulting from a dysregulated innate immune response, characterized by uncontrolled systemic inflammation triggered by an infection. This leads to widespread tissue and organ damage far from the original infection site. If not promptly treated, sepsis can progress to **septic shock**, where mortality rates rise significantly, from 15-25 to 30-50%. Sepsis begins with the host's recognition of PAMPs, activating inflammatory signaling pathways such as JAK/STAT, MAPK, and NF-κB along with the complement system. These pathways lead to the production of type I IFN and the early expression of inflammatory genes like TNF, IL-1, IL-12, and IL-18, triggering a cascade of cytokines and chemokines (e.g., IL-6, IL-8, IFN-γ, CCL2, CCL3, CXCL10). This systemic inflammation can affect multiple organs, causing neurological, pulmonary, cardiovascular, renal, hematological and/or hepatic damage. Following the acute phase, sepsis often results in an immunocompromised state due to the presence of immature polymorphonuclear leukocytes (PMNs), myeloid-derived suppressor cells (MDSCs), and monocytes differentiating into M2 macrophages. Currently, treatment focuses on infection control, hemodynamic stabilization, and modulating the septic response[96,97].

1.4.3. Chronic inflammation

Chronic inflammation is a persistent inflammatory response that can last for extended periods, leading to progressive tissue damage and organ dysfunction. It can be caused by the failure to eliminate a foreign agent, exposure to low levels of irritants, autoimmunity, defects in resolving inflammation, or recurrent episodes of acute inflammation. Chronic inflammation is associated with conditions such as chronic respiratory diseases, liver disorders, cancer, and others[72].

- **Asthma** is a heterogeneous disease characterized by chronic, ongoing airway immune activation and hyperreactivity, resulting in structural remodeling and expiratory airflow limitation, often accompanied by other allergic comorbidities. Its development involves strong genetic and environmental components, with sex and age playing key roles. Asthma involves viral-induced wheezing and allergen sensitization. Allergen detection in the airways activates DCs, which process the allergens and migrate to lymph nodes to interact with naive T cells, promoting their differentiation into Th2 cells. Th2 cell maturation and migration into the mucosa are influenced by the secretion of IL-25, IL-33, CCL17, CCL22 and TSLP by epithelial cells. Th2 cells and/or ILC type 2 then produce pro-allergic cytokines such as IL-3, IL-4, IL-5, IL-9, IL-13, and GM-CSF, leading to the IgE, mast cell and eosinophil responses characteristic of allergic asthma. Asthma can also be driven by basophils, monocytes/macrophages, and Th1 and Th17[98,99].

- **Chronic obstructive pulmonary disease (COPD)** is an inflammatory disease that predominantly affects peripheral small airways and lung parenchyma, causing persistent airflow limitation and emphysema. The most common cause of COPD is exposure to irritant agents such as cigarette smoke. Chronic inhalation of irritants activates TLRs, triggering an innate immune response that increases neutrophils and macrophages in the lungs, as well as activating airway epithelial cells and mucus secretion. Importantly, ROS activates NF-κB and p38 MAPK signaling pathways, resulting in increased expression of inflammatory genes and cytokines like TNF, IL-1, and IL-6. Subsequently, the activation of T and B cells coupled with macrophages that have reduced capacity for efferocytosis of apoptotic cells, results in a failure to resolve inflammation. Treatment for COPD relies on bronchodilators, including β_2-agonists and anticholinergics[100–102].

- **Liver cirrhosis** results from prolonged liver inflammation, leading to hepatic fibrosis, ultimately causing liver failure. It can stem from unresolved inflammation due to viral infections, alcohol consumption, metabolic or genetic factors, and autoimmune diseases. Cirrhosis is a systemic disease with varying degrees of inflammation and immune deficiency, depending on its stage. Kupffer cells, the liver's resident macrophages, clear apoptotic neutrophils and pathogens, activate T (polarizing them into Th1) and iNKT cells, and release cytokines involved in inflammation. Mucosal-associated invariant T (MAIT) cells perform antibacterial functions. Systemic inflammation is reflected by

elevated plasma levels of acute-phase proteins, endothelial activation markers (ICAM1, VCAM1 or VEGF), and cytokines like TNF, IL-1β, IL-6, IFN-γ and IL-17[103,104].

Liver cirrhosis progresses from an asymptomatic compensated phase to an **acute decompensated (AD)** phase, characterized by complications such as ascites, gastrointestinal bleeding, encephalopathy, or a combination of these, which increase the risk of hospitalization and mortality. Around 30% of hospitalized patients with AD cirrhosis develop **acute-on-chronic liver failure (ACLF)**, presenting the most severe immune alterations and the poorest prognosis[105].

1.4.3.1. Acute-on-chronic liver failure in acute decompensated cirrhosis

Acute-on-chronic liver failure (ACLF) is a syndrome associated with patients suffering **acute decompensated (AD)** cirrhosis. It is characterized by high-grade systemic inflammation, including oxidative stress, and immune cell paralysis, often triggered by precipitant events (e.g., infections, alcoholic hepatitis or injured tissue). Altogether, these factors contribute to intrahepatic circulatory dysfunction, worsening portal hypertension and impaired liver tissue perfusion, leading to hepatic and/or extrahepatic organ failure and high short-term risk of mortality[106–108].

In the context of liver cirrhosis, infection and tissue injury promote the recognition of PAMPs and DAMPs by PRRs (e.g., TLRs and NLRs) on the surface of peripheral monocytes and DCs, activating intracellular signaling pathway that lead to massive pro-inflammatory cytokine production (e.g., IL-6, IL-8, IL-1β, IL-1ra, TNF-α, sCD163, C-reactive protein). IL-6 drives Th17 responses and, together with IL-8, induces neutrophil chemotaxis, activation, and phagocytosis. Recruited neutrophils release reactive oxygen species (ROS) and proteolytic enzymes, contributing to immunopathology and cell death, resulting in high-grade systemic inflammation. A dysregulated, excessive innate immune response, contributes to hepatocyte apoptosis and liver damage, creating a self-perpetuating cycle of inflammation (Figure 5a). Simultaneously, the immune response is counterbalanced by an excessive compensatory anti-inflammatory response, where IL-10 modulates NF-kβ activity, reducing the secretion of pro-inflammatory cytokines like TNFα, IL-1, IL-6, IL-8, and IL-12. This, combined with the exhaustion of immune effector cells, leads to acquired immunodeficiency. This condition is characterized by decreased HLA-DR expression in monocytes, which impairs antigen presentation; elevated PD-1 and TIM-3 expression in T cells, resulting in lymphocyte anergy and reduced INF-γ production; and neutrophils with low phagocytic capacity and high levels of CXCR1/2 receptors (Figure 5b). Consequently, these factors significantly increase the risk of infections in ACLF patients. Indeed, ACLF severity, prognosis, and risk of mortality correlate with the extent of the dysfunctional immune response[109–111].

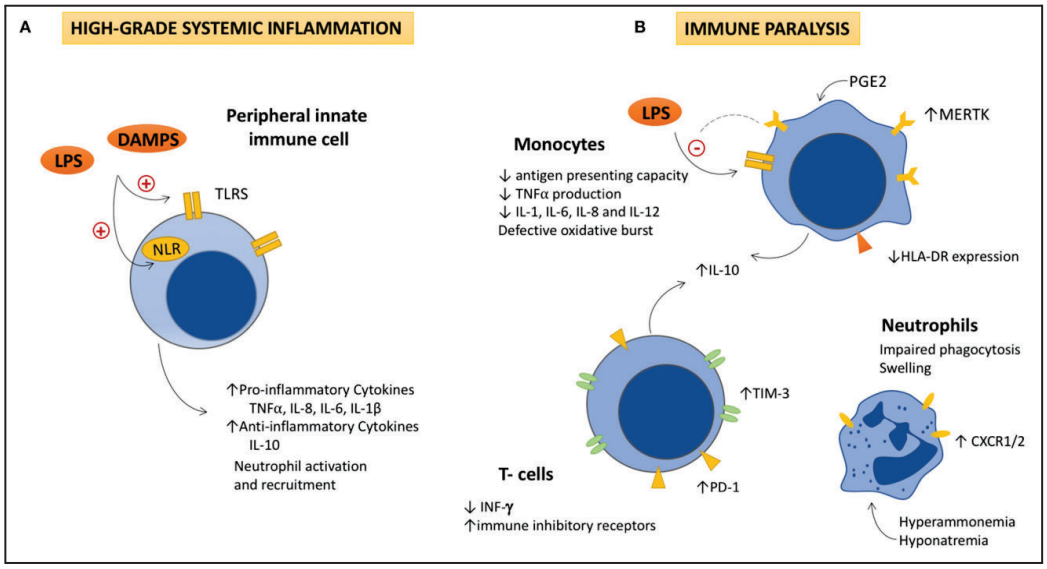


Figure 5. Overview of the dysfunctional immune response in ACLF. (A) Mechanisms of systemic inflammation. (B) Immune system cell paralysis. *Extracted from Martin-Mateos et al.*[109]

Although there is currently no effective treatment for patients with ACLF, disease management primarily focuses on addressing precipitating factors to prevent disease progression and mitigate associated complications. Liver transplantation remains the only definitive solution for patients with advanced ACLF, but it is associated with significant complications. To bridge the gap between supportive care and transplantation, albumin dialysis and plasma exchange have been proposed as potential therapeutic options[107,111]. Indeed, administration of human serum albumin (HSA) to patients with AD cirrhosis has demonstrated anti-inflammatory effects. Specifically, HSA is internalized by leukocytes, activating endosomal TLR3 and TLR4 signaling, leading to changes in the immune cell transcriptome, particularly in cytokine-related genes, without affecting phagocytosis, efferocytosis, or ROS production[112]. In parallel, other experimental therapies are being explored. For instance, IL-1 has been shown to reduce liver inflammation and neutrophil infiltration, while IL-22 exerts anti-apoptotic, proliferative, and antimicrobial effects, improving liver injury in both mouse models and ACLF patients, respectively[113,114]. Additionally, low-dosage beta-blockade in ACLF patients with portal hypertension has demonstrated anti-inflammatory benefits by reducing leukocyte counts and improving clinical outcomes. Other emerging therapies, such as granulocyte colony-stimulating factor and fecal microbiota transplantation, have also shown potential in promoting hepatic regeneration and restoring neutrophil function[107,115].

1.4.4. Infection

Infections represent a category of diseases in which the body's immune response is responsible for eliminating invading microorganisms and restoring tissue homeostasis. However, the nature and outcome of this response can vary depending on the pathogen involved (virus, bacteria, or fungi), the circumstances of infection, and the host's immune system's effectiveness.

Viral infections, for example, can range from acute, triggering robust and often widespread inflammation, such as influenza and COVID-19, to chronic involving long-term interactions between the pathogen and the host, such as those caused by hepatitis B virus and human immunodeficiency virus. Also, some viral infections can weaken the host's immune response, while others can trigger autoimmunity and cancer in genetically susceptible individuals. After entering host cells, viruses replicate at the site of infection, with some killing infected cells and releasing cellular contents like proteases and lysosomal enzymes, creating an inflammatory environment. Neutrophils are rapidly recruited, and alongside endosomal TLR activation, this leads to the production of pro-inflammatory cytokines, IFNs, cationic proteins, lipid mediators, and MMPs, activating the JAK-STAT pathways and triggering adaptive immunity. T cells, primarily Th1, Th17 and cytotoxic T cell subsets, either destroy virus-infected cells or release cytokines like TNF that cause tissue damage. Additionally, Th cells help B cells produce antibodies that form immune complexes, which may deposit in tissues and cause related diseases. If Tregs and inhibitory pathways fail to control Th and cytotoxic T cells, tissue damage can occur. Persisting viruses can also stimulate innate immune cells, such as DCs, NK cells and macrophages to produce anti-inflammatory molecules (IL-10 and TGF-β). If the anti-inflammatory responses outweigh pro-inflammatory ones, the virus may persist in the host[116–118].

- **Influenza**, also called the **flu**, is an infectious respiratory disease caused by influenza A and influenza B viruses. It primarily affects the upper respiratory tract with symptoms like fever and cough, but in some cases can lead to severe complications like pneumonia[119].

- **COVID-19** is an infectious respiratory disease caused by the SARS-CoV-2 virus. The virus targets the Angiotensin Converting Enzyme-2 (ACE2) receptor on host cells in the nasal cavity and lungs through the spike protein on its surface, leading to membrane fusion and infection. While it primarily affects the respiratory system, COVID-19 is also known to involve multiple organs, including the cardiovascular, immune, and endocrine systems, potentially causing not only short-term, but also long-term effects[120].

- **Hepatitis B virus (HBV)** interacts with the hepatic bile acid transporter NTCP via the large HBsAg-(L) protein, triggering viral internalization into hepatocytes. Once inside, HBV replicates and induces chronic liver inflammation, with intermittent episodes of acute inflammation. This leads to liver fibrosis, potentially progressing to cirrhosis, and in severe cases, decompensated liver disease or hepatocellular carcinoma. Current therapies include antiviral agents targeting viral replication and immunomodulators, such as IFN therapy, to manage infection and prevent progression to severe liver disease[121].

- **Human immunodeficiency virus (HIV)** invades host cells by attaching with its envelope glycoprotein (gp120/42) to CD4 receptors on T lymphocytes, using CCR5 or CXCR4 as co-receptors. As HIV replicates, it destroys the host cell, leading to immune system

activation and depletion of CD4+ T cells. HIV causes **acquired immunodeficiency syndrome (AIDS)** and subsequent susceptibility to opportunistic infections and malignancies. However, HIV is now considered a manageable chronic disease[122].

1.4.5. Solid tumors

Solid tumors are abnormal masses of tissue that arise from the uncontrolled growth of cells, and their development is often closely linked with chronic inflammation. Solid tumors can be broadly classified based on their tissue of origin (histological type) into four groups[123]:

- **Carcinomas:** Malignant neoplasms of epithelial origin, arising from either the internal or external lining of the body, including organs like the breasts, lungs, gastrointestinal tract, and skin. Examples of carcinomas considered in the **Publication 2** of this thesis are Breast cancer (BRCA), Colorectal Cancer (CRC), Nasopharyngeal carcinoma (NPC) and Head and Neck squamous cell carcinoma (HNSCC).
- **Sarcomas**: Tumors that develop in supportive and connective tissues such as bones, tendons, cartilage, muscle, fat, and blood vessels.
- **Lymphomas**: Solid cancers originating in the lymphatic system, which consists of a network of vessels, nodes, and organs such as the spleen, tonsils, and thymus.
- **Mixed types:** Neoplasms that exhibit characteristics of more than one category.

Tumors are characterized by their ability to invade surrounding tissues and potentially spread to other parts of the body, leading to metastasis. Tumor initiation and progression are significantly influenced by the tumor microenvironment (TME), the presence of chronic inflammation, which can provide a pro-inflammatory context, and the tumor's capacity to promote immune evasion[124–126]. Understanding the mechanisms driving each cancer type, as well as the unique characteristics of each patient, is crucial for developing personalized treatments to prevent disease progression and, ultimately, achieve remission.

1.5. Using inflammation as a diagnostic tool

Immune-driven diseases encompass a wide range of conditions where the immune system plays a central role in disease pathogenesis, often due to imbalances in inflammatory responses, whether excessive or insufficient[62]. Understanding the cellular and molecular mechanisms underlying these inflammatory processes, as well as understanding of the interplay between the innate and adaptive immune systems and potential dysregulation, is essential for advancing our knowledge of these diseases. Integrating these insights with clinical data can help researchers identify specific inflammatory profiles and develop personalized, effective therapies, ultimately improving patient outcomes and quality of life.

Assessing inflammation and immune status has traditionally relied on invasive tissue biopsies, which are often not feasible or practical for continuous monitoring. A less invasive alternative is the use of liquid biopsies, such as blood samples, which offer a dynamic and accessible way of tracking immune responses and inflammation[127,128]. Liquid biopsies are particularly suited for large-scale studies across multiple diseases, as they provide a comprehensive snapshot of immune status across various organs. They also offer valuable insights into immune cell composition, cell activity, cytokine levels, and circulating biomarkers, aiding in the early detection, monitoring, and management of inflammatory and immune-mediated diseases[129,130].

Additionally, liquid biopsies can also be used to investigate therapeutic approaches and monitor the immune system's response to treatments, either during clinical trials or *in vitro*, such as in patients-derived blood cell culture. They are particularly useful for understanding how therapies impact immune regulation at the molecular and cellular levels, highlighting drug effects on specific immune cell subsets and changes in transcriptional profiles[128,131–133].

These applications underscore the transformative potential of liquid biopsies in both research and clinical practice, to refine therapeutic strategies for more targeted and effective interventions, enabling a more personalized approach to managing immune-driven diseases.

2. Single-cell technology

In the past decade, we have witnessed a technological revolution that has transformed our ability to study the immune system[134]. Historically, biopsies were analyzed using bulk genomics techniques to characterize the molecular profile of all cell populations within the sample. However, with the advent of single-cell technologies, we are now able to examine the profiles of individual cells[135]. This advancement has greatly improved our understanding of sample heterogeneity, enabled the discovery of new immune cell types, and allowed us to capture cellular state dynamics across different physiological and pathological contexts[136].

2.1. Origin and evolution

In 2009, Tang *et al.* made a significant breakthrough in the field of transcriptomics by being the firsts to achieved whole-transcriptome profiling (mRNA-Seq) from a single mouse blastomere[137]. This advancement allowed for increased gene capture and the discovery of previously unknown transcript isoforms. While further adaptations were necessary to apply this technique to larger numbers of cells, tissues and other species[138], this work marks the beginning of a new era.

Over the last 15 years, many techniques have been developed to increase scalability (increasing from tens to thousands of cells profiled at a time) and sensitivity (capturing the cellular transcriptome) (Figure 6)[139,140]. In most cases, these protocols follow the same general steps: starting from a single-cell suspension, cells are captured and lysed, followed by reverse transcription of the mRNA into cDNA, amplification, and library preparation for sequencing. The isolation and capture of single cells can be achieved using plate-based methods (e.g., SMART-Seq[141], SMART-Seq2[142]), where each well captures a single cell, or droplet-based methods (e.g., inDrop[143], Drop-Seq[144], 10X Genomics[145]), where cells are encapsulated within droplets. While plate-based methods are known for their high sensitivity in gene detection and full-length transcript sequence, allowing for differential isoform capturing and variant calling, they have a lower cell throughput. In contrast, droplet-based methods offer high-throughput profiling of thousands of cells, which is useful for identifying populations or discovering rare cell types, albeit with reduced sensitivity and skewed transcript coverage[139]. Both technologies have addressed some of their limitations. Plate-based methods, for example, have been enhanced by combining multiple rounds of barcoding, making them scalable from hundreds to thousands of cells and more affordable per plate, as seen with methods like sciRNA or SPLIT-seq[146]. Droplet-based microfluidics approaches have also been paired with long-read sequencing technologies, such as Nanopore long-read sequencing, to obtain full-length transcript information (scNaUmi-seq[147]). Since 10X Genomics Chromium technologies emerged in 2017 as a single-cell droplet-based technology, it has dominated the market, However, new competitors have emerged, covering specific applications (e.g. DNA or protein profiling), like Tapestri Single-Cell from Mission Bio.

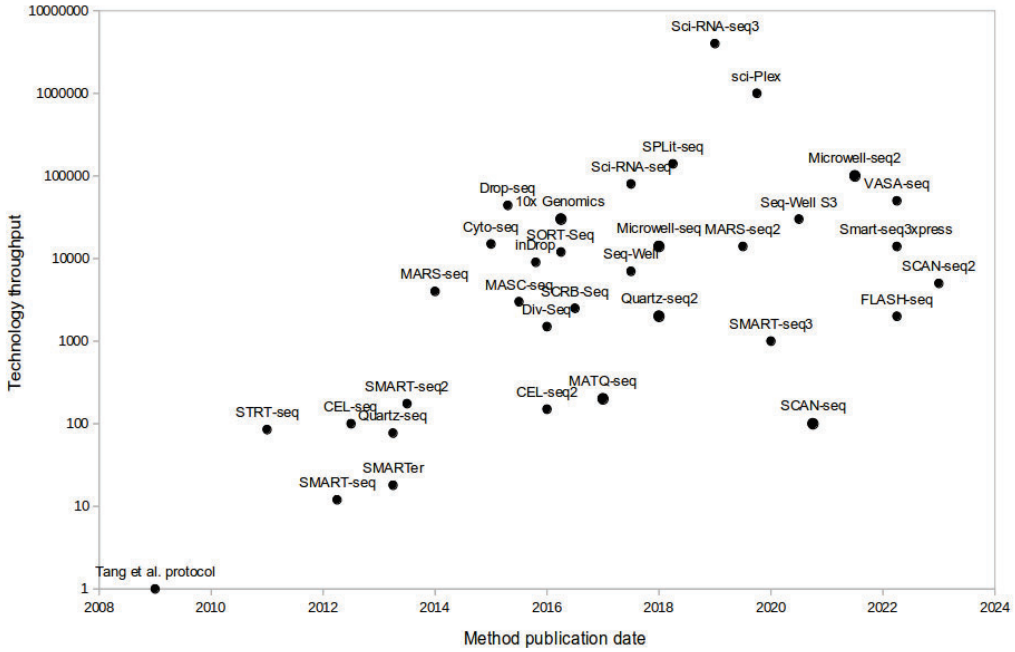


Figure 6. Timeline of scRNA-seq methods. Scatterplot shows the method publication date (years) and throughput of sequencing for each technology (number of cells per experiment run). The plot includes different plate-based and droplet-based technologies, as well as tag-based and full-length methods. *Adapted from Huang et al. and Svensson et al.*[139,148] .

Recently, several methods have been developed to simultaneously profile multiple layers of information from the same cells[149], including transcriptomics (scRNA-seq), chromatin accessibility (scATAC-seq)[150], surface proteins (scCITE-seq)[151], immune profile (scTCR/BCR-seq), and spatial location. Like scRNA-seq methods, these techniques have evolved over the past few years, enhancing their capabilities. These approaches offer a more comprehensive view and a deeper understanding of cellular behavior in complex scenarios.

With new protocols for single-cell profiling rapidly emerging, it is essential to systematically compare different methods to understand their strengths and weaknesses[152], and to choose the appropriate method depending on the research question. Among the various available methods for single-cell analysis, the droplet-based approach provided by 10X Genomics was selected as the technology for generating all the scRNA-seq data analyzed in this thesis. This choice was made due to its flexibility, high throughput, widespread within the research community, commercial availability, and strong client support. Accordingly, a comprehensive description of the technique, including its strengths and limitations, will be provided.

2.2. Experimental design

The success of any research project depends not only on the quality of the data, but also on its experimental design, which lays the groundwork for extracting the most value from the collected data. A well-crafted design involves carefully selecting the sample or biopsy, considering its collection and preservation, determining the most suitable single-cell modalities and protocols to

answer the scientific question, and deciding on processing methods, such as multiplexing[153]. By meticulously planning these aspects, technical variability and batch effects can be significantly reduced, leading to more robust downstream computational analysis.

2.2.1. Single-cell RNA-sequencing (scRNA-seq)

Single-cell RNA-sequencing (scRNA-seq) technology enables the analysis of transcriptomes on a cell-by-cell basis using scalable microfluidic platforms. 10X Genomics provides several scRNA-seq 3' assays specifically designed to capture the transcriptomic profiles. For example, the standard 3' v3.1 assay allows profiling 500 to 10,000 individual cells in a single run, while the high-throughput (HT) 3' v3.1 assay scales this capacity to 2,000 to 20,000 cells[145].

The process begins with a single-cell suspension, which is combined with a master mix of reagents (lysis buffer) and gel beads coated with barcoded oligonucleotides, including a 10X barcode that marks each RNA molecule's cell of origin and a Unique Molecular Identifier (UMI) that gives each transcript a unique 12-base sequence fingerprint, a poly(dT) sequence for capturing mRNA, and adapter sequences for downstream next-generation sequencing (Figure 7a). These components, along with oil, form simultaneous reaction vesicles known as Gel Beads-in-Emulsion (GEMs) when loaded into a channel of the *Chromium Next GEM Chip* (Figure 7b). After Poisson-based distribution loading, around 10% of functional GEM contain a single cell, a unique gel bead, and the necessary reagents. Following GEM generation, a single cell within each droplet is lysed, releasing its content. The Gel Bead dissolves, releasing barcoded oligonucleotides and reagents into the solution, initiating the reverse transcription (RT) of polyadenylated (polyA) mRNA to generate cDNA. At this point, all cDNAs from a single GEM share the same barcode. The droplets are then lysed, and the barcoded cDNA is purified and amplified via PCR (Figure 7c,d). After cDNA quantification, libraries are prepared by fragmenting, size-selecting, and indexing (adding Illumina P5 and P7 amplification primers). Finally, the libraries are sequenced using Illumina sequencing instruments, producing approximately 20,000 paired-end reads per cell that include the 10X GEM cell barcode, the UMI, the polyA fragment, and a 90 base pair sequence from the 3' end of the captured mRNA (Figure 7d).

Importantly, each gel bead in the system not only has a unique 10X barcode but also contains distinct UMIs. UMIs help to mitigate and correct amplification biases after sequencing by grouping amplicons with the same UMI into a single count. This process addressed biases introduced during PCR amplification, which can arise from factors such as GC content, transcript length, and primer affinity. Because cDNA synthesis is inefficient and many low-abundant mRNAs are not RT, this amplification can distort the data, impacting the biological interpretation[154]. Thus, UMIs are now a standard component in almost all sequencing protocols due to their effectiveness in improving data accuracy and minimizing downstream biases.

Although only 3' scRNA-seq assays were described, they share a common protocol with most of the other assays offered by 10X Genomics, with only minor adaptations. These assays can be used to study not only cell transcriptomic profiles but also immune profiling (T- and/or B-cell receptors), chromatin accessibility, and cell surface proteins. Additionally, they provide support for capturing nuclear transcriptomics, adapting to different sample types (fresh, frozen, FFPE), and enabling sample multiplexing using barcoded oligonucleotides[145].

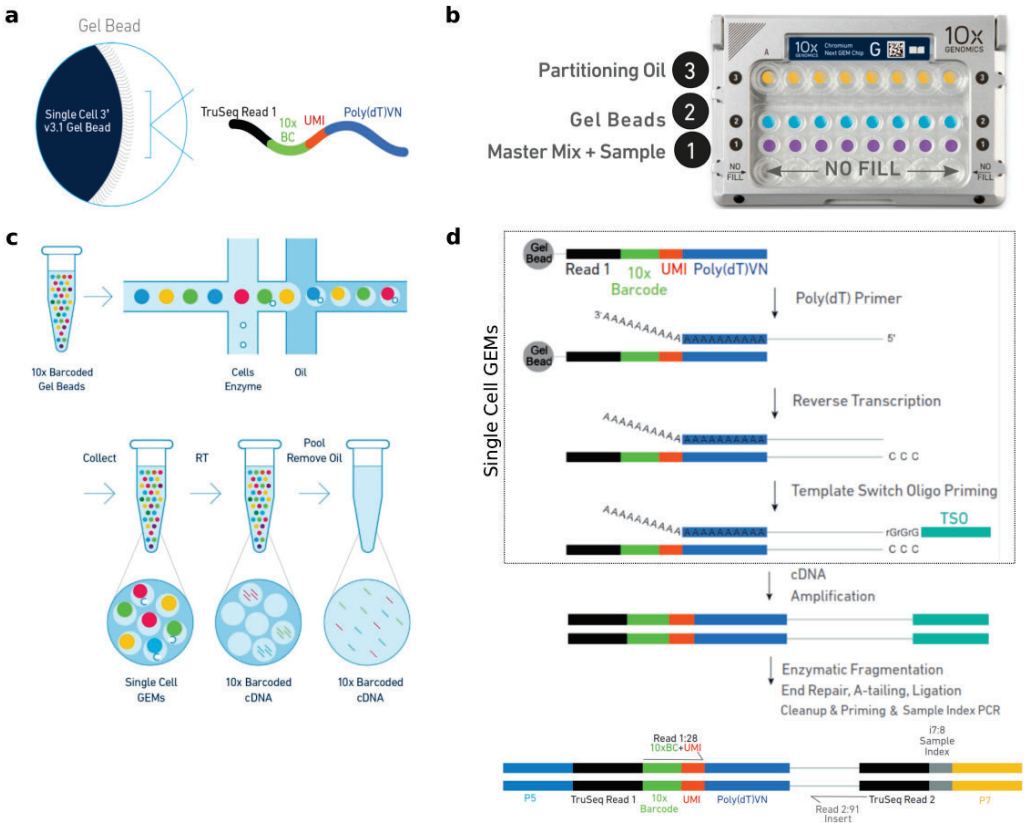


Figure 7. scRNA-seq 10X Genomics technology. (**a**) Single Cell Gel Bead (3' GEX v3.1). (**b**) Chromium Next GEM microfluidics chip. (**c**) Droplet-based cell encapsulation scheme. (**d**) Chromium Single Cell 3' Gene Expression V3.1 library preparation for sequencing with an Illumina sequencer. *Adapted from 10X Genomics*[145].

2.2.2. Sample input

Sample collection, preservation, transportation, and preparation are crucial steps to avoid compromising the quality of scRNA-seq data.

Sample collection and storage

When an experimental design is planned in advance, samples can be prospectively collected, allowing for careful and well-designed collection and preservation according to the research objectives. In contrast, retrospective studies using biobank samples depend on pre-existing biopsies (e.g., cryopreserved or paraffin-embedded samples), which may limit the ability to control for such biases. Additionally, although samples can be collected and processed freshly at

hospitals or research centers, in some research projects involving multi-center sample collection, study designs require disconnecting the time and site of sampling from downstream processing steps. In such cases, a tailored experimental design and accurate sample handling involving preservation methods is crucial to minimize technical alterations (e.g., cell clumping or lysing, or altered cellular transcriptomes) and ensuring the detection of subtle biological variability.

Cell suspension preparation

Single-cell RNA sequencing requires a suspension of viable single cells as input, making it crucial to minimize delays in chip loading, reduce the presence of cellular aggregates, dead cells (ensuring cell viability above 70%), non-cellular nucleic acids, and biochemical inhibitors. To maximize the number of GEMs containing a single cell and maintain a low multiplet rate (the fraction of two or more cells associated to a unique GEM), cell suspensions should be loaded at a limiting dilution. Given that cell encapsulation has about 65% efficiency, it is important to start with enough cells to achieve the desired target cell recovery[145]. While this process is relatively straightforward for liquid biopsies, it becomes more challenging with solid tissue biopsies, as dissociation steps are required.

Delicate tissues and sensitive cell types present particular challenges in single-cell analysis. For instance, granulocytes have relatively low RNA content and high levels of RNases, resulting in fewer detected transcripts in GEMs. Neutrophils, in particular, are prone to rapid degradation post-collection and cannot be isolated effectively[155], unlike other blood cells such as Peripheral Blood Mononuclear Cells (PBMCs). Additionally, epithelial and stromal cells are less resistant to sample processing and tissue dissociation[156], leading to changes in cell composition in the downstream analyses compared to their tissue of origin.

Technical alterations during sample handling

The success of single-cell experiments relies on effective sample handling. Stress-related factors such as collection, transport, storage, and downstream processing can alter the natural state of cells, influencing the outcome. Other critical factors include the skills of lab personnel, library preparation workflows, and sequencing technologies, all of which can lead to technical differences[153]. Fortunately, the single-cell research community has made significant efforts to address some of these limitations[157]. Cryopreservation has emerged as a method to manage sample logistics in decentralized analysis, preserving the transcriptional profile of cells in suspension and some tissues, such as PBMCs[158] or bone marrow aspirates[159]. However, cryopreservation introduces two challenges: the time before cryopreservation can induce stress-related effects, and it can reduce the viability of sensitive cells, leading to considerable changes in sample composition[158]. Alternatively, other preservation methods, such as alcohol-based or cross-linking fixatives, have been proposed, but are not compatible with

single-cell assays[160,161]. Additionally, downstream processing steps involving mechanical or enzymatic dissociation cause cellular stress, modifying the transcriptional activity and promoting the expression of heat shock and stress-response genes[156,162]. In this scenario, we optimized a protocol for early sample preservation to decouple sampling from downstream processes, minimizing stress-related artifacts and making it suitable for single-cell technologies. This methodology is presented in **Publication 1** of this thesis.

Multiplets

Multiplets occur when two or more cells aggregate into a single droplet (doublets or multiplets) during the cell encapsulation step in scRNA-seq protocols. This aggregation can result from incomplete cell dissociation during experimental processing or from cell overloading of the experiment. Since the presence of multiplets can lead to false discoveries by complicating the differentiation between biological signals and technical artifacts, it is important to understand their causes, impact, and, most importantly, how to mitigate them.

The multiplet rate follows a Poisson distribution, depending on the scRNA-seq assay (number of GEM barcodes) and the targeted cell recovery. According to 10X Genomics, the multiple rate is approximately 0.8% for every 1,000 cells targeted in the standard 3' v3.1 assay or 0.4% for the HT 3' v3.1 assay. The probability of multiplets in the HT assay is double that of the standard assay, as the HT assay has roughly twice the number of GEMs[145]. Understanding the multiplet rate of each assay, as will be discussed in the following section, improves experimental design.

Sample multiplexing

While several *in-silico* batch correction algorithms have been developed for single-cell data, it is always preferable to design experiments that minimize batch differences from the start. Since 2017, a variety of sample multiplexing techniques for scRNA-seq experiments have been introduced as effective strategies to address batch effects and capture subtle differences across conditions[163], along with their corresponding computational methods -a topic that will be discussed later. Additionally, sample multiplexing offers advantages such as increased sample throughput in a single GEM well (by overloading commercial droplet-based assays), easier identification of multiplets, reduced experimental costs, and the ability to handle large-scale sample projects more efficiently.

Sample multiplexing involves pooling together cells from different samples, and running a single-cell experiment, where all cells are processed together in a single GEM well. Following cell encapsulation, library preparation, and sequencing, this approach allows for the assignment of each cell to its original sample and the identification of droplets that initially contained either a single cell (singlet) or multiple cells (multiplets). The main methods for sample multiplexing are barcode-based and those using genetic polymorphisms (e.g., SNPs).

- In **barcode-based multiplexing**, as shown in Figure 8, samples are labeled either with unique DNA barcodes that targets the cell (or nuclei) surface proteins, a method known as cell hashing (HTO)[164], or with lipid/cholesterol-modified oligonucleotide (LMO/CMO) that tag the cell membrane, a technique known as CellPlex[145].

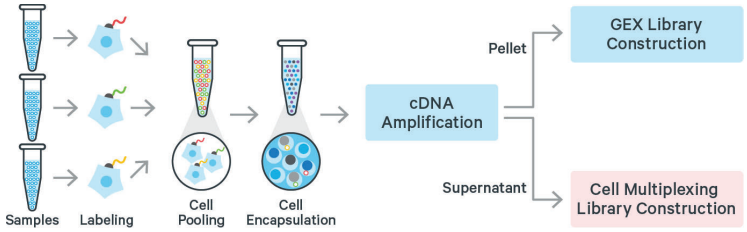


Figure 8. Schema of 10X Genomics scRNA-seq experiment with barcode-based sample multiplexing. *Extracted from 10X Genomics*[145].

- In **SNP-based multiplexing**, multiplexed samples are distinguished based on their natural genetic landscapes. Since this approach requires samples with distinct genetic signatures (e.g. SNPs), it cannot be used to multiplex samples obtained from the same individuals at different time points, developmental stages, or experimental conditions.

Despite the benefits of sample multiplexing, it is crucial to balance the reduced library preparation costs from overloading GEM channels with increased library sequencing costs due to reads lost for sequenced multiplets. For example, increasing from 10,000 cells in a standard 3' assay to 30,000 cells would result in an increase in multiplets from 8% to 24%. To aid in this decision, the researchers Hafemeister and Satija developed the *Cost per cell* tool [https://satijalab.org/costpercell/], which helps to calculate costs and determine the most effective strategy.

2.3. Raw data pre-processing

As outlined in the previous experimental design section, the 10X Genomics libraries generated must be sequenced, providing sequencing reads as input raw data for the pre-processing steps to obtain a gene-by-cell count matrix and determine the sample-of-origin cell identity.

2.3.1. From sequencing reads to count matrices

Cell Ranger[165] is a command-line tool designed to process 10X Genomics single-cell data, offering a suite of analysis pipelines for barcode processing, read alignment, single cell 3' gene counting, and feature barcode analysis.

- The CellRanger *mkfastq* pipeline demultiplexes raw base call (BCL) files generated by unique runs in Illumina sequencers into different FASTQ files, utilizing a wrapper around Illumina's *bcl2fastq* module.

- The CellRanger *multi* pipeline processes FASTQ files from a combination of libraries (e.g. GEX, or CMO) generated from a single GEM well, performing alignment, filtering, barcode counting, and UMI counting. When only GEX libraries are generated, the CellRanger *count* pipeline can be used as an alternative.

  GEX reads are first trimmed to remove non-essential Template Switch Oligo (TSO) nucleotides from the 5' end and the polyA tail from the 3' end. The cleaned reads are then aligned to the reference genome of interest using the splicing-aware STAR aligner[166], differentiating between exonic, intronic or intergenic regions. Exonic reads are further mapped to annotate transcripts. Subsequently, reads are grouped by GEM-barcodes, UMI, and gene annotation. Following this, two algorithms *OrdMag* and *EmptyDrops*, are used to evaluate droplets with high and low RNA content, distinguishing between background and cellular RNA containing UMIs, which allows for the identification of cell-containing droplets from empty droplets[145]. This step is crucial for quantifying the mRNA detected for each gene in each cell, resulting in a count matrix with genes in the rows and barcodes (cell-containing droplets) in the columns, serving as the starting point for the data processing. CMO reads enable the classification of cell-containing droplets as singlets, multiplets, or blanks (with low tag counts). Only GEM-barcodes identified as singlets are retained in the count matrix[167].

- The cellRanger *aggr* pipeline processes outputs from multiple runs, normalizing them to same sequencing depth and recomputing feature-barcode matrices for combined data.

2.3.2. Sample demultiplexing: sample-of-origin cell identity

Given that sample multiplexing was employed for data generation, it is necessary to discuss the tools used for sample demultiplexing, which involves separating pooled samples in a single library back to their original sources. While several alternative algorithms are available, only the methods applied to the data associated with this thesis will be detailed.

Barcode-based demultiplexing

- **Cell hashing (HTO) libraries** were demultiplexed using the *HTODemux* function[164] from Seurat. Briefly, it performs *k-medoids* clustering on the centered log ratio (CLR) normalized HTO values to separate cells into *K (number of samples) + 1* clusters. It then calculates a negative distribution for each HTO and uses the cluster with the lowest average value as the negative group. For each HTO, a negative binomial distribution is fitted to the negative cluster, with the 0.99 quantile serving as the threshold for classifying a GEM-barcoded cell as positive or negative. Finally, GEM-barcoded cells positive for a single HTO are classified as singlets, those positive for multiple HTOs are annotated as multiplets, and those negative for all HTOs are classified as negative.

- **CMO cell membrane tag libraries** were demultiplexed using the CellRanger *multi* pipeline provided by 10X Genomics. Starting from the GEM-barcoded cells, CMO tag counts are modeled using a normal distribution for each CMO, with the signal-to-noise mean ratio and variance being estimated. For each GEM-barcode, the likelihood of each state -single CMO, multiplets, or background noise- is calculated. When the probability of a particular state exceeds 0.9 (default), it is assigned to the corresponding GEM-barcode; otherwise, it remains unassigned[145].

SNP-based demultiplexing

SNP-based demultiplexing involves two steps: first, requires piling up (or genotyping) a list of common variants for each cell, followed by donor deconvolution (Figure 9). Given the experimental design used in the laboratory and the need to standardize the demultiplexing process, *Vireo*[168], in combination with *cellSNP-lite*[169], was employed in this thesis.

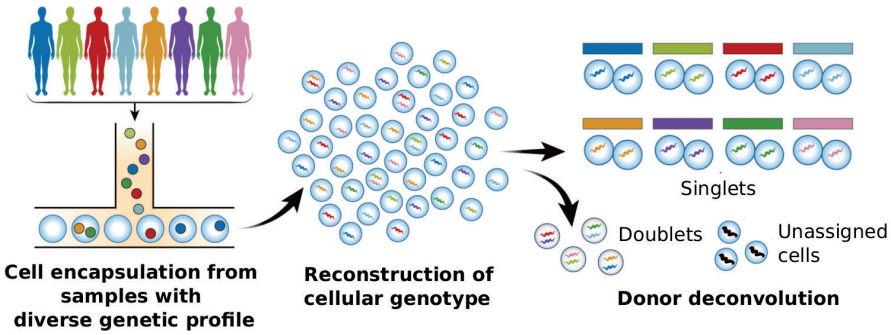


Figure 9. SNP-based multiplexing and demultiplexing pipeline. Individuals with diverse genetic profiles are multiplexed and loaded onto a droplet-based scRNA-seq instrument. Cellular genotypes are then computationally reconstructed, grouping cells based on their genotype, which enables donor deconvolution (demultiplexing). This process identifies singlets, doublets, and unassigned cells. *Adapted from Ye Lab [https://yelab.org/].*

Starting from aligned sequenced reads (BAM files), a list of human common SNPs identified by the 1000 Genome Project [https://sourceforge.net/projects/cellsnp/files/SNPlist/] is used by *cellSNP-lite*[169] to reconstruct the genotype of each cell. This genotypic information is subsequently used by *Vireo*[168], which employs a Bayesian approach to demultiplexed pooled cells into groups based on the presence or absence of expressed alleles. When donor genotype references are available, these cell pools can be assigned back to their corresponding donors.

Although SNP-based sample demultiplexing can be performed with no, partial, or complete genotyping data, genotyping data is essential for accurately mapping cells back to their original individuals. However, combining barcode and SNP-based multiplexing approaches can lead to an enhanced sample demultiplexing. This strategy has been conceptualized, developed, and applied in **Publication 2** of this thesis.

2.4. scRNA-seq data processing

Fifteen years after the introduction of the first single-cell methods and the generation of millions of datasets, the field has achieved considerable maturity, and numerous bioinformatics solutions have been proposed to address the various challenges associated with this technology[170]. In this context, a wealth of publications, tutorials and guides are constantly updated to benchmark these tools and establish the basis for accurate scRNA-seq data analysis[171–176].

After obtaining the count matrices, the data is ready to be processed and explored. Importantly, multiple frameworks have been developed to facilitate the analysis of single-cell data, with the most popular being *Seurat*[177] (R-based) and *Scanpy*[178] (Python-based). While these frameworks differ in their programming languages, data structures, and task-specific analysis tools, the community has ensured good interoperability between them, allowing for flexible and case-driven single-cell analysis. Interestingly, when I began my thesis back in 2020, I started by analyzing single-cell data with the R framework. However, due to the exponential growth of single-cell data, later in this thesis I shifted to the Python-based to address scalability issues and to easily integrate machine learning algorithms.

2.4.1. Data quality control

Data quality control (QC) is a crucial step in scRNA-seq analysis. While the data processing begins with an initial QC, it is essential to continuously apply QC throughout the analysis to ensure high-quality data and prevent technical artifacts from masking biological variability.

Poor-quality libraries and cells

Poor sample quality or experimental failures in the laboratory often result in low-quality libraries. Therefore, when a low number of cells and/or an unreliable transcriptomic profile are obtained, the best approach is to discard these libraries to prevent misleading results. However, sometimes not all of the library may be affected, so a more detailed and tailored QC is performed to remove poor-quality cells or genes with low expression.

Cells of interest are those of high-quality, characterized by a sufficiently rich transcriptomic profile to reveal both their biological identity and sample-specific characteristics. Conversely, poor-quality cells may include empty droplets -having evaded detection by CellRanger's *EmptyDrops*-, cells that are damaged or dying, as well as multiplets. To evaluate cell quality, three key covariates are typically computed and analyzed together per droplet: the number of UMI counts (count depth), the number of genes detected (transcriptomic complexity), and percentage of counts attributed to mitochondrial genes[173]. For instance, droplets with low count depth combined with low transcriptomic complexity can indicate empty droplets that might have mistakenly captured ambient RNA, or they may represent neutrophils[179]. Additionally, if these

droplets also exhibit a high percentage of mitochondrial mRNAs, it could suggest that they have encapsulated damaged or dead cells. Cells under stress often experience compromised cell membranes, which facilitates the leakage of both cytosolic and mitochondrial mRNAs, the latter of which is rarely released[173]. In this context, a droplet is classified as poor-quality based on specific thresholds derived from the distributions of the three covariates, but thresholds should be tailored to each sample and the biological context being studied. It is advisable to be relatively permissive in the initial stages to avoid filtering out viable cells or rare populations.

Multiplet detection and removal

Multiplets are an inherent consequence of single-cell assays. In experiments with a single sample, multiplets (i.e., droplets containing mixed transcriptomes) can result from either cells with dissimilar gene expression (heterotypic multiplets) or cells with similar transcriptomes (homotypic multiplets). In multiplexed experiments, multiplets can originate from multiple samples (e.g. donors) or a single one. While the multiplet expected rate can be predicted, detecting and removing them remains a challenging task, depending on the experimental strategy employed.

During sample demultiplexing, bioinformatics tools can flag and remove multiplets generated from different samples of origin thanks to the barcodes or genetic information, whereas multiplets from the same sample may remain uncaptured. To address this, several methods[180] such as *Scrublet*[181] or *DoubletFinder*[182] have been developed to evaluate whether the transcriptome of a given droplet resembles a mixture of two other random droplets from the dataset and assign a corresponding confidence score. Since these methods rely on estimating the doublet probability for each droplet, it is advisable to flag the cells rather than remove them at this stage. Additionally, QC covariates can be used to identify potential multiplets. When multiple cells are encapsulated in a single droplet, they can result in higher count depth and increased transcriptomic complexity. Therefore, detecting abnormally high values in the distributions can suggest the presence of a multiplet. A particular case is heterotypic multiplets, which can be easily identified during downstream analysis, either after clustering cells or by performing a differential expression analysis (DEA). This is because lineage markers (e.g., *CD3D* for T cells, *MS4A1* for B cells, and *CD14* for myeloid cells) are typically unique to each cell type. Therefore, the co-expression of these markers within a droplet indicates the presence of multiple cells.

Heterotypic multiplets can introduce misleading heterogeneity that may obscure the true biological variation within the sample, whereas homotypic doublets are not necessarily identifiable and considered innocuous. Upon confident identification of multiplets, it is advisable to remove them from the dataset and reprocess the dataset to ensure a more accurate representation of the single-cell landscape.

2.4.2. Normalization

In single cell analysis, differences in the total number of counts across cells present a significant challenge. These variations can arise from biological factors, such as differences in cell size or activity status, or from technical factors, like variations in transcriptomic capture or other experimental processes[183]. While several methods exist for normalizing transcriptomic data, they have needed to be adapted to address the particularities of single-cell data, which is sparse and heterogeneous.

A common normalization approach is library-size log normalization. In this method, following the count-per-million (CPM) normalization concept, the library size of each cell is calculated, and the absolute mRNA counts of its genes are divided by this value to obtain relative mRNA frequencies, which are then multiplied by a scaling factor (usually 10,000). The data is subsequently *log(x+1)*-transformed to address certain limitations: this reduces the mean-variance relationship, mitigates data skewness by fitting it into a normal distribution assumed by many downstream methods, and facilitates the computation of differences in gene expression as log-fold changes (where the quotient of logarithms equals the difference in logarithms)[173,184].

While log-transformation is widely used in scRNA-seq, it has limitations, particularly due to the sparsity typical in this kind of data (dropout events). Moreover, this method may be inadequate for computing differentially expressed genes, as normalizing by library size alone might not fully account for compositional biases, potentially leading to misleading differences between cells. One option to address this issue is to remove genes expressed in fewer than ~10-20 cells in the dataset before running normalization. Due to these challenges, more sophisticated methods such as *SCRAN*[183] or *SCTransform*[185] have been developed, although they can suffer from scalability problems. Despite multiple efforts to develop and benchmark new methods[186], no single method has emerged as an absolute winner. The choice of method should be tailored to the specific dataset or project and its research goals to ensure the findings reflect the biological variability.

2.4.3. Dimensionality reduction

Single-cell data captures mRNA from around 30,000 genes, and even though it is standard practice to exclude genes that are not expressed in a certain number of cells during data normalization, the data remains quite sparse. Specifically, many of the remaining genes contain mostly zero counts across cells, may not be informative, or are expressed uniformly across all cells; these genes might not represent meaningful biological entities or variation across samples, as the “curse of dimensionality” concept[187] suggests. Therefore, it is necessary to reduce the dimensions of the normalized matrix by excluding uninformative genes while preserving relevant biological information and decreasing the demand on computational resources for large datasets.

Feature selection

The first step in dimensionality reduction is feature selection, which involves choosing highly informative genes based on their variability (highly variable genes, HVGs) or their average expression (highly expressed genes, HEGs). However, due to the mean-variance relationship, genes with higher expression levels als tend to have higher variance. To avoid bias in HVG selection, genes are grouped into bins according to their mean expression. Within each bin, genes are ranked by variance, and the top-ranked ones are selected[49]. Typically, 1,000 to 5,000 HVG are selected, though this number is highly dependent on the dataset of the study[173]. Finally, to equalize the variance among all HVG, *Z-score* scaling is applied to the data.

When batch effects are present in a dataset, since variable genes across the entire dataset might capture batch effects rather than true biological signals or specific cell types unique to certain samples, it is crucial to select HVGs in a way that accounts for these effects. A batch-aware approach involves identifying HVGs separately within each batch and then combining these lists by selecting genes that are consistently highly variable across the majority of batches.

Linear dimensionality reduction: PCA

According to the “curse of dimensionality” concept, dimensionality reduction algorithms are required to further eliminate inherent data redundancy due to the nature of cellular regulation, where multiple genes are co-regulated, and signaling pathways involve many genes[189]. This enhances our ability to find relevant information to distinguish between cell populations.

A widely used pre-processing method for dimensionality reduction is principal component analysis (PCA). Starting from scaled HVGs, PCA uses an orthogonal transformation to combine correlated groups of genes from the original dataset into principal components (PCs), which are then ranked in decreasing order of variance, with the first PC having largest variance[189]. The appropriate number of PCs for downstream analysis, typically between 20 and 30, can be determined, for example, by identifying the inflection point (“elbow” heuristics) where the variance explained by each PC stops decreasing significantly, meaning that adding more PCs does not increase the explained variability in the dataset[173]. PCA is highly interpretable because it provides the gene loadings associated with each PC, representing the contribution of each gene to the variance explained by the PC, and it is computationally efficient[190]. However, it lacks the capability to capture and visualize the complex structure of scRNA-seq data, for which non-linear dimensionality reduction algorithms are a better option[173].

Non-linear dimensionality reduction: Data visualization

Non-linear dimensionality reduction methods aim to project high dimensional data into 2D or 3D spaces, enabling data visualization. In these methods, while global distances may not be

meaningful, local distances are preserved; cells with similar profiles in the high-dimensional space remain closely embedded. The most commonly used methods are t-distributed stochastic neighbor embedding (t-SNE)[191] and Uniform Manifold Approximation and Projection (UMAP)[192].t-SNEs and UMAPs allow cells embedded in 2D space to be colored based on any covariate, providing valuable insights into which variables drive transcriptomic variability between cells. However, these embeddings should be used with caution, avoiding over-interpretation or inferring direct relationships between clusters. Although t-SNE and UMAP facilitate the clustering process, it is crucial to understand both its potential and limitations[173,193].

2.5. Identifying cellular structures

To this point, cells have been embedded and visualized based on their underlying properties. The next step after PCA is to identify the cellular structure within scRNA-seq data, meaning finding cell identities that resemble known cell types or states. For this purpose, the use of graph representation methods enable encoding the relationships between cells and their underlying data topology. Based on these graphs, unsupervised community detection algorithms group similar cells into clusters. Finally, the annotation step aims to infer the cell type or state identity for each cluster.

2.5.1. KNN graph representation

The K-Nearest Neighbors (KNN) algorithm calculates a distance value for all node pairs in a graph, establishing new relationships between each node and its K nearest neighbors. In the context of scRNA-seq data, each node of the graph represents a cell, connected to its K most similar cells in a lower-dimensional space, where K typically ranges from 2 to 100[193]. While several metrics can be used to measure the similarity between cells in the lower-dimensional space, Euclidean distance is commonly computed between PCs for all cells[173].

KNN graphs are the preferred method and a starting point for many downstream processes, such as unsupervised clustering or non-linear dimensionality reduction techniques like t-SNE or UMAP, due to three key advantages. First, KNN is highly scalable, making it suitable for large datasets, as it only requires neighborhood queries for each cell. Second, KNN provides a non-linear representation of scRNA-seq data, enabling the detection of complex relationships and patterns between cells. Finally, these graphs can be represented using adjacency matrices, where rows and columns correspond to cells, and values are encoded as ones for neighboring cells or zeroes for non-neighbors, allowing for the application of linear algebra and graph theory methods[193].

2.5.2. Unsupervised clustering: Leiden algorithm

The goal of unsupervised cell clustering is to detect dense regions in the KNN-graph using community detection methods such as the Leiden algorithm[194], which is widely used for scRNA-seq data analysis. These methods aim to identify clusters or communities of densely connected cells that are minimally connected to cells outside the cluster. Specifically, the Leiden algorithm detects clusters by identifying enrichment over the expected background cell connectivity distribution. The resolution parameter of the algorithm can be adjusted to control the scale of the clustering; higher resolution values result in the detection of a higher number of clusters. Additionally, the algorithm allows for sub-clustering within a specific cluster, meaning it can identify finer cell types[195]. However, this can also be achieved by treating the cluster as a new, separate dataset and re-processing it from scratch for a more detailed analysis.

2.5.3. Cell identity annotation

The term “cell annotation” refers to the process of assigning cellular phenotypes (biological identities) to previously defined clusters. Ideally, cell types should be robust across datasets, identifiable based on the expression of well-known specific gene markers, and linked to specific functions. Additionally, cells can be further classified into cell states, which are cells that may alter their expression under particular conditions, such as the activation of a specific pathway[195,196]. While some cell types, like immune cells, are found across different tissue biopsies, others may be sample-specific. Understanding which cell types are expected and deciding at what level of detail to capture cell identity is essential for effective data analysis.

Manual cell annotation

Manual cell annotation relies on a robust set of marker genes (or expression thresholds) that are well-known to be associated with particular cell types or states; this information can be obtained from protein-level studies or from published scRNA-seq atlases. After selecting a specific clustering resolution, one can examine which clusters express a specific list of marker genes. However, due to data sparsity, some gene markers may be expressed in only a subset of cells within the cluster. Alternatively, differential expression analysis (DEA) can be performed to identify genes that are highly expressed in the defined clusters and then determine if these genes can be used to associate clusters with cell types based on the canonical gene markers. In this context, DEA can be done using methods such as Wilcoxon rank-sum test[197] or MAST[198], which use normalized data and can compute either “one vs all” or “one vs one” across all clusters. It is important to note that while log-fold change indicates the strength of a gene as a cluster marker, the p-values obtained may be inflated because the statistics are computed on the same data used for clustering[199].

Marker-based annotation is sensitive to the chosen cluster resolution and the robustness of the gene markers, thus requiring prior biological knowledge, extensive literature review and experience in scRNA-seq annotation. Consequently, with the increasing number of scRNA-seq atlases published, the field is gradually shifting towards automated cell annotation approaches.

Automated cell annotation

For marker gene-based automatic classifiers, such as *CellAssign*[200] or *scGate*[201], cells are annotated based on their expression levels of a specific, predefined set of markers. Alternatively, *CellTypist*[202] uses a larger set of genes and is trained on previously annotated datasets.

The "query-to-reference" mapping approaches enable the project of new datasets into well-annotated references (using existing low-dimensional representation), allowing both datasets to be visualized together. This process then facilitates label transfer from the query to the reference using the joint embedding. By using an existing latent space, it avoids the need for data reintegration and batch correction but does so at the cost of not capturing unseen variations present in the query. For this reason, transferred labels have an associated uncertainty score that indicates the accuracy of the annotation; low scores correspond to high uncertainty, which may help to identify condition-specific cell states not found in the reference. The success of annotation predictions depends on the quality of the "training data" (reference) and the resolution of its annotations, the robustness of the classifier (e.g., neural network-based, support vector machines, or linear regression models), and how well the "query data" (dataset of interest) matches the training dataset[203]. Although many query-to-reference packages are available[204], the most common ones are *Seurat v4*[177], which assigns the majority label from its KNN graph, *Symphony*[205], a linear mixture framework that captures model components into minimal reference elements, and *scArches*[206], which is compatible with variational autoencoders like *scANVI*[207].

In summary, automated cell annotation methods are fast and straightforward, requiring no pre-processing of training data. They are independent of clustering methods, leverage validated knowledge, and contribute to harmonizing cell type profiles across the community, leading to a more consensus definition of cell identities. However, should be used with caution and considered a starting point rather than as a final data annotation, as manual marker-based approaches remain the most accepted method[173].

However, manual cell identity annotation is not a straightforward step, and achieving the most accurate results requires following a top-down strategy involving recursive processing steps[196,208]. The process begins with the entire dataset, where an initial high-level or low-resolution cell annotation is performed, typically identifying major cell lineages. Then, within each of these major cell groups, the same processing steps are repeated, involving data dimensionality reduction (such as highly variable gene selection and PCA), constructing KNN graphs, and performing

unsupervised clustering. This allows for fine-grained or high-resolution cell annotation, corresponding to specific cell types or states. Essentially, this recursive approach helps remove background noise and better captures biological differences within populations that would otherwise be difficult to detect.

2.6. Batch-effect removal and data Integration

Batch effects arise from differences in experimental design groups rather than the conditions being tested. These effects can result from technical variations during the handling of cells in different groups (or batches) at any stage of the scRNA-seq experiment, from data collection to sequencing. However, batch effects are not always present, making it essential to assess the extent of the batch effect before attempting any corrections.

The most effective approach to minimize technical variation and prevent batch effects is through careful experimental design, standardized protocol workflows, controlling for confounders, and including sufficient replicates. When prevention for batch effects is not possible, especially in retrospective studies or despite a well-designed experiment, batch-effect removal methods are necessary for joint analysis across different data batches. However, these correction methods risk over-correcting confounders, potentially leading to the loss of true biological variation.

Batch-effect removal and data integration methods involve data modeling to eliminate confounding technical effects. Some methods improve performance by starting with dimensionality reduction, while others finish by projecting the data back into the original high-dimensional space, resulting in a batch-corrected gene expression matrix. In general, batch correction methods can be classified into global models (originated from bulk RNA-seq methods), linear embedding models, graph-based methods, and deep learning approaches.

- **Linear embedding models** project cells into a low dimensional embedding (latent space), which is used to identify groups of similar cells in other batches, and to compute a correction vector across batches using, for instance, mutual nearest neighbors (MNN). Some methods project the data back into gene expression space while others only provide a corrected embedding. Examples: *Scanorama*[209]*, Seurat*[184] and *Harmony*[210].

- **Graph-based methods** represent the data from each batch using a nearest-neighbor graph. To correct batch effects, they force connections between cells from different batches while accounting for differences in cell composition.

- **Deep learning (DL) approaches** rely on autoencoder (AE) networks. These methods either use a conditional variational autoencoder (CVAE), conditioning the dimensionality reduction on a specific batch covariate, or apply a locally linear correction. The methods are iteratively trained through the networks to achieve convergence, resulting in a

corrected latent space. Some DL approaches incorporate cell identity labels to learn and preserve biological variation while avoiding its correction as a batch-effect. However, since batch correction is a preprocessing step, it might not be suitable when labels are not yet available. Examples: *scVI*[211], *scANVI*[207], which refines *scVI* trained model with cell annotations, and *scGen*[212].

When selecting an integration method, several factors must be considered: the method and its parameters, the level at which batch covariates occur (e.g., samples, donors, datasets), the dimension of the dataset, and the required output format. Although integration methods have been benchmarked to highlight their strengths and weaknesses, no single method is optimal for all scenarios; for example, while *Harmony* is recommended for simple integration tasks, *Scanorama* and *scVI* perform well on more complex datasets. However, if cell annotations are available, *scANVI* and *scGen* perform outperform the others[174,213]. Evaluating these methods involves assessing their ability to remove batch effects (e.g., Local Inverse Simpson's Index (LISI), kBET metrics) while preserving biological variation (e.g., Normalized Mutual Information (NMI), Adjusted Rand Index (ARI) metrics), but this is challenging due to the lack of a true biological reference for comparison. To assist with method selection, researchers have developed evaluation pipelines like *scIB*[213] and *BatchBench*[214]. Performing data integration and batch-effect removal in complex, large datasets has been a technical challenge addressed in **Publication 2** of this thesis.

2.7. scRNA-seq data analysis

The field of single-cell data analysis has seen a rapid expansion, with an increasing number of tools and packages developed for both data processing and analysis. These tools are crucial for understanding sample complexity and extracting biologically meaningful insights.

2.7.1. Comparative analysis of experimental conditions

Compositional analysis

Compositional analysis is conducted to identify differences in cell abundance across different experimental conditions. It is crucial that these analyses include sufficient numbers of samples (replicates) and cells to effectively account for background variation and achieve statistically significant outcomes. Furthermore, when comparing differential abundance populations, it is necessary to use cell proportions (always summing up to one) rather than absolute cell counts, resulting in data that is inherently compositional. In such cases, the "sum-to-one" constraint can introduce misleading correlations between cell types that have not undergone any changes.

Over the years, several methods have been employed for compositional analysis; some overlook the compositional nature of the data, while others struggle with a limited number of replicates. A

recent approach that effectively addresses these challenges to assess differences in known cell-types is *scCODA*[215], a Bayesian approach that accounts for uncertainty in cell proportions and corrects the negative correlative bias. scCODA provides unique results that must be interpreted relative to a reference cell type, which should maintain a constant relative abundance across conditions. When compositional changes are assessed in transitional states or within a cell population, KNN-based methods, such as Milo[216], are more powerful.

Differential Expression Analysis (DEA)

Differential Expression Analysis (DEA) can be utilized not only to identify cellular structures but also to assess the magnitude and significance of gene expression differences between various experimental conditions, such as comparing disease states to healthy references. Although these analyses can be conducted on whole-sample data using the popular *t-student test*, scRNA-seq data allows to explore differences at the level of specific cell types or states. However, scRNA-seq has unique challenges, such as pseudo-replication (multiple replicate cells) introducing correlations between measurements, dropout events leading to matrix zero-inflation, and high cell-to-cell variability. To address these issues, various tools have been developed. For instance, naïve methods like the *Wilcoxon rank-sum test* (*Seurat)*[177] incorrectly assume sample independence. Other approaches involve generating condition "pseudobulks" (using sum or mean aggregation) and applying bulk RNA-seq methods like *edgeR*[217] or *limma*[218]. Additionally, some methods specifically account for individual cells as a random effect using Generalized Mixed Effect (GME) models, such as *MAST*[198]. However, there is little consensus on the best approach, but some studies suggest that pseudobulk methods perform best, while scRNA-seq specific methods are prone to incorrectly labeling highly expressed genes as differentially expressed[219,220]. The outcome of such analyses typically includes the log2 fold-change (Log2FC) and the adjusted p-value for each gene in the comparisons. These results can be further ranked to identify the genes that contribute the most to potentially explaining any observed phenotypic differences across conditions.

Gene Set Enrichment Analysis (GSEA)

Following DEA, it is common to perform Gene Set Enrichment Analysis (GSEA) based on the DEGs to identify gene programs (e.g., biological processes, gene ontologies, or regulatory pathways) that are overrepresented in a specific experimental condition compared to others[221]. Thus, to perform GSEA, curated gene sets that define biological processes or pathways based on previous studies or immunological knowledge are required. While most gene sets are derived from bulk RNA-seq data, like those in the *Molecular Signatures Database (MSigDB)*[222], some databases now compile information from scRNA-seq data, such as *CellMarker*[223] and *PanglaoDB*[224]. A common implementation for GSEA can be found in *fgsea*[225] or *DecoupleR*[226] packages, which compute enrichment statistics based on a pre-ranked list of DEGs against the

gene sets of interest, providing log-fold changes (logFC) and p-values adjusted for multiple hypothesis testing. Interestingly, smaller gene sets can impair the performance of pathway analysis methods because they have higher gene variance. Additionally, these methods are particularly sensitive to normalization procedures, and that do not automatically account for batch effects. Therefore, it is recommended to filter out gene sets with low gene counts and ensure good data quality free from batch effects[227,228].

Gene Set Activity Inference (GSAI)

Similarly to conducting GSEA between conditions using whole-data or specific cell-types, it is also possible to infer the activity level of gene sets in each individual cell independently of gene expression in other cells or conditions. This approach, known as Gene Set Activity Inference (GSAI), can be implemented using *AddModuleScore()* from *Seurat*[177], *Ucell*[229], or *AUCell*[230] to compute enrichment scores. Additionally, more complex tools, such as Multivariate Linear Models (MLM) or Univariate Linear Models (ULM) from *DecoupleR*[226], fit models to predict observed gene expression based on pathway-gene interactions, providing a score that corresponds to the t-value of the slope (with a positive value indicating pathway activity). Similar to GSEA, these GSAI tools do not account for batch effects. Therefore, it is crucial to ensure that the data has been properly processed, taking into account the intrinsic dataset characteristics.

2.7.2. Machine Learning (ML) for single-cell genomics

Single-cell technologies have led to the exponential generation of high-throughput genomic data. Combined with increased computational power and resources, more sophisticated analytical methods, including machine learning (ML) algorithms and, even Deep Learning (DL) models, have become feasible. These advanced techniques, already applied in other fields like computer vision or natural language processing[231], are now applied to single-cell data[232,233].

DL methods are increasingly employed due to their flexibility and capacity to handle the complexity of scRNA-seq data, which involves high-dimensionality (thousands of genes and millions of cells), sparsity, noise, heterogeneity (i.e., diverse cell populations with varying gene expression profiles), and inherent nonlinear interactions among genes[232,233]. These approaches can be divided into three main categories based on the amount of information available and the task to be performed: supervised learning is used when complete information is available for tasks such as classification, unsupervised learning is employed to identify underlying patterns such as cell clustering, whereas semi-supervised learning assists in guiding supervised tasks with minimal information[233]. DL approaches mainly use Deep Neural Networks (DNNs) to analyze large and complex datasets. These models aim to improve the outcome of scRNA-seq analysis steps discussed in previous sections, such as data denoising, normalization, batch effect removal and data integration, supervised cell annotation, trajectory inference, gene network reconstruction, biomarker discovery, and even multimodal data analysis[232,233] (Figure 10).

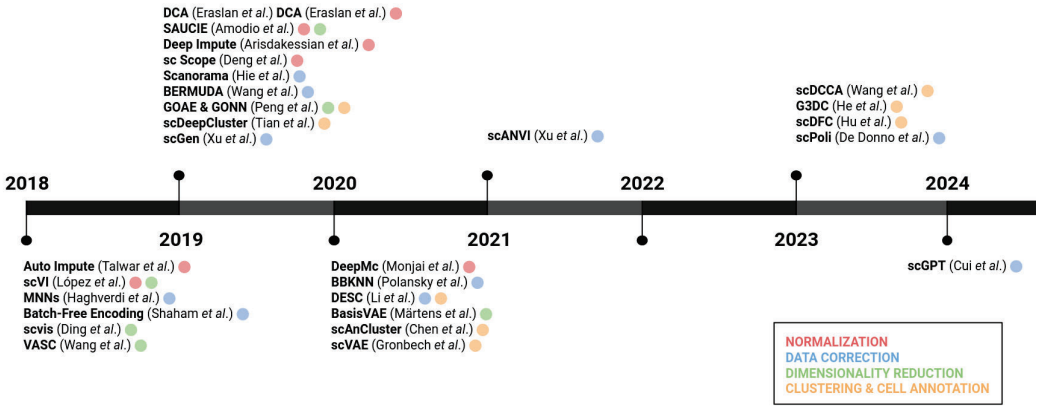


Figure 10. Timeline of Deep Learning (DL) methods developed for scRNA-seq data analysis. The plot displays the publication year and the first author. Circles indicate the specific analysis step each method was developed for, including data normalization, data correction (batch-effect removal and data integration), dimensionality reduction, clustering and cell annotation. Only the steps relevant to the work of this thesis are included. *Generated with data from* Erfanian *et al*.[233]

Among DL architectures, autoencoders (AEs) are widely used in the single-cell analyses field. AEs are flexible DNNs that learn from input data, such as gene expression profiles, and encode data into low-dimensional hidden layers. Data then follows a decoding procedure, with the aim of obtaining input values. After model training, parameters learned in the lowest dimensional layers could be used to embed cells into a latent space (similar to PCA but with non-linear combinations of genes) (Figure 11). For instance, the scVI[211] batch-correction method leverages this approach by incorporating a "condition" where data compression into the latent space is guided by an auxiliary batch covariate[232,233].

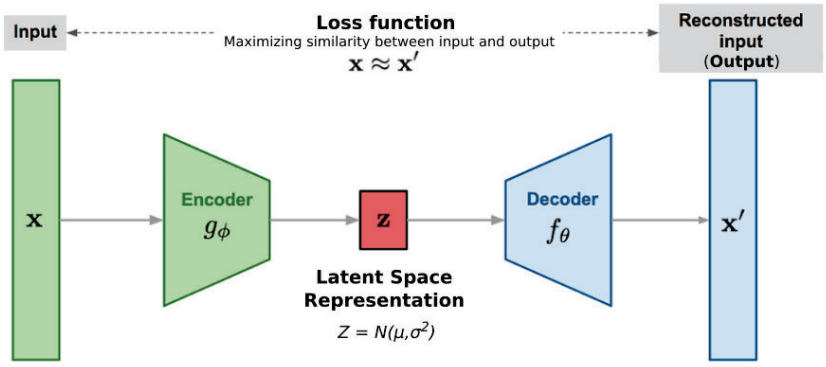


Figure 11. Schematic representation of a basic autoencoder (AE) architecture. Starting from the input data (X), the encoder function (g) compresses the high-dimensional data into a lower-dimensional latent space (Z), capturing key features in the hidden layers. The decoder function (f) then reconstructs the original input (x') from the compressed representation. The objective is to minimize the difference between the input and the reconstructed output, typically measured using a reconstruction loss function. *Adapted from* Lil'log [https://lilianweng.github.io].

The complex, multi-layer architecture of DL models turns them into "black boxes", where it is not straightforward to know the relations learned by the model (e.g., how gene expression is linked to latent features). This has led to a growing focus in research on improving the interpretability of

DL models in the context of scRNA-seq analysis, where it is essential to understand the molecular mechanisms underlying biological processes like the immune response[234]. Those models require post-hoc methods such as SHapley Additive exPlanations (SHAP)[235], model-agnostic techniques that extract important biological features from the trained model[234].

Previous mentioned approaches have already shown promising results for uncovering new aspects of cellular behavior and understanding cellular heterogeneity in biological systems when applied on scRNA-seq data. However, robust and systematic benchmarking is critically needed[232,233,236]. While some studies compare DL against previous state-of-the-art methods, comprehensive evaluations across different ML models are lacking. In fact, a recent benchmarking study found that DL methods do not always outperform classical ML approaches in cell annotation tasks[237]. Therefore, such evaluations are crucial for identifying the potential and limitations of these methods and determining which ML approach performs best for specific tasks or conditions.

Objectives

The main objective of this PhD thesis, titled "**Towards a large-scale human inflammation atlas at single-cell resolution**", is to design new experimental and computational approaches for large-scale, decentralized studies, focusing on analyzing human inflammatory profiles using single-cell genomics technologies.

The specific goals of this work are the following:

1. Develop and validate a new experimental protocol for sample fixation at the time of collection to be used in the context of single-cell genomic studies.

2. Implement strategic, high-throughput experimental designs and develop computational approaches to analyze single-cell data from large-scale studies.

3. Generate a reference atlas of inflammation from circulating immune cells across multiple diseases and in healthy individuals using single-cell transcriptomics.

4. Characterize inflammation-associated profiles in diseases and healthy status.

5. Identify inflammatory biomarkers for specific cell types and disease.

6. Design a new framework for classifying patients based on their inflammatory profile.

7. Investigate the immunomodulatory effects of albumin treatment in patients with acute decompensated cirrhosis at single-cell resolution.

Publications

THESIS DIRECTOR'S REPORT

As the director of this Doctoral Thesis, titled "**Towards a large-scale human inflammation atlas at single-cell resolution**", conducted by Laura Jiménez Gracia, I confirm that it is presented as a compendium of 3 publications.

The candidate's contribution to each publication are below:

PUBLICATION 1

FixNCut: single-cell genomics through reversible tissue fixation and dissociation.

Laura Jiménez-Gracia, Domenica Marchese, Juan C. Nieto, Ginevra Caratù, Elisa Melón-Ardanaz, Victoria Gudiño, Sara Roth, Kellie Wise, Natalie K Ryan, Kirk B. Jensen, Xavier Hernando-Momblona, Joana P. Bernardes, Florian Tran, Laura Katharina Sievers, Stefan Schreiber, Maarten van den Berge, Tessa Kole, Petra L. van der Velde, Martijn C. Nawijn, Philip Rosenstiel, Eduard Batlle, Lisa M. Butler, Ian A. Parish, Jasmine Plummer, Ivo Gut, Azucena Salas, Holger Heyn†, Luciano G. Martelotto†. ***Genome Biol*** 25, 81 (2024).

Impact Factor 2023: 10.1 (Q1)

In this paper, Laura made substantial contributions to the conception, design, and analysis of this study. She played a key role in analysis and interpreting the data, drafting significant portions of the manuscript, and revising it critically for important intellectual content. Laura's expertise in computational biology was instrumental in ensuring the accuracy and integrity of the analytical methods applied. She has reviewed the final version of the article and has provided her approval for its submission, agreeing to be accountable for all aspects of the work.

PUBLICATION 2

Interpretable Inflammation Landscape of Circulating Immune cells.

Laura Jiménez-Gracia*, Davide Maspero*, Sergio Aguilar-Fernández*, Francesco Craighero*, Maria Boulougouri, Sara Ruiz, Domenica Marchese, Ginevra Caratù, Jose Liñares-Blanco, Miren Berasategi, Ricardo O Ramirez Flores, Angela Sanzo-Machuca, Ana M. Corraliza, Hoang A. Tran, Rachelly Normand, Jacquelyn Nestor, Yourae Hong, Tessa Kole, Petra van der Velde, Frederique Alleblas, Flaminia Pedretti, Adrià Aterido, Martin Banchero, German Soriano, Eva Román, Maarten van den Berge, Azucena Salas, Jose Manuel Carrascosa, Antonio Fernández Nebro, Eugeni Domènech, Juan Cañete, Jesús Tornero, Javier Pérez-Gisbert, Ernest Choy, Giampiero Girolomoni, Britta Siegmund, Antonio Julià, Violeta Serra, Roberto Elosua, Sabine Tejpar, Silvia Vidal, Martijn C. Nawijn, Ivo Gut, Julio Saez-Rodriguez, Sara Marsal, Pierre Vandergheynst, Alexandra-Chloé Villani, Juan C. Nieto†, Holger Heyn†. ***BioRxiv*** *(2023)*

Manuscript under revision.

In this paper, Laura collected and preprocessed all the data included in the study. She contributed to data integration and cell identity annotation to generate the inflammation atlas. Additionally, she computed the inflammation-related signature activity scores and performed a comprehensive analysis across cell types and diseases. Davide Maspero and Francesco Craighero made significant contributions by developing and refining the machine learning models used in the analysis. Their work was essential for the accurate prediction and classification within the study. Sergio Aguilar-Fernández played a role in the analysis of gene regulatory networks and transcription factor identification. He also contributed to the data integration and cell identity annotation, with a particular focus on understanding the specific role of B cells across inflammatory diseases, emphasizing Systemic Lupus Erythematosus (SLE).

PUBLICATION 3

Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis.

Joan Clària* , Ferran Aguilar* , Juan-José Lozano*, **Laura Jiménez-Gracia***, Juan C. Nieto*, Berta Romero-Grimaldo, Xavi Marcos-Fa , Emma Giarracco, Emmanuel Weiss, Jonel Trebicka, Inmaculada Hernàndez, Javier Fernandez, Mireia Casulleras, Cristina López-Vicario, Sinan Muldur, Alex Hopke, Alexandru Vlagea, Ana M. Aransay, Domenica Marchese, Mauro Bernardi, Rajiv Jalan, Paolo Angeli, Giuliana Magri, Andrea Cerutti, Daniel Irimia, Holger Heyn†, Vicente Arroyo†, Richard Moreau†. Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis. ***JHEP Reports*** 6, 1-15 (2024).

Impact Factor 2023: 9.5 (Q1)

In this paper, Laura conducted all the computational analysis of the single-cell RNA sequencing (scRNA-seq) data. Her role was pivotal in processing and interpreting the scRNA-seq dataset to derive the meaningful insights reported in the publication. However, she did not handle patient collection, perform any in vitro or in vivo experiments, process samples for bulk or single-cell RNA sequencing, analyze bulk RNA-seq data, or conduct other functional assays.

Dr. Holger Heyn

Publication 1. FixNCut: Single-cell genomics through reversible tissue fixation and dissociation[1]

Laura Jiménez-Gracia, Domenica Marchese, Juan C. Nieto, Ginevra Caratù, Elisa Melón-Ardanaz, Victoria Gudiño, Sara Roth, Kellie Wise, Natalie K Ryan, Kirk B. Jensen, Xavier Hernando-Momblona, Joana P. Bernardes, Florian Tran, Laura Katharina Sievers, Stefan Schreiber, Maarten van den Berge, Tessa Kole, Petra L. van der Velde, Martijn C. Nawijn, Philip Rosenstiel, Eduard Batlle, Lisa M. Butler, Ian A. Parish, Jasmine Plummer, Ivo Gut, Azucena Salas, Holger Heyn†, Luciano G. Martelotto†. **FixNCut: single-cell genomics through reversible tissue fixation and dissociation.** *Genome Biol* **25**, 81 (2024).

† Joint senior authors

DOI: https://doi.org/10.1186/s13059-024-03219-5

Impact Factor 2023: 10.1 (Q1)

Protocol: https://www.protocols.io/view/fixncut-v1-0-14egn3xjql5d/v1

Raw data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE229944

Processed data: https://zenodo.org/records/10468192

Code: https://github.com/LJimenezGracia/FIXnCUT_benchmarking

[1] In this paper, I conducted the full analysis of the scRNA-seq data and performed all computational analysis. However, I did not develop the FixNCut method, process the samples for single-cell sequencing, perform the flow cytometry, or work with spatial staining. Therefore, I will only discuss the results from the scRNA-seq analysis and the improvements achieved through the application of the FixNCut protocol, which I can take credit from.

Jiménez-Gracia *et al. Genome Biology* (2024) 25:81
https://doi.org/10.1186/s13059-024-03219-5

Genome Biology

METHOD

Open Access

FixNCut: single-cell genomics through reversible tissue fixation and dissociation

Laura Jiménez-Gracia[1,2], Domenica Marchese[1,2], Juan C. Nieto[1,2], Ginevra Caratù[1,2], Elisa Melón-Ardanaz[3,4], Victoria Gudiño[3,4], Sara Roth[5,6,7], Kellie Wise[8,9,10], Natalie K. Ryan[9,11,12], Kirk B. Jensen[8,9,10], Xavier Hernando-Momblona[13,14], Joana P. Bernardes[15], Florian Tran[15,16], Laura Katharina Sievers[15,16], Stefan Schreiber[15,16], Maarten van den Berge[17,18], Tessa Kole[17,18], Petra L. van der Velde[18,19], Martijn C. Nawijn[18,19], Philip Rosenstiel[15], Eduard Batlle[13,14,20], Lisa M. Butler[9,11,12], Ian A. Parish[5,6], Jasmine Plummer[21], Ivo Gut[1,2], Azucena Salas[3,4], Holger Heyn[1,2,22*†] and Luciano G. Martelotto[8,9,22*†]

†Holger Heyn and Luciano G. Martelotto are joint senior authors.

*Correspondence: holger.heyn@cnag.eu; luciano.martelotto@adelaide.edu.au

[22] Omniscope, Barcelona, Spain
Full list of author information is available at the end of the article

Abstract

The use of single-cell technologies for clinical applications requires disconnecting sampling from downstream processing steps. Early sample preservation can further increase robustness and reproducibility by avoiding artifacts introduced during specimen handling. We present FixNCut, a methodology for the reversible fixation of tissue followed by dissociation that overcomes current limitations. We applied FixNCut to human and mouse tissues to demonstrate the preservation of RNA integrity, sequencing library complexity, and cellular composition, while diminishing stress-related artifacts. Besides single-cell RNA sequencing, FixNCut is compatible with multiple single-cell and spatial technologies, making it a versatile tool for robust and flexible study designs.

Keywords: Single-cell genomics, RNA sequencing, Sample fixation, Tissue dissociation, Cellular stress

Introduction

Single-cell sequencing has revolutionized our understanding of the complexity of life, allowing researchers to study tissues, organs, and organisms with unprecedented resolution [1]. However, most single-cell techniques are designed for freshly prepared specimens, which can present logistical challenges for decentralized study designs that require disconnecting the time and site of sampling from downstream processing steps. In this regard, preservation methods have been developed that enable sample collection and storage, expanding the applications of single-cell sequencing to personalized medicine and collaborative research. In addition to facilitating flexible study designs, early

© Crown 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

sample preservation can improve robustness and reproducibility by reducing artifacts introduced during sample handling, such as differences in lab personnel skills, library preparation workflows, and sequencing technologies. Furthermore, preservation methods can mitigate cellular stress caused by external factors, such as sample collection, transport, storage, and downstream processing steps involving mechanical or enzymatic dissociation, which can alter transcriptomic profiles [2–4]. Such cellular stress can impact sample quality and confound downstream analyses by inducing early stress-response genes and altering the natural state of the cell. Therefore, early sample preservation can enhance the quality and reliability of single-cell sequencing studies, while enabling flexible and decentralized study designs.

Dissociation-induced artifacts can be mitigated by the use of cold-active proteases active at low temperatures (6 °C) to decrease transcriptional activity and the expression of heat shock and stress-response genes. However, digestion at low temperatures can result in changes in cell type abundance due to incomplete tissue dissociation [2]. Alternatively, inhibitors of transcription and translation have been shown to reduce gene expression artifacts by minimizing the impact of dissociation-induced stress [5]. To overcome the challenges associated with sample logistics in single-cell studies, cryopreservation has been established as a storage method that preserves transcriptional profiles of cells in suspension and solid tissues [4]. However, cryopreservation can result in reduced cell viability and induce considerable changes in sample composition, such as the depletion of epithelial cells, myeloid suppressor cells, and neutrophils [3, 6–8]. Alternatively, cells can be fixed using alcohol, such as ethanol or methanol, but this can cause structural damage due to dehydration, protein denaturation, and precipitation, potentially affecting transcriptomic profiles. Nevertheless, alcohol-based fixation has been shown to better maintain cell composition compared to cryopreservation in specific contexts [3, 9]. More recently, ACME (ACetic-MEthanol) fixation has been developed as a solution to simultaneously dissociate and fix tissues, resulting in high cellular and RNA integrity [10]. Although ACME has been demonstrated to be effective when combined with cryopreservation for cnidarian samples, its value for sample preparation across other species, including mouse and human, remains to be shown. Cross-linking fixatives, such as formaldehyde and paraformaldehyde (PFA), are used with specialized single-cell assays, but are incompatible with commonly applied high-throughput single-cell RNA sequencing (scRNA-seq) protocols measuring transcriptome by standard polyA-based expression capture (3′ end sequencing). Moreover, formaldehyde-based fixation generally impedes the application of single-cell multiome analysis, cellular indexing of transcriptomes, and epitopes sequencing (CITE-seq) [11, 12] or immune repertoire profiling, which relies on 5′ end sequencing. Finally, post hoc computational tools such as machine learning algorithms have been developed to reduce or remove dissociation-induced artifacts [13]. For a more comprehensive description of methods to mitigate dissociation-induced artifacts, we recommend referring to the review by Machado et al. [14]. However, due to the often larger biological compared to technical variability and the fact that not all cell types within a sample suffer the same stress, it is difficult to generalize bias correction across cells and samples [15].

To overcome the challenges discussed above, it was crucial to develop a standardized protocol for sample collection, cell stabilization, and tissue processing that allows for

fixation prior to sample processing. We present a workflow, called FixNCut, which uses Lomant's Reagent (dithiobis(succinimidyl propionate) (DSP)), a reverse crosslinker fixative, to enable tissue fixation prior to sample digestion steps. Such order of events prevents changes in gene expression during digestion and processing, while disconnecting sampling time and location from subsequent processing sites. DSP is a homo-bifunctional N-hydroxysuccinimide ester (NHS ester) crosslinker that contains an amine-reactive NHS ester at each end of an 8-carbon spacer arm. NHS esters react with primary amines at pH 7–9 to form stable amide bonds and release N-hydroxy-succinimide. Proteins typically have multiple primary amines in the side chain of lysine residues and the N-terminus of each polypeptide that can serve as targets for NHS ester crosslinking reagents. DSP is lipophilic and membrane-permeable, making it applicable for intracellular and intramembrane conjugation. However, it is insoluble in water and must be dissolved in an organic solvent before adding to the reaction mixture. The presence of a disulfide bond in the center of DSP makes its crosslinking reversible via reducing agents like DTT, which is present in most reverse transcription buffers of single-cell sequencing applications (e.g., 10x Genomics assays). To date, DSP has only been successfully used to preserve cells in suspension (cell culture or PBMCs) for single-cell sequencing in applications such as CLint-Seq, nanofluidic systems, or phosphoprotein-based cell selection [16–18], but it has not been employed in tissues prior to dissociation.

Here, we demonstrate the versatility of the FixNCut protocol to overcome key limitations in generating single-cell data across multiple tissues. We provide evidence that FixNCut preserves RNA integrity, library complexity, and cellular composition, while allowing for cell labeling or sample hashing prior to single-cell analysis. To illustrate its potential, we applied FixNCut to fix and digest mouse lung and colon tissue, as well as human colon biopsies from inflammatory bowel disease (IBD) patients, demonstrating its clinical utility. Additionally, we show that DSP fixation can be used in the context of spatial-omics, specifically multiplexing tissue imaging for spatial proteomics (i.e., Phenocycler).

Results

Reversible fixation of human cells in suspension

Extending previous studies using DSP to preserve cell lines for RNA sequencing [16], we initially confirmed its applicability for single-cell analysis of cells in suspension (human peripheral blood mononuclear cells; PBMCs) and microfluidics systems (10x Genomics Chromium Controller), before combining fixation and dissociation of complex solid tissues. To this end, we compared cell morphology, RNA integrity, and reverse transcription efficiency of fresh and DSP-fixed PBMCs. Fixed cells showed highly similar morphology compared to fresh PBMCs in bright-field microscopy, with no evident changes in cell phenotypes or sample clumping after fixation (Fig. S1a). Next, we captured and barcoded single cells from both fresh and DSP-fixed samples using the Next GEM Single Cell 3′ Reagent Kits v3.1 from 10x Genomics. Bioanalyzer profiles of the amplified cDNA from both samples were virtually identical, demonstrating DSP fixation not to affect RNA integrity or reverse transcription performance (Fig. S1b,c). After sequencing, we confidently mapped over 80% of the reads from both sequencing libraries to the human reference genome, with over 50% of exonic reads usable for quantifying

gene expression levels (Fig. 1a). We observed a comparable correlation between the number of detected genes or the number of unique molecular identifiers (UMIs) and sequencing depth for fresh and fixed samples (95% CI, 3.65e−05 ± 1.11e−05 and 0.44 ± 0.06 vs. 3.56e−05 ± 1.14e−05 and 0.35 ± 0.06, respectively), indicating DSP fixation to conserve library complexity (Fig. 1b). Briefly, we captured a total of 22,481 genes in both conditions, together with 1667 and 1482 specific for fresh and fixed samples, respectively. We confirmed this observation at the single-cell level, where we found a similar relationship between sequencing depth and the number of detected UMIs or genes per cell (Fig. S1d). In line, we observed similar gene counts in single blood cells in fresh and fixed samples (Fig. 1c). After filtering out low-quality cells, we found a similar distribution of the main quality control (QC) metrics between both protocols (Fig. S1e), except for a few specific cell subpopulations (Fig. S1f). These results suggest DSP fixation to conserve the ability to detect gene transcripts in single cells compared to fresh samples in scRNA-seq experiments.

To further assess potential technical variation between protocols, we identified highly variable genes (HVGs) independently in fixed and fresh PBMCs. We found that 70% of HVGs were shared between the two protocols, indicating a conserved representation of the transcriptome and suitability for joint downstream processing (Fig. S1g). Additionally, when we examined the variation captured in the main principal components (PCs) and displayed single-cell transcriptomes in two dimensions (uniform manifold approximation and projection; UMAP), we did not observe any notable outliers due to the sampling protocol (Fig. 1d). Cells clustered together based on biological differences rather than preparation protocol, suggesting fixed and fresh cells to have similar capacity for cellular phenotyping. The pseudo-bulk gene expression profiles between fixed and fresh samples were highly correlated ($R^2 = 0.99$, $p < 2.2e{-}16$) (Fig. 1e), indicating DSP fixation not to alter the expression of specific genes. This was further confirmed at the cell population level (Fig. S1h). Moreover, the biological processes such as apoptosis, hypoxia, reactive oxygen species (ROS), cell-cycle (G2/M checkpoint), unfolded protein response (UPR), and inflammation hallmarks remained unchanged across libraries (Fig. 1f).

Next, we performed a joint analysis of 17,483 fresh and fixed cells, which were clustered to define 19 distinct cell populations (Fig. 1g; Fig. S1i). All cell types and states were found across both protocols at similar proportions, except for classical monocytes and NK cells, which showed small but significant differences, being slightly increased in fresh and fixed, respectively (Fig. 1h). Fixation did not affect differential expression analysis (DEA), with only four upregulated genes representing hemoglobin subunits (*HBA1*, *HBA2*, and *HBB*) and a mitochondrial gene (*MT-NDL4*) (Fig. 1i; Additional file 2: Table 1). These genes were consistently found across all cell populations (Additional file 3: Table 2), a phenomenon also observed when performing digestion protocols at low temperatures [3]. The FixNCut protocol may prevent erythrocyte lysis, leading to their co-encapsulation with nucleated blood cells and the detection of specific transcripts.

Importantly, we observed a reduction in technical artifacts introduced during sample processing prior to single-cell experiments upon fixation. Specifically, gene expression alterations previously defined to correlate with ex vivo PBMC handling [19] showed a significant reduction upon fixation ($p < 2.2e{-}16$) (Fig. 1j). Moreover, a sampling-time gene signature obtained from single-cell benchmarking studies [4] also showed a

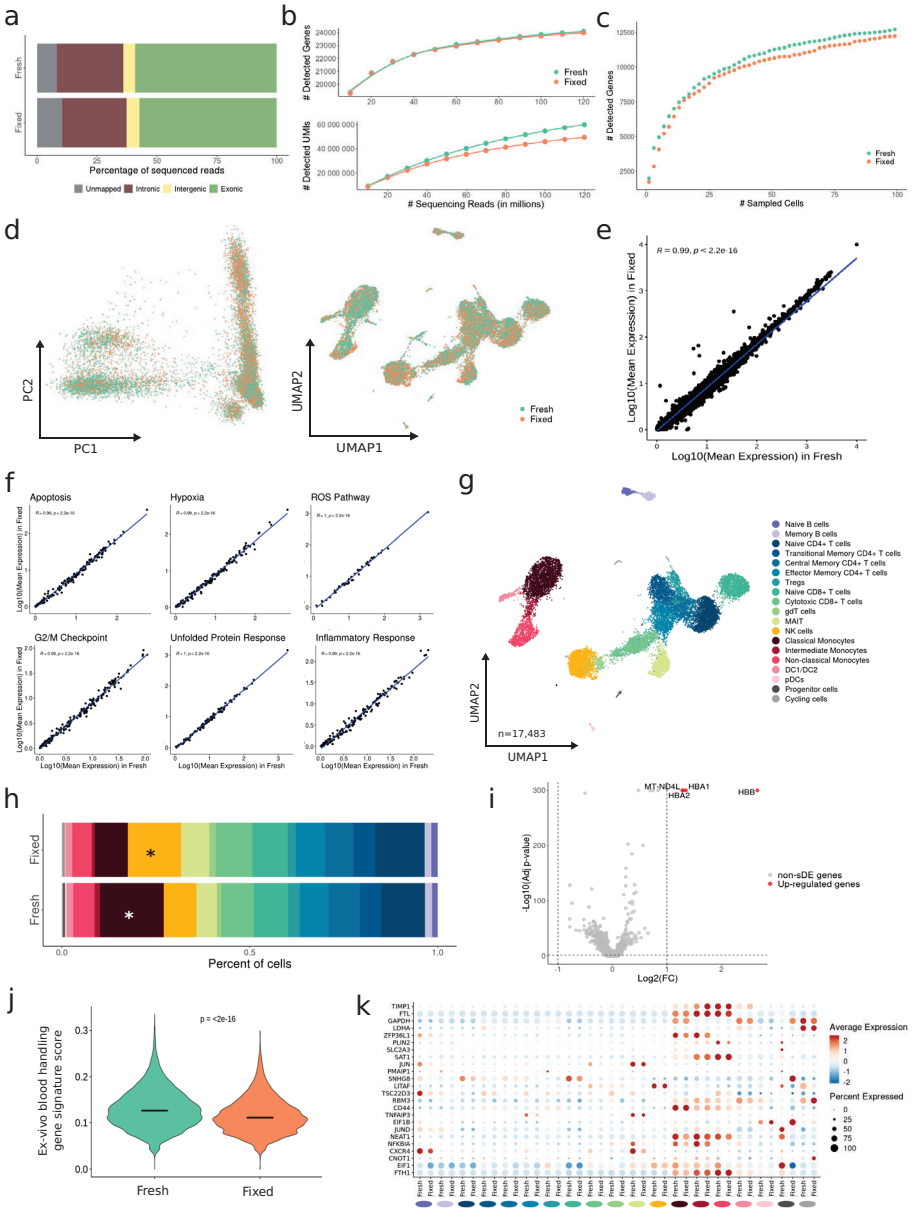


Fig. 1 FixNCut protocol in human peripheral blood mononuclear cells (PBMCs). **a** Mapping analysis of sequencing reads within a genomic region. **b** Comparative analysis of the number of detected genes (*top*) and UMIs (*bottom*) across various sequencing depths. **c** Cumulative gene counts analyzed using randomly sampled cells. **d** Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) representation of gene expression profile variances of fresh and fixed samples. **e**, **f** Linear regression model comparing average gene expression levels of expressed genes (**e**) and main biological hallmarks, including apoptosis, hypoxia, reactive oxygen species (ROS), cell cycle G2/M checkpoint, unfolded protein response (UPR), and inflammatory response genes (**f**). The coefficient of determination (R^2) computed with Pearson correlation and the corresponding *p*-value are indicated. **g** UMAP visualization of 17,483 fresh and fixed PBMCs, colored by 19 cell populations. **h** Comparison of cell population proportions between fresh (n = 9754) and fixed (n = 7729) PBMCs with the Bayesian model scCODA. Asterisks (*) indicate credible changes. **i** Differential gene expression analysis between fresh and fixed samples. The top differentially expressed genes (DEGs) with significant adjusted *p*-values (FDR) < 0.05, upregulated (red), and downregulated (blue) with Log2FC > |1| are indicated. **j** Violin plot of ex-vivo blood handling gene signature score [19] for fresh and fixed human PBMCs. Statistical analysis between fixed and fresh cells was performed using the Wilcoxon signed-rank test. **k** Dotplot showing the average expression of sampling-time DEGs for Fresh (*y*-axis) for all 19 cell types (*x*-axis) split by protocol. The dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level

significant reduction in the fixed PBMCs ($p < 2.2e{-}16$) (Fig. 1k). Interestingly, T lymphocytes appeared to be particularly affected ($p < 0.0001$; except for gdT cells, $p < 0.05$), showing the strongest protection from sampling artifacts in fixed cells (Fig. S1j). DSP also protected against the general reduction of gene expression activity, previously reported during PBMC sample processing [4]. Notably, more than 30% of genes from the sampling-time signature were also detected as enriched in the fresh PBMCs (Fig. S1k). The results suggest that fixed PBMCs have comparable cellular composition and gene expression profiles to freshly prepared samples, while reducing gene expression artifacts introduced during sample preparation.

FixNCut protocol applied on mouse solid tissues

Beyond the benefits of cell fixation in standardizing sample processing and preserving gene expression profiles of cells in suspension, the FixNCut protocol was specifically designed for solid tissues. Specifically, it allows for fixation and subsequent digestion, which is particularly advantageous for complex and logistically challenging study designs, such as clinical trials. Here, sampling artifacts, including biases in gene expression and cell type composition, are frequently observed in fragile solid tissue types. For example, differentiated colonic epithelial cells (e.g., secretory or absorptive cells), tightly connected adult neurons, or processing-sensitive adipocytes are more susceptible to cell damage and death as a result of common tissue dissociation protocols [20, 21]. Fixation prior to digestion using the FixNCut protocol can reduce these artifacts. Thus, we next evaluated the effectiveness of the FixNCut protocol with subsequent scRNA-seq readout in different solid mouse tissues before extending its application to challenging human patient samples, such as tissue biopsies.

Fresh mouse lung samples were minced, mixed, and split into two aliquots, one processed fresh and the other fixed using the FixNCut protocol with subsequent 30-min tissue digestion using Liberase TL. The fixed sample showed a slight decrease in cell size and an increase in DAPI+ cells, but overall, cell morphology was similar to the fresh sample (Fig. S2a). Single-cell encapsulation and scRNA-seq (10x Genomics, 3′ RNA v3.1) showed comparable proportions of reads mapped to the mouse reference genome and exonic genomic regions for both fresh and fixed samples (Fig. 2a). We further observed a similar correlation between the number of detected genes or UMIs and sequencing depth in fresh and fixed samples (95% CI, $2.31e{-}05 \pm 8.03e{-}06$ and 0.40 ± 0.06 vs. $2.64e{-}05 \pm 7.86e{-}06$ and 0.40 ± 0.06, respectively) (Fig. 2b). At the cell level, we confirmed the similar complexity of fixed libraries, as reflected by the number of detected UMIs and genes (Fig. S2b). Genes identified in both libraries ($n = 20{,}684$) were mostly protein-coding genes (76%). Conversely, genes exclusively captured in either fixed ($n = 1383$) or fresh ($n = 1157$) samples were largely non-coding RNA genes, specifically lncRNAs (52% vs 47%) (Fig. S2c). We further observed that more genes were captured for the fixed samples after accumulating information from a few number of individual lung cells (Fig. 2c). Importantly, after filtering out low-quality cells, the main QC metrics in fixed samples showed consistent distributions across all characterized cell types (Fig. S2d,e), suggesting that FixNCut protocol preserves the capacity for scRNA-seq profiling after fixation and digestion.

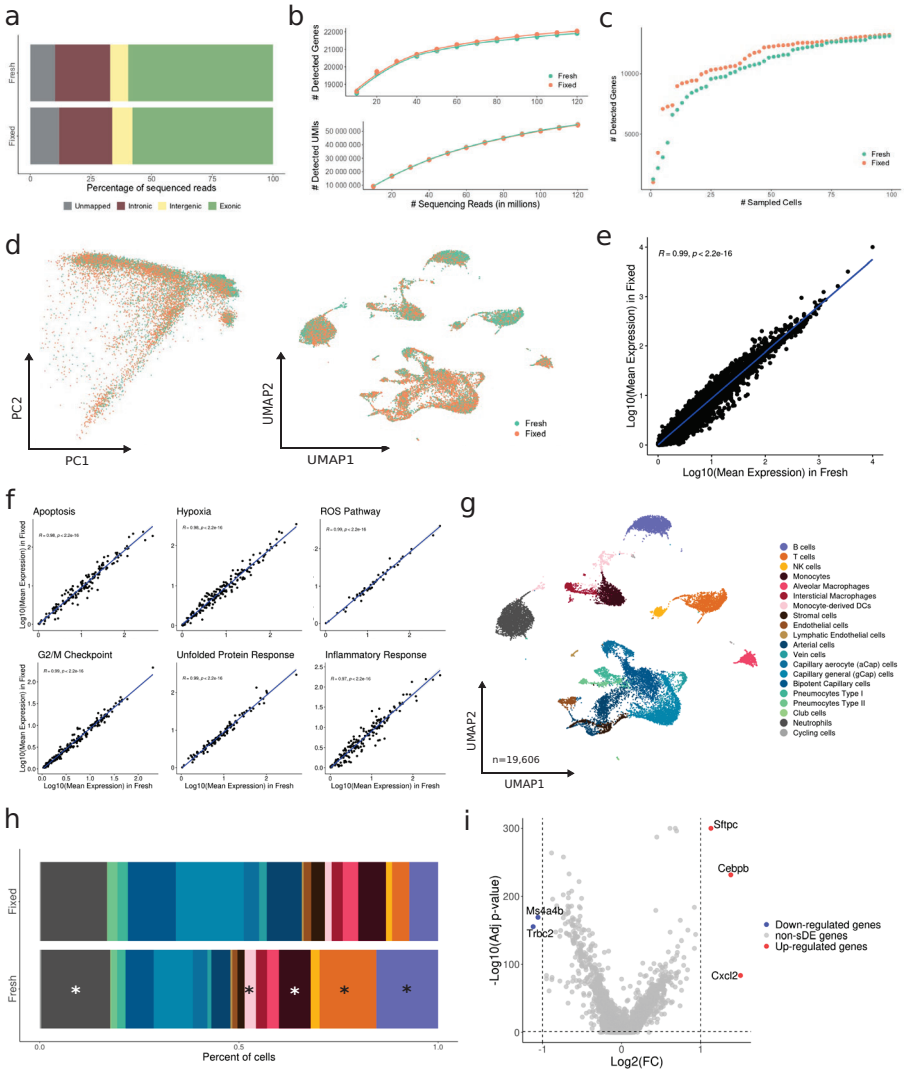


Fig. 2 FixNCut protocol tested in mouse lung samples. **a** Mapping analysis of sequencing reads within a genomic region. **b** Comparative analysis of the number of detected genes (*top*) and UMIs (*bottom*) across various sequencing depths. **c** Cumulative gene counts analyzed using randomly sampled cells. **d** Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) representation of gene expression profile variances of fresh and fixed samples. **e** Linear regression model comparing average gene expression levels of expressed genes between protocols. The coefficient of determination (R^2) computed with Pearson correlation is indicated. **f** Hierarchical clustering of coefficient of determination (R^2) obtained for all pair comparisons between protocols for biological hallmarks, including apoptosis, hypoxia, reactive oxygen species (ROS), cell cycle G2/M checkpoint, unfolded protein response (UPR), and inflammatory response genes. **g** UMAP visualization of 19,606 fresh and fixed mouse lung cells, colored by 20 cell populations. **h** Comparison of cell population proportions between fresh (n = 10,289) and fixed cells (n = 9317). The top figure shows the difference in cell population proportions between fresh and fixed samples, and the bottom figure shows the results of the compositional cell analysis using the Bayesian model scCODA. Asterisks (*) indicate credible changes, upregulated for the fresh sample. **i** Differential gene expression analysis between fresh and fixed samples. The top differentially expressed genes (DEGs) with significant adjusted p-values (FDR) < 0.05, upregulated (red), and downregulated (blue) with Log2FC > |1| are indicated

An overlap of almost 80% of sample-specific HVGs was found when comparing the fresh and FixNCut protocols (Fig. S2f). The absence of batch effects linked to protocols was demonstrated by the PCA and UMAP representations (Fig. 2d), indicating bias-free transcriptome profiles after cell fixation and digestion. Highly comparable

profiles of mean gene expression values were observed between fresh and fixed mouse lung samples ($R^2 > 0.99$, $p < 2.2e{-}16$) (Fig. 2e), a finding also confirmed at the population level (Fig. S2g). Moreover, the high correlation across gene programs supported the absence of alterations in major biological processes (Fig. 2f).

We then performed a joint analysis of all 19,606 mouse lung cells, which were segregated into 20 distinct cell populations, encompassing both lung and tissue-resident immune cells (Fig. 2g; Fig. S2h). All characterized cell types were detected in both fresh and fixed samples, with slight variability in cell type proportions between both protocols. The fixed protocol showed an improved representation of tightly connected epithelial and endothelial cell types, while immune cells (B and T cells, monocytes, monocyte-derived DCs, and neutrophils) were proportionally increased in the fresh sample (Fig. 2h). To validate preserved gene expression profiles in fixed tissues, we performed differential expression analysis (DEA) between the two protocols. We observed upregulation of genes related to pneumocytes (*Sftpc*), myeloid enhancer binding protein (*Cebpb*), and endothelial cells promoting cell migration (*Cxcl2*) for the fixed protocol, which could be largely explained by the enrichment of this population upon fixation. In contrast, fresh samples were enriched in genes related to inflammatory and immune processes (*Ms4a4b* and *Trbc2*), in accordance with the increased proportion of recovered immune cells (Fig. 2i; Additional file 4: Table 3). At the cellular level, we found a uniform enrichment in stress-related genes in non-immune populations from the fresh sample, while the fixed sample showed this enrichment in immune populations (Additional file 5: Table 4). Gene set enrichment analysis (GSEA) stratified by cell type revealed that freshly prepared endothelial cells were enriched in ROS, apoptosis, and cellular response to external stimuli, whereas the opposite patterns were observed upon fixation (Additional file 6: Table 5). Overall, these results suggest the global conservation of library complexity and quality, along with the inclusion of tightly connected, challenging-to-isolate cell types in fixed mouse lung samples.

We further evaluated the performance of the FixNCut protocol in a different challenging solid tissue context. To do so, we minced and mixed mouse colon samples that were split and subjected to scRNA-seq after digestion of either fresh or fixed tissues. Our results indicate that FixNCut provides several benefits, including improved transcriptome capture accuracy, as evidenced by a higher number of total reads mapped to the reference and a higher exonic fraction (Fig. 3a). Additionally, the fixed sample exhibited a higher non-significant library complexity based on the total number of detected genes (95% CI, $8.81e{-}05 \pm 2.83e{-}05$ vs. $9.03e{-}05 \pm 3.16e{-}05$) coupled with an increased number of total UMIs at deeper sequencing (95% CI, 0.71 ± 0.05 vs. 0.73 ± 0.07) (Fig. 3b), with fixed cells showing increased numbers of detected UMIs and genes (Fig. S3a). Genes identified in both libraries ($n = 18{,}314$) were mostly protein-coding genes (81%). Conversely, genes exclusively captured in fixed samples ($n = 2225$) compared to those from fresh ($n = 1011$) showed a larger percentage of protein-coding (45% vs 35%) coupled with a smaller fraction of lncRNA (44% vs 53%) (Fig. S3b), indicating that FixNCut enhances the gene capture efficiency, potentially allowing for a more fine-grained cell phenotyping after sample fixation. Notably, the

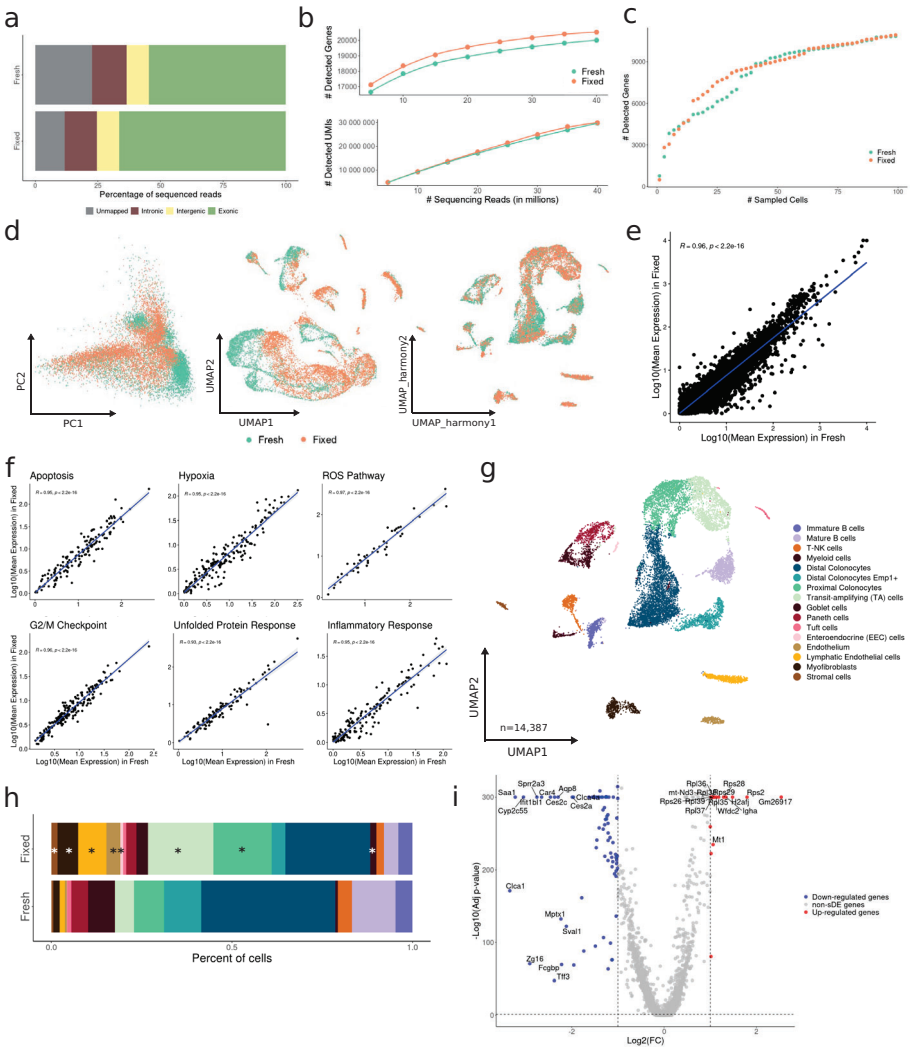


Fig. 3 FixNCut protocol tested in mouse colon samples. **a** Mapping analysis of sequencing reads within a genomic region. **b** Comparative analysis of the number of detected genes (*top*) and UMIs (*bottom*) across various sequencing depths. **c** Cumulative gene counts analyzed using randomly sampled cells. **d** Principal component analysis (PCA), uniform manifold approximation and Projection (UMAP) prior data integration, and *harmony* integrated UMAP representation of gene expression profile variances of fresh and fixed samples. **e**, **f** Linear regression model comparing average gene expression levels of expressed genes (**e**) and biological hallmarks, including apoptosis, hypoxia, reactive oxygen species (ROS), cell cycle G2/M checkpoint, unfolded protein response (UPR), and inflammatory response genes (**f**). The coefficient of determination (R^2) computed with Pearson correlation and the corresponding p-values are indicated. **g** UMAP visualization of 14,387 fresh and fixed mouse colon cells, colored by 16 cell populations. **h** Comparison of cell population proportions between fresh ($n = 6009$) and fixed ($n = 8378$) mouse colon samples with the Bayesian model scCODA. Asterisks (*) indicate credible changes, upregulated for the fixed sample. **i** Differential gene expression analysis between fresh and fixed samples. The top differentially expressed genes (DEGs) with significant adjusted *p*-values (FDR) < 0.05, upregulated (red), and downregulated (blue) with Log2FC > |1| are indicated

cumulative gene count was greater for the fixed colon, particularly when considering a larger number of sampled cells (Fig. 3c), and we observed improved QC metrics for the FixNCut sample after filtering out low-quality cells, which held true across all cell populations (Fig. S3c,d).

The overlap of HVGs between the fresh and fixed colon samples was slightly lower than that observed in lung tissues (> 60%) (Fig. S3e). Further, we identified noticeable differences in the transcriptomic profile, as demonstrated in both PCA and UMAP representations (Fig. 3d), which were attributed to the aforementioned improvements in library complexity after DSP fixation. Given that the overall cellular transcriptomic profile remains intact, confirmed by a high correlation in mean gene expression values between the two protocols ($R^2 = 0.96$, $p < 2.2e{-}16$) (Fig. 3e, f), we applied sample integration to collectively annotate cells and to address technical differences at cell type level (Fig. 3d). Notably, cell populations that exhibited a diminished correlation between the fresh and fixed samples coincided with cell types that were specifically enriched in the fixed sample (Fig. S3f).

We captured a total of 14,387 cells that were clustered into 16 cell populations, representing both immune and colon-epithelium cells (Fig. 3g; Fig. S3g). All cell types were detected in both conditions, but we observed a clear shift in cell type composition with an enrichment of sensitive epithelial and stromal cells in the fixed sample (Fig. 3h). Differential expression analysis revealed a higher representation of ribosomal protein and mitochondrial genes in the fixed sample, mostly explained by the larger capture of actively cycling epithelial cell population known as transit-amplifying (TA) cells (Fig. 3i; Additional file 4: Table 3). In line, both epithelial and stromal populations were also enriched in mitochondrial and ribosomal protein genes (Additional file 7: Table 6). Additionally, GSEA by cell population showed enrichment of ribosomal-dependent and metabolic processes pathways in fixed cells, especially in the sensitive populations (Additional file 8: Table 7). Together, our results demonstrate the FixNCut protocol to enhance library complexity and quality metrics, while also capturing fragile epithelial and stromal cell populations from delicate tissues, such as the colon. Thus, DSP-based fixation preserves the integrity of tightly connected cell types that are otherwise difficult to isolate for single-cell experiments, resulting in an improved representation of cell types and states in these solid tissues.

Long-term storage of fixed tissues

Conducting multi-center clinical studies can be challenging due to centralized data production and the need for storage and shipment. To address this challenge, we evaluated the combination of the FixNCut protocol with cryopreservation (cryo; 90% FBS and 10% DMSO; see the "Methods" section) to allow for the separation of sampling time and location from downstream data generation, while preserving sample composition and gene expression profiles. To test this, fresh mouse lung samples were minced, mixed, and split into three pools for fixation-only, cryo-only, and fixation+cryo sample processing. After single-cell capture and sequencing (10x Genomics, 3′ RNA v3.1), all three libraries showed comparable statistics of mapped and exonic reads across conditions, indicating successful preservation of the transcriptome (Fig. 4a). We also observed a similar relationship between the number of detected genes and the sequencing depth for all three protocols (95% CI, 2.88e−05 ± 8.64e−06 cryo vs. 2.37e−05 ± 8.07e−06 fixed/cryo), although the number of detected UMIs was statistically significantly reduced in the fixed/cryo sample compared to cryo-only (95% CI, 0.46 ± 0.06 cryo vs. 0.24 ± 0.06 fixed/cryo) (Fig. 4b), which was consistent considering the detected UMIs and genes

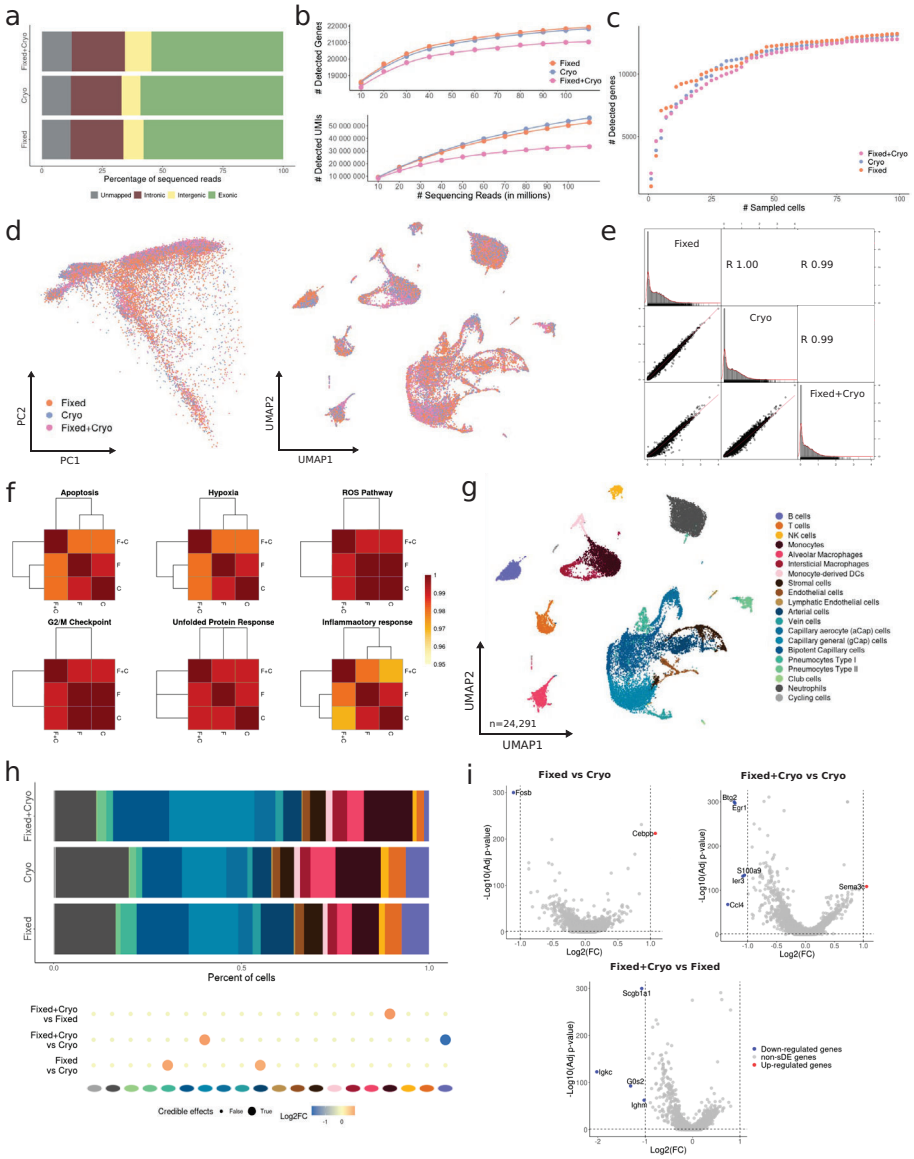


Fig. 4 Long-term storage of fixed mouse lung samples. **a** Mapping analysis of sequencing reads within a genomic region. **b** Comparative analysis of the number of detected genes (*top*) and UMIs (*bottom*) across various sequencing depths. **c** Cumulative gene counts analyzed using randomly sampled cells. **d** Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) representation of gene expression profile variances of fixed, cryopreserved, and fixed/cryopreserved samples. **e** Linear regression model comparing average gene expression levels of expressed genes across protocols used. The coefficient of determination (R^2) computed with Pearson correlation is indicated. **f** Hierarchical clustering of coefficient of determination (R^2) obtained for all pair comparisons across protocol for biological hallmarks, including apoptosis, hypoxia, reactive oxygen species (ROS), cell cycle G2/M checkpoint, unfolded protein response (UPR), and inflammatory response genes. **g** UMAP visualization of 24,291 fixed, cryo, and fixed/cryo mouse lung cells, colored by 20 cell populations. **h** Comparison of cell population proportions between fixed ($n = 10{,}256$), cryopreserved ($n = 8609$), and fixed/cryopreserved cells ($n = 5426$). The top figure shows the difference in cell population proportions between fixed, cryo, and fixed/cryo samples, and the bottom figure shows the results of the compositional cell analysis using the Bayesian model scCODA. Credible changes and Log2FC are indicated. **i** Differential gene expression analysis across conditions: fixed vs cryo (*top-left*), fixed/cryo vs cryo (*top-right*), and fixed/cryo vs fixed (*bottom*). The top differentially expressed genes (DEGs) with significantly adjusted *p*-values (FDR) < 0.05, upregulated (red), and downregulated (blue) with Log2FC > |1| are indicated

across individual cells (Fig. S4a), but hardly noticeable when accumulating gene counts across multiple cells (Fig. 4c). Genes identified across all libraries (n = 19,509) were mostly protein-coding genes (78%). Conversely, genes exclusively captured in fixed/cryo samples compared to those from cryo-only showed a similar percentage of protein-coding genes (34% vs 36%) and comparable fractions of lncRNA (52% vs 50%) (Fig. S4b). After removing low-quality cells, we found highly comparable distributions for the main QC metrics across all samples. However, we noticed a small increase in the percentage of mitochondrial gene expression detected in the fixed/cryo sample (Fig. S4c). Similarly, the different cell populations showed consistent QC across conditions (Fig. S4d).

We confirmed the absence of DSP-fixation biases after cryopreserving fixed samples, as indicated by a high overlap (> 70%) of HVGs across all three protocols (Fig. S4e). In addition, both PCA and UMAP dimensionality reduction plots showed no discernible biases between preservation protocols (Fig. 4d). We also observed highly comparable expression profiles and gene programs when correlating the mean gene expression values for all protocol comparisons ($R^2 > 0.99$, $p < 2.2e{-}16$) (Fig. 4e). Moreover, there was no appreciable alteration in biological processes at the gene program or population level when comparing across protocols (Fig. 4f; Fig. S4f).

We next analyzed 24,291 mouse lung cells processed with the three different protocols and annotated 20 lung and tissue-resident immune cell populations (Fig. 4g; Fig. S4g), and all cell types and states were found across the three conditions at similar proportions. However, we observed slight changes in composition, with fixed/cryo samples showing a decrease of B cells coupled with an increase of gCap compared to the cryo sample, and an increase of Monocytes compared to fixed-only. Comparing fixed-only with cryo, we found an increase in arterial and pneumocyte type I cells compared to the cryo sample (Fig. 4h). Additionally, we observed downregulation of genes associated with immune function (e.g., *Igkc*, *Ccl4*, *Scgb1a1*) in the fixed/cryo, explained by the aforementioned shift in cell type composition. Importantly, the cryo-only sample showed upregulated genes related to stress response, such as *Fosb* (Fig. 4i; Additional file 4: Table 3). A closer inspection of the different cell populations validated the expression of stress-related genes across all cells in the cryo-only compared fixed/cryo samples, specifically in non-immune cells (Additional file 9: Table 8). Accordingly, GSEA detected an enrichment of regulatory or response pathways for almost all cryopreserved cell types compared to fixed/cryo samples (Additional file 10: Table 9). These results support the feasibility of cryopreservation after fixation to combine the robustness and logistical advantages of the respective methods for scRNA-seq experiments.

Minimization of technical artifacts in FixNCut tissue samples

Fixing tissues after sample collection preserves the natural state of a cell and avoids technical biases, previously shown to affect bulk and single-cell transcriptomics analysis [2–4, 22]. In addition to the abovementioned differences in stress-response genes, we further aimed to demonstrate the ability of FixNCut to preserve gene expression profiles by examining previously identified artifact signatures. Specifically, we investigated condition-specific gene signatures from published studies using our mouse lung and colon data (see the "Methods" section).

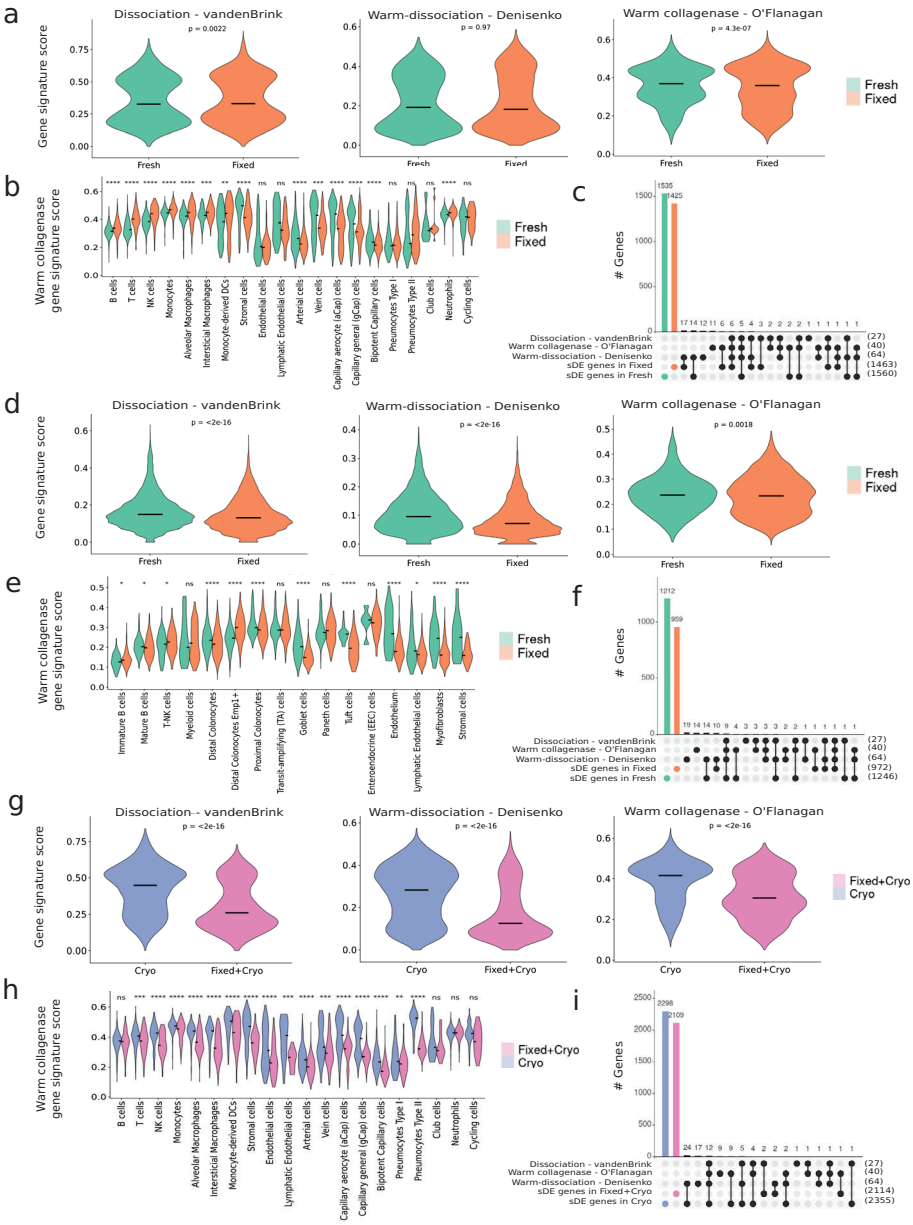


Fig. 5 Minimization of technical artifacts using FixNCut protocol on mouse tissues. This figure shows the impact of various dissociation-induced gene signature scores, including dissociation on mouse muscle stem cells [22], warm dissociation on mouse kidney samples [3], and warm collagenase dissociation on mouse primary tumors and patient-derived mouse xenografts [2], across mouse tissues and processing protocols used. All statistical analyses between protocols were performed using the Wilcoxon signed-rank test; significance results are indicated with the adjusted *p*-value, either with real value or approximate result (ns, $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$). **a** Violin plots of dissociation-induced gene signatures scores for fresh and fixed mouse lung. **b** Score of warm collagenase gene signature for fresh and fixed mouse lung samples across cell populations. **c** Overlap of differentially expressed genes in the fresh and fixed mouse lung sample with genes from the three dissociation-induced signatures. **d** Violin plots of dissociation-induced gene signatures scores for fresh and fixed mouse colon. **e** Score of warm collagenase gene signature for fresh and fixed mouse colon samples across cell populations. **f** Overlap of differentially expressed genes in the fresh and fixed mouse colon sample with genes from the three dissociation-induced signatures. **g** Violin plots of dissociation-induced gene signatures scores for cryo and fixed/cryo mouse lung. **h** Score of warm collagenase gene signature for cryo and fixed/cryo mouse lung samples across cell populations. **i** Overlap of differentially expressed genes in the cryopreserved and Fixed/Cryo mouse lung sample with genes from the three dissociation-induced signatures

After analyzing mouse lung samples, we found that fixed samples had comparable dissociation/temperature-signature scores, except for the warm collagenase which was significantly lower ($p < 4.3e{-}07$) compared to fresh samples (Fig. 5a). We observed that external tissue stressors had a greater impact on fresh lung resident cells compared to the infiltrating immune cell fraction (Fig. 5b). Interestingly, the signature scores for these populations displayed a bimodal-like distribution, indicating an uneven effect within cell populations (Fig. 5b). Additionally, stress-signature genes were not only found to be differentially expressed in the fresh lung but also in the fixed samples, regardless of their level of expression (Fig. 5c).

Similarly, fixed colon samples showed a significantly larger decrease in dissociation/temperature-stress signature scores compared to fresh samples (Fig. 5d). Here, we also observed a lineage-dependent impact of cell stress; colonocytes were greatly affected with differences in subtypes, whereas immune cells largely escaped stress biases (Fig. 5e). Endothelial and stromal cells suffered the largest dissociation-related stress in the fresh samples, which was drastically reduced upon fixation (Fig. 5e). Moreover, stress-signatures genes were also differentially expressed in the fresh colon sample, while largely absent in the fixed sample (Fig. 5f).

Furthermore, we demonstrated the effectiveness of FixNCut for long-term sample storage by examining the dissociation/temperature-stress signature scores in cryo-only and fixed/cryo mouse lung samples. Our results showed that cryopreserved samples had significantly higher stress-related signature scores compared to fixed/cryo ($p < 2.2e{-}16$) (Fig. 5g). Interestingly, the stress signature score for endothelial and stromal cells exhibited a bimodal distribution exclusively in the cryo-only sample, with cells showing larger dissociation-related effects in the same population (Fig. 5h), consistent with previous observations. Over 70% of signature-specific genes were significantly differentially expressed in the cryo-only sample, an even higher proportion compared to fresh lung and colon samples, whereas the fixed/cryo samples had almost no overlapping DEGs (Fig. 5i).

Moving towards the use of FixNCut on clinical samples

As a proof-of-concept for a multi-center clinical research study focused on autoimmune diseases, we evaluated the performance of FixNCut on human patient biopsies. To this end, we obtained fresh colon biopsies from two IBD patients in remission. The biopsies were mixed and split into four aliquots, which were processed as follows: fresh, fixed-only, cryo-only, or fixed/cryo. The fixed human colon samples exhibited a similar proportion of reads mapped to the reference genome and exonic regions as mouse colon tissues (Fig. 6a) and displayed comparable library complexity for short-term (fixed vs. fresh) and long-term (fixed/cryo vs. cryo) conditions, considering the total number of detected genes (95% CI $6.29e{-}05 \pm 2.03e{-}05$ fresh vs. $5.53e{-}05 \pm 2.13e{-}05$ fixed; $6.65e{-}05 \pm 2.23e{-}06$ cryo vs. $6.10e{-}05 \pm 2.11e{-}05$ fixed/cryo) and captured UMIs (95% CI 0.48 ± 0.07 fresh vs. 0.37 ± 0.08 fixed; 0.47 ± 0.07 cryo vs. 0.39 ± 0.08 fixed/cryo) (Fig. 6b). A similar pattern was observed comparing the number of captured genes and UMIs at the cell level (Fig. S5a). The cumulative detected gene count was highest for the fixed biopsy, with fresh being the worst condition (Fig. 6c). Genes identified across all libraries ($n = 21{,}637$) were mostly protein-coding genes (70%), followed by

non-annotated genes (21%). Genes exclusively captured in one condition (fresh, fixed, cryo-only, or fixed/cryo) showed a highly similar distribution of gene features (Fig. S5b). After removing low-quality cells, the main quality control metrics had similar distributions, with slightly improved median UMI, gene counts, and reduced mitochondrial gene percentage for the fixed sample (Fig. S5c), which were consistently observed across almost all populations (Fig. S5d). We observed an overlap (> 50%) of HVGs across all conditions (Fig. S5e) and although subtle protocol-associated effects were found, we ensure a consistent cell annotation across samples after successful sample integration (Fig. 6d). Gene expression profiles and gene programs significantly correlated across all samples ($R^2 > 0.96$, $p < 2.2e{-}16$) (Fig. 6e, f), with a slightly reduced correlation observed in M0 macrophages and stromal cells (Fig. S5f).

By jointly analyzing 17,825 IBD colon cells across protocols, we identified 21 major cell types, including both colon and tissue-resident as well as infiltrated immune cells (Fig. 6g; Fig. S5g). All cell types and states were found across all protocols at similar proportions, although the number of high-quality cells was reduced in the cryo-only sample (Fig. 6h). Notably, the FixNCut protocol captured larger proportions of CD4+ T and B cells compared to fresh or cryo-only samples, among other minor changes (Fig. 6h). We also found that fixed samples had downregulation of heat-shock proteins (HSP), such as *HSPA1A*, *HSPA1B*, and *DNAJB1*, and upregulation of B cell-specific genes, including *MS4A1*, *HLA-DRA*, *HLA-B*, and *CXCR4*, when compared to fresh and cryo-only human colon samples (Fig. 6i; Additional file 11: Table 10). In line with the findings from the mouse experiment, we observed that fixed human colon samples exhibited increased expression of ribosomal genes (related to TA cells) at the cell population level, whereas fresh samples showed higher mitochondrial expression. Apart from increased HSP genes in the cryo-only sample, no other significant differentially expressed genes were found between the conditions (Additional file 12: Table 11; Additional file 13: Table 12). GSEA revealed the cryo-only sample to be enriched in stress pathways (response to external stimuli such as stress, temperature,

(See figure on next page.)

Fig. 6 FixNCut protocol tested in human colon biopsies. **a** Mapping analysis of sequencing reads within a genomic region. **b** Comparative analysis of the number of detected genes (*top*) and UMIs (*bottom*) across various sequencing depths. **c** Cumulative gene counts analyzed using randomly sampled cells. **d** Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) representation of gene expression profile variances prior data integration, and *harmony* integrated UMAP representation of gene expression profile variances of fresh, fixed, cryopreserved, and fixed/cryopreserved samples. **e** Linear regression model comparing average gene expression levels of expressed genes across protocols used. The coefficient of determination (R^2) computed with Pearson correlation is indicated. **f** Hierarchical clustering of coefficient of determination (R^2) obtained for all pair comparisons across protocols for biological hallmarks, including apoptosis, hypoxia, reactive oxygen species (ROS), cell cycle G2/M checkpoint, unfolded protein response (UPR), and inflammatory response genes. **g** UMAP visualization of 17,825 fresh, fixed, cryopreserved, and fixed/cryopreserved human colon cells, colored by 21 cell populations. **h** Comparison of cell population proportions between fresh ($n = 5759$), fixed ($n = 4250$), cryo ($n = 3489$), and fixed/cryo ($n = 4327$) cells. The bottom figure shows the results of compositional cell analysis using the Bayesian model scCODA. Credible changes and Log2FC are indicated. **i** Differential gene expression analysis across conditions: fixed vs fresh (*top-left*), fixed vs cryo (*top-right*), fixed/cryo vs cryo (*bottom-left*), and fixed/cryo vs fixed (*bottom-right*). Significant adjusted *p*-values (FDR) < 0.05, upregulated (red), and downregulated (blue) genes with Log2FC > |1| are indicated. The top DE genes are included in the plot. **j** Violin plots for stress-related gene signature score [2, 3, 22] for human colon biopsies across protocols. Statistical analysis between protocols was performed using the Wilcoxon signed-rank test

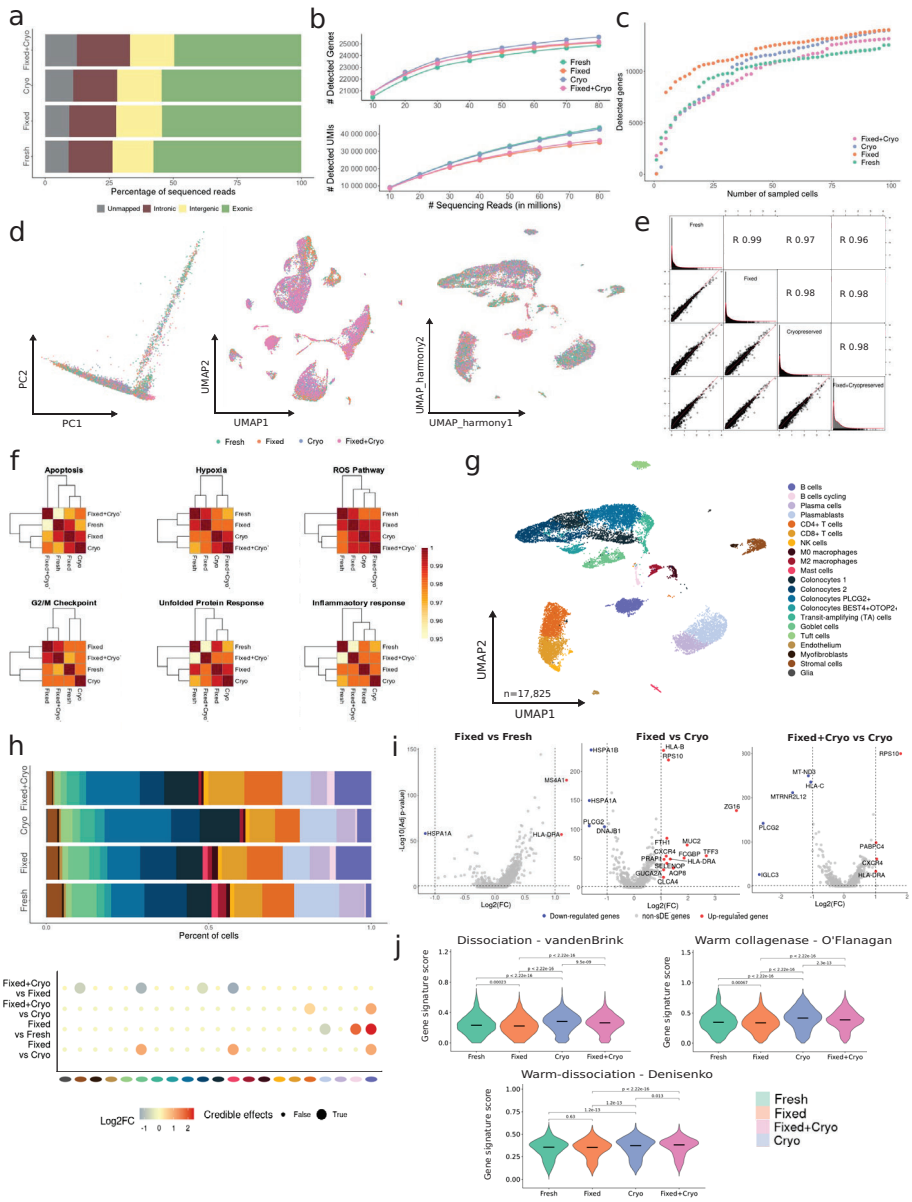


Fig. 6 (See legend on previous page.)

oxidative stress, and protein folding), a pattern also observed in the fixed/cryo sample compared to fixed, but restricted to immune cell types (Additional file 14: Table 13; Additional file 15: Table 14).

By comparing published stress signatures (see the "Methods" section) across fresh, fixed-only, cryo-only, and fixed/cryo samples, we found no significant differences between the fixed and fresh samples, whereas the cryo-only sample had significantly higher scores than the fresh sample, with the fixed/cryo sample presenting a lower stress scores than the cryo-only sample, indicating a reduction of gene expression artifacts by fixation for long-term storage in patient biopsies (Fig. 6j). We observed no significant cell population-specific effect; however, stromal cells had the highest score compared to other (Fig. S5h). These findings provide proof-of-concept evidence for the applicability

and value of FixNCut in improving the robustness of data generation in the clinical setting.

Expanding the application of FixNCut to various single-cell techniques

Next, we assessed the compatibility of the FixNCut protocol with single-cell application variants that involve cell labeling with antibodies or lipids targeting the cell membrane (e.g., FACS, CITE-seq, cell hashing). To this end, we stained fresh, cryopreserved, and fixed PBMCs and colon tissue samples with fluorescent antibodies or lipid-modified oligos (LMOs), which were analyzed by flow cytometry.

First, we analyzed a cohort of cryopreserved and cryo+fixed human PBMCs ($n = 3$) stained with anti-CD3, CD4, CD8, and CD19 monoclonal antibodies (mAbs). The analysis of cell morphology and viability showed that DSP fixation induced slight changes in cell size, internal complexity, and membrane permeability of PBMCs. Specifically, lymphocytes showed a decrease in forward scatter (cell size), while monocytes showed a decrease in side scatter (internal complexity) (Fig. S6a,b). Next, cells were gated based on the expression of CD3, CD4, and CD8 to characterize all T cell subtypes, and CD19 for B cells. We confirmed that DSP fixation did not alter the percentage of any of the immune cell types analyzed (Fig. 7a). However, using the same amount of antibodies, the mean fluorescence intensity (MFI) was higher in the cryopreserved compared to the cryo+fixed PBMCs (Fig. 7b). To ensure cryopreservation did not introduce any bias, we analyzed another cohort of fresh and fixed human PBMCs ($n = 4$) stained with anti-CD45, CD3, CD4, and CD8 mAbs and against ubiquitously expressed human surface proteins (β2M and CD298) and LMOs for cell hashing and multiplexing. DSP fixation led to an increased binding of DAPI, Annexin V (apoptotic cell marker), and propidium iodine (necrotic cell marker) in PBMCs, indicating reduced membrane integrity, particularly after storing cells at 4 °C for 2 days (Fig. S6c). These changes in morphology and membrane integrity should be taken into account when working with mixed study designs including both fixed and fresh samples. Despite observing a minor decrease in the MFI for most of the tested antibodies in fixed samples, we did not detect significant differences in cell type composition comparing fresh to fixed cells (Fig. 7c; Fig. S6d). Similarly, whereas PBMCs labeled with β2M and CD298 antibodies showed no differences in MFI between protocols, cells stained with LMOs revealed a minor but noticeable reduction in signal strength in fixed cells (Fig. 7d).

Similarly, staining dissociated human colon biopsies with anti-CD45, CD3, CD11b, and EpCAM antibodies showed similar MFI in cryopreserved and fixed cells (Fig. 7f; Fig. S6e). Our results showed that DSP fixation is compatible with cell labeling of cells with antibodies. Nevertheless, we recommend the optimization of labeling conditions and flow cytometry protocol of fixed samples, depending on antibody sensitivity, antigen abundance, and downstream applications, to further improve results.

Finally, we explored the potential applicability of DSP-fixed tissues in spatial technologies. We applied DSP-fixed tissue to spatial proteomics using multiplexed immunofluorescence tissue imaging, formerly known as co-detection by indexing (CODEX) [23]. We analyzed a DSP-fixed paraffin-embedded prostate cancer sample and its formalin-fixed counterpart using the commercial Phenocycler instrument (Akoya Biosciences). DSP-fixed sections exhibited a pattern of expression and signal intensity comparable

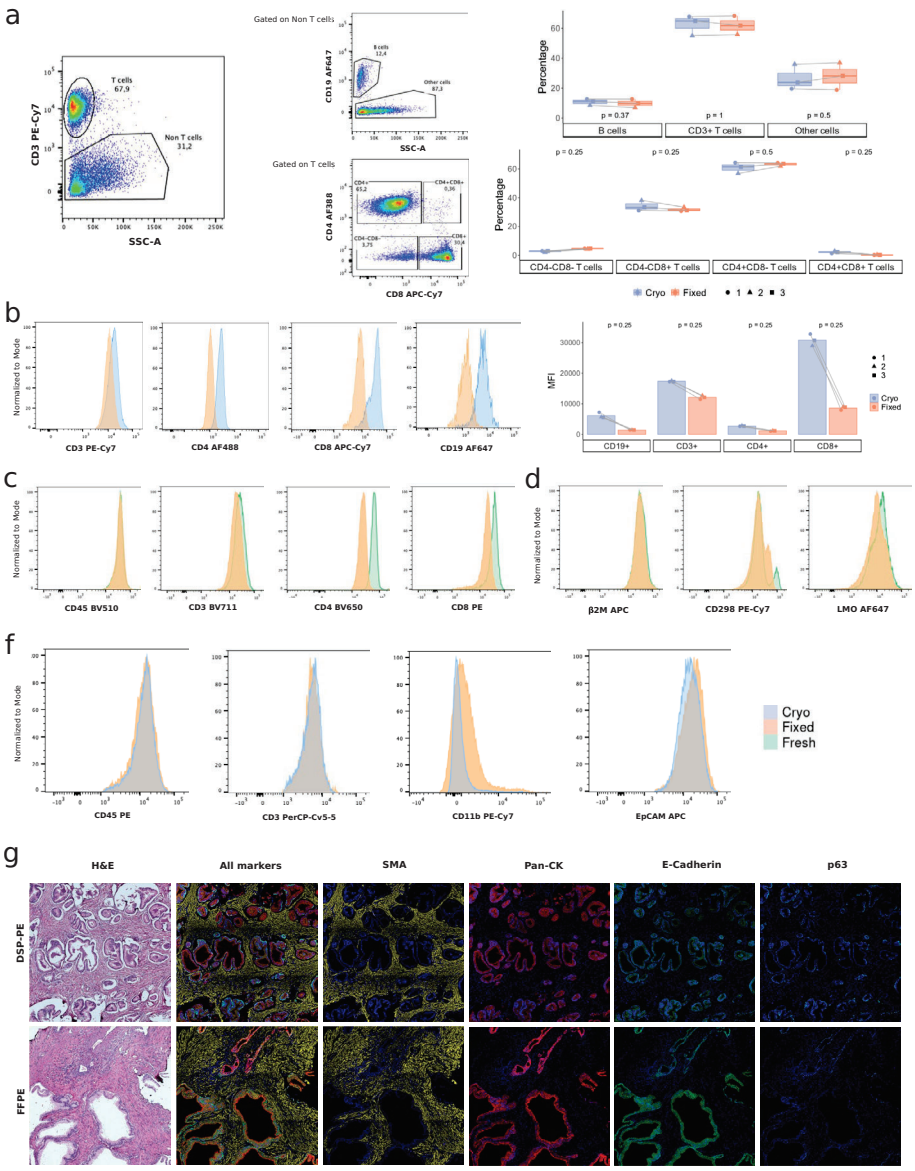


Fig. 7 Fluorescent antibody labeling of membrane proteins in fixed cells and tissues in mouse and human. **a** Representative gating strategy of one experiment analyzed by flow cytometry with cryopreserved and cryopreserved+fixed PBMCs from healthy donors ($n = 3$). PBMCs were stained with anti-human CD45, CD3, CD19, CD4, and CD8 monoclonal antibodies (mAbs). T cells were selected by the positive expression of CD3, whereas B cells were selected from the CD3-negative fraction (Non T cells) and by the positive expression of CD19. CD4-positive and CD8-positive T cells were selected from CD3-positive T cells. Box plots show the percentage of positive cells in each subpopulation for cryo (blue) and cryo+fixed (orange) PBMCs analyzed by flow cytometry. **b** Representative histograms of the mean fluorescent intensity (MFI) with cryopreserved and cryopreserved+fixed PBMCs from healthy donors ($n = 3$) stained with anti-human CD45, CD3, CD19, CD4, and CD8 mAbs. Bar plots show the MFI expression for three fresh and fixed PBMC samples analyzed by flow cytometry. **c**, **d** Representative histograms of MFI of fresh and fixed PBMCs from healthy donors ($n = 4$) stained with anti-human CD45, CD3, CD4, and CD8 mAbs, anti-β2M, anti-CD298, and LMOs analyzed by flow cytometry. **f** Representative histograms of the MFI from cryopreserved (blue) and cryopreserved+fixed (orange) human colon samples processed and stained with anti-human CD45, CD3, CD11b, and EpCAM mAbs and analyzed by flow cytometry. **g** Multiplex fluorescence tissue imaging of a human prostate cancer section, DSP-fixed (*top*) or formalin-fixed (*bottom*) paraffin-embedded, captured using Phenocycler. Images show hematoxylin and eosin staining, five-color overlay, and individual SMA, Pan-CK, E-cadherin, and p63 antibody staining

to formalin-fixed sections, the latter considered to be the "gold standard" (Fig. 7g). This study showcased the potential of DSP fixation in spatial-omics, serving as a proof-of-concept for its application to a broader array of spatial techniques. Additional efforts are underway to extend the use of DSP-fixed tissues to other spatial-omics technologies, such as Visium/Xenium (10x Genomics), GeoMx/CosMx (Nanostring), STereo-seq (BGI), and MERSCOPE (Vizgen).

Discussion

In this study, we introduced the FixNCut protocol, a novel approach that combines sample fixation with subsequent tissue dissociation to overcome several limitations in generating single-cell data. While DSP fixation has previously been used to fix K562 cells and keratinocytes prior to sequencing [16, 17], or in combination with single-cell technologies to explore the adaptive immune repertoire [18], our study is the first to utilize the reversible properties of the DSP fixative with standard enzymatic dissociation on solid biopsies in the context of single-cell transcriptomic studies.

The FixNCut protocol offers significant advantages, including the ability to fix tissue prior to digestion, providing a snapshot of the cell transcriptome at sampling time and minimizing technical artifacts during tissue processing. As the reversible fixative targets proteins rather than nucleic acids, our approach ensures RNA integrity, library complexity, cellular composition, and gene expression comparable to those of the gold-standard fresh RNA sequencing assays. With FixNCut, time and location constraints for sample collection and processing are removed, making it an ideal strategy for research studies involving multiple groups, institutions, and hospitals worldwide.

Standardizing protocols for clinical biopsy collection and downstream processing poses a significant challenge. While single-cell profiling technologies have proven to be highly useful and straightforward for PBMCs and other cells loosely retained in secondary lymphoid tissues, the effective isolation of single cells from solid tissues, such as tumors, remains a technical hurdle. In such cases, single cells may be tightly bound to extracellular scaffolds and neighboring cells, making dissociation and isolation difficult [24]. Additionally, preserving the single-cell transcriptome before scRNA-seq is crucial, particularly when processing multiple samples from biological replicates simultaneously, as it can reduce the need for immediate time-consuming single-cell isolation protocols, such as dissociation, antibody staining, or FACS-based isolation [24]. Recent studies have highlighted the impact of collection and dissociation protocols on cell type proportions and transcriptome profiles in multiple tissue contexts [2–4]. Furthermore, dissociation-related effects have also been observed in cryopreserved human gut mucosa samples [25] and renal biopsies [26]. To overcome these issues, a one-step collagenase protocol was used for intestinal biopsies when no cell type enrichment was required [25]. Meanwhile, the use of cold-active proteases on kidney samples resulted in fewer artifacts, but inefficient tissue dissociation [26]. In addition, studies have shown that neural cell populations (NCPs) suspension without methanol preservation also experiences alterations in cellular composition and gene expression [27]. For fragile tissue biopsies, such as the pancreas or skin, which contain delicate cell populations, cryopreservation and cellular dissociation steps may introduce biases on the cellular composition. We

previously conducted testing of VivoFix [28] with variable results, but acknowledge the potential utility of the protocol for fixation and dissociation.

To address these challenges, we developed FixNCut, an approach that involves reversible fixation of the tissue at the time of collection to prevent further transcriptional changes during downstream processing. This is followed by standard dissociation and storage procedures. At the core of our approach is the use of Lomant's reagent/DSP, a reversible fixative that can easily penetrate cell membranes and preserve tissue characteristics. In this study, we compared fresh and fixed lung and colon samples from different species and experimental scenarios and found comparable results. Additionally, we demonstrated the versatility of the FixNCut for long-term storage by cryopreservation following sample fixation, making it a suitable protocol for its use in more complex and challenging research scenarios. While the lung represents a more resilient tissue for sample processing, without the introduction of major changes in gene expression or cellular composition, colon tissue is very sensitive. Here, we demonstrated decreased RNA quality and a shift in cellular composition in fresh compared to fixed samples, even under standard experimental conditions. We predict that under more stressful conditions, such as therapeutic intervention models, these biases will be even more pronounced.

In single-cell analysis, antibodies are commonly used to select cells of interest by standard FACS enrichment or to quantify cell surface proteins using sequencing (e.g., CITE-Seq). Unlike other fixative agents, such as methanol, which induces protein unfolding and precipitation, or formalin, which non-selectively cross-links proteins, DNA, and RNA (reducing immunoreactivity with target-specific antibodies), the FixNCut protocol has the advantage that cells can be readily stained with antibodies and LMOs. This allows cell labeling or hashing before single-cell analysis. Hence, we assume that this protocol is adaptable and compatible with multiple single-cell modalities, including but not limited to CITE-seq, as well as other droplet or microwell platforms. This versatility makes it a powerful tool for designing flexible and robust studies, being applicable to different tissues, species, or disease conditions.

The FixNCut protocol offers a straightforward way to preserve biopsies for various research contexts, including animal models at research institutes and patient biopsies collected at hospitals. We have demonstrated that FixNCut can be applied in clinical settings, where samples are collected at separate locations and time from their downstream processing steps. The FixNCut protocol shows potential compatibility with multiple single-cell and spatial applications for both single-cell and single-nuclei sequencing, making it a promising versatile tool in various basic, translation, and clinical areas (e.g., oncology and autoimmunity). However, further validation efforts have to reinforce its utility in these diverse applications.

Conclusions

We demonstrate the FixNCut protocol to preserve the transcriptional profile of single cells and the cellular composition of complex tissues. The protocol enables the disconnection of sampling time and location from subsequent processing sites, particularly important in clinical settings. The protocol further prevents sample processing artifacts by stabilizing cellular transcriptomes, enabling robust and flexible study designs for collaborative research and personalized medicine.

Methods

Sample collection

Human PBMC isolation

Peripheral venous blood samples were collected from voluntary blood donors using ACD tubes and stored at 4°C. PBMC isolation was performed using Ficoll density gradient centrifugation. Briefly, 10 mL of blood were diluted with an equal volume of 1× PBS and carefully layered onto 15 mL of Lymphoprep (PN. 15257179, STEMCELL Technologies) followed by centrifugation for 20 min at 800×*g* and room temperature (RT) (with acceleration and brake off). After centrifugation, PBMCs were collected with a sterile Pasteur pipette, transferred to a 15-mL tube, and washed twice with 10 mL of 1× PBS by centrifugation for 5 min at 500×*g* at RT. PBMCs were resuspended in 1× PBS + 0.05% BSA, and cell number and viability were measured with LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem).

Mouse lung collection

C57BL/6 mice were purchased from Janvier Laboratories at 6 weeks of age and sacrificed between weeks 7 and 9 by CO_2 asphyxiation. Lung samples were perfused prior to collection. To perfuse the lungs, a 26-G syringe was used to inject 3 mL of cold Hank's Balanced Salt Solution (HBSS) into the right ventricle of the heart, which resulted in the lungs turning white after injection. Mice were then carefully dissected for further processing.

Mouse colon collection

Mice were sacrificed using as described above, and the colon was collected and washed with HBSS using a syringe to remove feces. The collected colon samples were transported from the facility to the lab in a complete DMEM medium on ice. Upon arrival, the samples were extensively washed with ice-cold PBS and then cut into 3 × 3 mm pieces on a Petri dish using a sterile razor blade. The tissue pieces were then fixed as previously described.

Human colon biopsies

Colonic biopsies were collected from an ulcerative colitis patient in remission and placed in HBSS (Gibco, MA, USA) until processing, which was completed within an hour. The biopsies were split into four different conditions: fresh, fixed, cryopreserved, and fixed/cryopreserved. For fixation, the biopsies were treated as previously described for mouse lung tissue.

Human prostate tissue

Human prostate tissue was collected with informed consent from a 75-year-old patient who underwent a radical prostatectomy procedure for prostate cancer.

Sample preparation: fixation and cryopreservation

Preparation of DSP fixation buffer

A 50× stock solution of DSP (50 mg/mL) was prepared in 100% anhydrous DMSO and stored at −80 °C. Prior to use, 10 µL of the 50× DSP was added dropwise to 490 µL of RT

PBS in a 1.5-mL tube while vortexing to prevent DSP precipitation. This working solution (1× DSP fixation buffer) was then filtered through a 40-µm Flowmi Cell Strainer (PN. BAH136800040-50EA, Sigma-Aldrich). Table 1 provides detailed instructions for using the FixNCut protocol with both the DSP stock solution and working dilution.

PBMCs fixation

One million cells were split into two separate 1.5-mL tubes, with one tube used fresh (as a non-fixed control sample) while the other was subjected to cell fixation. For fixation, cells were centrifuged at 500×*g* for 5 min at 4 °C, and the resulting pellet was resuspended with 500 µL of 1× DSP fixation buffer and incubated at RT. After 15 min, the cells were mixed by pipetting and incubated for an additional 15 min. Fixation was stopped by adding 10 µL of 1 M Tris-HCl pH 7.4, and the sample was briefly vortexed and incubated at RT for 5 min. Both fresh and fixed samples were centrifuged for 5 min at 500×*g* at 4 °C, and contaminating erythrocytes were eliminated by resuspending the pellets and incubating at RT for 5 min with 1× Red Blood Cell lysis solution (PN. 130-094-183, Miltenyi Biotec). Both samples, fresh and fixed, were centrifuged for 5 min at 500×*g* at 4 °C, and contaminating erythrocytes were eliminated by resuspending the pellets in 500 µL of PBS and incubating at RT for 5 min upon addition of 10 times volume of 1× Red Blood Cell lysis solution (PN. 130-094-183, Miltenyi Biotec). Cells were then resuspended in an appropriate volume of 1× PBS + 0.05% BSA in order to reach optimal concentration for cell encapsulation (700–1000 cells/µL) and filtered using a pluriStrainer Mini 40 µm (PN. 43-10040-70 Cell Strainer). Cell concentration was verified with LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem).

Cryopreservation

Cryopreservation of fresh or fixed biopsies was done by transferring them into 1 mL of freezing media (90% FBS + 10% DMSO, Thermo Scientific) and storing them at −80 °C in a Mr. Frosty™ Freezing Container to ensure gradual freezing.

Table 1 Recommendations on working with DSP stock and working dilution, as described by Attar et al. [16]

Preparation of 50× DSP stock
• Equilibrate DSP vial at RT for 30 min and then prepare a 50× stock solution of DSP (50 mg/mL) in anhydrous dimethyl sulfoxide (Sigma, Cat. N. 276855-100ML).
• Dispense the stock into single-use aliquots (e.g., 100 µL aliquots, but volume depends on your use) and store in a bag and dry environment (with silica/desiccant if possible) at −80 °C.
• Be mindful of not opening and closing the tubes at −80 °C.
Preparation of DSP working dilution
• Thaw the 50× DSP stock reagent from −80 °C and equilibrate at RT no longer than 10 min before fixation.
• Immediately before use prepare 500 µL of 1× DSP working solution in molecular biology grade 1× PBS as follows: aliquot 10 µL of 50× DSP stock reagent in a 1.5-mL tube and while vortexing (VERY IMPORTANT) add 490 µL of PBS dropwise using a P200 pipette. 1× DSP must be used within 5 min of preparation.
• *Note: Do not prepare larger volumes (e.g., if you need to fix two samples, prepare each aliquot of 500 µL separately, DO NOT prepare 1 mL and then aliquot into two tubes). You should notice some thin white rings on the walls of the tube once diluted the 50× stock. This is expected and will be cleared during filtration. Stronger precipitation indicates insufficient solving of DSP and preparation of a new dilution is strongly recommended.*
• Filter the 1× DSP working dilution using a 40-µm Flowmi strainer (Sigma, cat. no. BAH136800040-50EA). After filtration 1× DSP can be pooled.
• Do not re-freeze the leftovers of the 50× DSP. Always use a freshly thawed aliquot to prepare the 1× working solution.

Single-cell RNA-seq experimental design (scRNA-seq)

Human PBMC 3′ scRNA-seq

Cells from both fresh and fixed PBMCs were processed for single-cell RNA sequencing using the Chromium Controller system (10X Genomics), with a target recovery of 8000 total cells. The Next GEM Single Cell 3′ Reagent Kits v3.1 (PN-1000268, 10X Genomics) were used to prepare cDNA sequencing libraries, following the manufacturer's instructions. Briefly, after GEM-RT clean-up, cDNA was amplified for 11 cycles and then subjected to quality control and quantification using an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). The Dual Index Kit TT Set A (PN-1000215, 10X Genomics) was used for indexing cDNA libraries by PCR. The size distribution and concentration of 3′ cDNA libraries were verified again on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). The cDNA libraries were sequenced on an Illumina NovaSeq 6000 using the following sequencing conditions: 28 bp (Read 1) + 8 bp (i7 index) + 0 bp (i5 index) + 89 bp (Read 2) to generate approximately 40,000 read pairs per cell.

Mouse lung cryopreservation, fixation, and cryopreservation upon fixation

Mouse lungs were harvested and transferred into ice-cold complete DMEM medium. Samples were extensively washed with ice-cold PBS, transferred into a Petri dish, and cut into ~3 × 3 mm pieces using a razor blade. Tissue pieces were divided into four tubes for each condition. Tissue pieces in tube 1 were cryopreserved in freezing media (50% DMEM + 40% FBS + 10% DMSO) and placed into a Mr. Frosty™ Freezing Container and transferred to a −80 °C freezer to ensure a gradual freezing process. The tissue pieces in tubes 2 and 3 were triturated with a razor blade (~1 × 1 mm) on ice and fixed by submerging them in 500 µL of 1X DSP fixation buffer (freshly prepared, within 5 min of use) and incubating at RT for 30 min. After incubation, 10 µL of 1 M Tris-HCl pH 7.5 was added to stop the fixation, and the samples were vortexed for 2–3 s and incubated at RT for 5 min. After a brief centrifugation, the supernatant was removed, and the tissue pieces were washed once with 1 mL of PBS. Tissue pieces in tube 2 were stored at 4 °C in PBS supplemented with 2 U/µL of RNAse inhibitor (Cat. N. 3335402001, Sigma-Aldrich) until the following day, while tissue pieces in tube 3 were cryopreserved by adding 10% DMSO to the PBS, transferring the cell suspension into a cryotube, which was placed into a Mr. Frosty™ Freezing Container, and stored at −80 °C. The tube containing fresh tissue was stored in complete DMEM on ice and washed again with ice-cold PBS before tissue dissociation.

Mouse lung dissociation and scRNA-seq

The day after the collection and storage of samples, the cryopreserved and fixed/cryopreserved samples were quickly thawed in a 37 °C water bath and washed with PBS. Similarly, the fixed sample was washed with PBS before tissue dissociation. The samples were then transferred to a Petri dish on ice and triturated using a razor blade. Next, the small tissue pieces were incubated in 1 mL of digestion media (200 µg/mL Liberase TL, 100 µg/mL DNase I in HBSS with $Ca^{2+}Mg^{2+}$) at 37 °C with shaking at 800 rpm. After 15 min of incubation, the samples were mixed by pipetting, followed by another 15 min of incubation. The cells were then filtered using a pluriStrainer Mini 70 µm (PN. 43-10070-40 Cell Strainer), and the strainer was washed with 10 mL of cold 1X HBSS. The samples

were centrifuged at 500×*g* for 5 min at 4 °C, and the cell pellets were resuspended in 100 µL of PBS + 0.05% BSA. Contaminating erythrocytes were lysed using the previously described method. The cells were washed once with 10 mL PBS + 0.05% BSA, resuspended in an appropriate volume of the same buffer, and filtered using 40-µm strainers. The total number of cells was determined using the LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem). The cell concentration of each sample was adjusted to 700–1000 cells/µL and 7000–10,000 cells were loaded into a 10X Chromium controller. Single-cell RNA-sequencing was performed as described above.

Mouse colon dissociation

Fresh and fixed colon samples were incubated in 1 mL of digestion media (200 U/mL Collagenase IV, 100 µg/mL DNase I in HBSS w/$Ca^{2+}Mg^{2+}$) at 37 °C, shaking at 800 rpm for 30 min. Samples were mixed by pipetting every 10 min during incubation. After the incubation, samples were filtered through a pluriStrainer Mini 70 µm (PN. 43-10070-40 Cell Strainer) and the strainer was washed with 10 mL of cold 1X HBSS. The samples were then centrifuged at 500×*g* for 5 min at 4 °C and cell pellets were washed twice with cold PBS + 0.05% BSA. Finally, the cell pellets were resuspended in an appropriate volume of the same buffer and filtered using 40-µm strainers. The total cell number was determined with the LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem). The cell concentration of each sample was adjusted to 700–1000 cells/µL and 7000 cells were loaded onto a 10X Chromium controller. Single-cell RNA-seq was performed as described above.

Human colon biopsies dissociation

Digestion of biopsies to single-cell suspension was achieved through mechanical and enzymatic digestion as previously described by Veny et al. [29]. Briefly, biopsies were washed in complete medium (CM) (RPMI 1640 medium (Lonza, MD, USA) supplemented with 10% FBS (Biosera, France), 100 U/mL penicillin, 100 U/mL streptomycin and 250 ng/mL amphotericin B (Lonza), 10 µg/mL gentamicin sulfate (Lonza), and 1.5 mM HEPES (Lonza)) before being triturated and incubated in 500 µL of digestion media (CM supplemented with Liberase TM (0.5 Wünsch U/mL) (Roche, Spain) + DNase I (10 µg/mL) (Roche, Spain)) with agitation (250 rpm) for 1 h at 37 °C. After incubation, the cell suspension was filtered using 50-µm and 30-µm cell strainers (CellTrics, Sysmex, USA) to remove cell aggregates and debris. Cell viability and concentration were determined with the LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem). Approximately 7000 cells were loaded onto the Chromium controller (10x Genomics, CA, USA).

Flow cytometric analysis

Flow cytometry analysis of human PBMCs

Anti-human CD3, CD4, CD8, and CD19 antibodies were tested as follows: Cryopreserved PBMCs obtained from three healthy donors were rapidly thawed in a 37 °C water bath. Thawed samples were washed in pre-warmed RPMI media supplemented

with 10% FBS (Thermo Fisher Scientific) and centrifuged at 500×*g* for 5 min at RT. The supernatant was discarded, and pellets were washed in 10 mL of 1× PBS + 0.05% BSA, centrifuged at 500×*g* for 5 min at 4 °C, and resuspended in 1 mL of PBS + 0.05% BSA. The cell suspension was then filtered through a 40-µm cell strainer. Cell viability and concentration were verified using a LUNA-FL™ Dual Fluorescence Cell Counter (LogosBiosystem). Each sample was split into two separate tubes, and half of the cells were fixed with 1× DSP fixation buffer as previously described. After fixation, cells were washed and resuspended in 100 µL of Cell Staining Buffer (PN-420201, Biolegend) and stained with 5 µL of each of the four following primary antibodies: anti-human CD3, CD4, CD8, and CD19 antibody for 15 min at RT in the dark. Detailed information on the antibodies and reagents used in this study is provided in Table 2. Samples were washed twice with Cell Staining Buffer and resuspended in 0.5–1 mL of PBS + 0.05% BSA. 10 µg/mL DAPI (PN-564907, BD Bioscience) was added to determine cell viability before flow cytometric analysis using the BD FACS Melody Automated Cell Sorter (BD Bioscience) and the BD FACSChorus™ Software. Post-acquisition analysis was performed using FlowJo version 10 (FlowJo LLC).

Anti-human CD45, CD3, CD4, and CD8 antibodies and anti-CD298 and β2-microglobulin antibodies were tested as follows: Human PBMCs were isolated from normal donor human buffy coats provided by the Australian Red Cross Blood Service by Ficoll-Paque® density gradient centrifugation. Fresh and fixed PBMCs were incubated with Human BD Fc Block for 10 min at 4°C and then stained for cell surface markers for 30 min at 4 °C according to the manufacturer's recommendation. Annexin V-FITC and PI staining were used to determine viability. Detailed information on the antibodies and reagents used in this study is provided in Table 2. The acquisition was performed on LSR Fortessa X 20 (BD) and analyzed via FlowJo software (FlowJo LLC).

Table 2 Flow cytometry antibodies and reagents

Antibody	Clone	Conjugate	Company
Human BD Fc Block	Fc1	NA	BD # 564219
Anti-human CD45	HI30	BV510	BD # 563204
	HI30	PE	BD Pharmingen
Anti-human CD3	UCHT1	BV711	BD # 563725
	SK7	PE-Cy7	Biolegend PN-344816
	OKT3	PerCP	Biolegend
Anti-human CD4	SK3	BV650	BD # 563875
	RPA-T4	Alexa Fluor® 488	Biolegend PN-300519
Anti-human CD8	HIT8a	PE	BD #555635
	SK1	APC-Cy7	Biolegend PN-344714
Anti-human CD19	6E10	Alexa Fluor® 647	Biolegend PN-302222
Anti-human CD298	LNH-94	PE-Cy7	BioLegend #341707
Anti-human β2-microglobulin	2M2	APC	BioLegend #316311
Anti-human CD11b	ICRF44	PeCy7	Biolegend
Anti-human EPCAM	9C4	APC	Biolegend
Annexin V		FITC	BD #556420
Propidium Iodine			Sigma-Aldrich #P4864

LMO sample preparation

The labeling of PBMCs was performed following the protocol previously described by McGinnis et al. [30]. Briefly, 5 × 105 fresh and fixed PBMCs were washed twice with PBS and labeled with a 1:1 molar ratio of anchor LMO and barcode oligonucleotide for 5 min on ice. Subsequently, both samples were incubated with a co-anchor and Alexa 647 fluorescent oligo feature barcodes at concentrations of 200 nM and 400 nM, respectively, for another 5 min on ice. The cells were then washed twice with ice-cold 1% BSA in PBS. Acquisition was performed on the LSR Fortessa X 20 (BD) and analysis was carried out using FlowJo software (FlowJo LLC). Detailed information on the LMOs used in this study is provided in Table 3.

Flow cytometry analysis of human colon biopsy

The single-cell suspensions obtained after biopsy digestion were labeled with the following antibodies: anti-CD45, anti-CD3, anti-CD11b, and anti-EPCAM, according to the manufacturer's instructions. Detailed information on the antibodies used in this study can be found in Table 2. Cell viability was assessed using the Zombie Aqua Fixable Viability Kit (BioLegend). The cells were then fixed with the Stabilizing Fixative (BD) before being analyzed using the FACSCanto II flow cytometer (BD).

Multiplexing fluorescence tissue imaging

Tissue preparation

The 2 cm × 10 mm tissue sample was divided into two equal halves lengthwise. One half was fixed in 2 ml of 10% neutral buffered formalin (NBF), while the other half was placed in 500 µL of 1X DSP fixation buffer. The NBF-fixed sample was incubated at RT for 4 h, stored overnight at 4 °C, washed 3 times with 1 mL of milliQ water, and then stored in 1 mL of 70% ethanol at 4 °C. The DSP-fixed sample was treated with freshly-made DSP fixation buffer, which was replaced every 60 min for 4 h. The fixed sample was then neutralized with 10 µL of 1M Tris-HCl pH 7.4 for 15 min at RT, washed 3 times with 1 mL of milliQ water, and placed in 1 mL of 70% ethanol at 4°C. Both NBF- and DSP-fixed samples were embedded in paraffin overnight. Five-micrometer-thick sections were cut from both the formalin-fixed, paraffin-embedded (FFPE) and the DSP-fixed paraffin-embedded (DSP-PE) tissues and mounted onto a single poly-L-lysine coated coverslip (22 × 22 mm, #1.5, Akoya Bioscience #7000005).

Table 3 Flow cytometry lipid-modified oligos (LMOs)

Reagent	Sequence	Company
LMO001A Lignoceric Anchor with DNA Oligo	Commercial source	Sigma-Aldrich
LMO001B Palmitic Co-anchor with DNA Oligo	Commercial source	Sigma-Aldrich
A647_Oligos	5′-CCTTGGCACCCGAGAATTCCA-A_647-3 (complementary to the Anchor) or /5Alex647N/CCTTAGCCGCTAATAGGTGAGC (Complementary to Capture sequence of Barcoded Oligo CCTTGGCACCCG AGAATTCCA[Barcode]GCTCACCTATTAGCGGCTAAGG)	Oligo from IDT

Antibody staining

The coverslip mounted section was baked at 60 °C on a heat block for 1 h to remove paraffin, then deparaffinized in 1× Histo-Choice clearing agent (ProSciTech #EMS64110-01) and rehydrated in ethanol before washing in milliQ water. Antigen retrieval was performed in a pressure cooker on the highest setting for 20 min in 1X citrate buffer, pH 6 (Sigma, #C9999-1000ML). The tissue was blocked using buffers from the commercially available Phenocycler staining kit (Akoya Bioscience, # 7000008) and stained with a 7-antibody panel at RT for 3 h. Detailed information on the antibodies used in this study can be found in Table 4. The antibodies were used at a dilution of 0.9:200 for commercially available Akoya antibody-oligo conjugates and 3.7:200 for antibodies custom-conjugated by Akoya Bioscience (Spatial Tissue Exploration Program (STEP)). After staining, the coverslip was subjected to a post-fixation step. DAPI staining was used to visualize cell nuclei and locate regions of interest on each tissue sample with the Zeiss Axio Observer 7 fluorescent inverted microscope.

Phenocycler image acquisition

The Phenocycler microfluidic microscope stage was programmed to acquire two 3 × 3 tiled regions on each tissue using a 20× objective lens, with each tile consisting of a 7-image Z-stack illuminated by LED light to specifically excite either DAPI (for 10 ms in all cycles) or one of three fluorescently labeled reporter oligos (Cy3, Cy5, and Cy7). The software was set to acquire images over 5 cycles, with each cycle consisting of the addition of a set of reporter oligos complementary to the antibody-oligo conjugates detailed in Table 4. During the first and last cycles, no reporter oligos were added to allow for background fluorescence subtraction. The exposure times for each antibody are also provided in Table 4. After imaging was completed, the sample was manually stained with Hematoxylin and Eosin (H&E) following the UofA histology protocol, and bright-field images were captured using the Zeiss Axio Observer 5 fluorescent inverted microscope.

Image processing

The acquired images were processed using the Phenocycler CODEX Processor software (Akoya version 1.8.3.14) to deconvolve and stitch them together, resulting in a set of multi-channel QPTIFF files for each region. The levels for each channel were adjusted using QuPath 0.4.0 and the final images were saved as RGB tiff files. However, some antibodies did not produce sufficient signal or acceptable images after processing and were therefore excluded from further analysis.

Table 4 Antibody and reporter cycle layout, including LED intensities and exposure times for each marker

Cycle #	Cy3 (LED 100%)	Exposure time (ms)	Cy5 (LED 100%)	Exposure time (ms)	Cy7 (LED 60%)	Exposure time (ms)
1	BLANK	250	BLANK	350	BLANK	500
2	e-cadherin-RX014	750	ERG-RX031	850	SMA-RX013	500
3	CD8-RX026	750	CD3e-RX045	850	PanCK-RX019	400
4	Empty	250	p63-RX046	700	Empty	500
5	BLANK	250	BLANK	350	BLANK	500

Data processing

To profile the cellular transcriptome, we processed the sequencing reads using the CellRanger software package (version 6.1.1) from 10X Genomics Inc. We mapped the reads against either the mouse mm10 or the human GRCh38 reference genome (GENCODE v32/Ensembl 98), depending on the samples. In order to avoid any artifacts on the downstream analysis due to differences in sequencing depth among samples, we normalized libraries for effective sequencing depth using "cellranger aggr". This subsampling approach ensures that all libraries have the same average number of filtered reads mapped confidently to the transcriptome per cell.

Data analysis

All analyses presented in this manuscript were conducted using R version 4.0.5, along with specific analysis and data visualization packages. For scRNA-seq analysis, we used the Seurat R package (version 4.0.0) [31], SeuratObject package (version 4.0.1), and other packages specified in the subsequent sections.

scRNA-seq quality control

To compare the library complexity (total captured genes and Unique Molecular Identifiers, or UMIs) across libraries, we investigated the relationship between the cumulative number of detected genes and UMIs with the library sequencing depth. To achieve this, we loaded the "molecule_info.h5" information using the function "read10xMolInfo" from the DropletUtils package (version 1.10.3). Then, we downsampled the library sequenced reads assigned to a barcode (excluding background and noisy reads) using the function "downsample_run", implemented in *Rcpp*, which ensures read sampling without replacement and simultaneously updates the sampling frequency. We utilized various depths for downsampling (steps of 5 M or 10 M reads, depending on the library), which emulates differences in sequencing depth per cell. To assess differences in library complexity along sequencing depth between protocols under study, we fitted a linear regression model ($Y \sim X$) to each curve. Then, we compared the confidence intervals (95% CI) of the independent variable (X, "sequenced reads") across libraries, considering the differences statistically significant when the confidence intervals between conditions did not overlap. Moreover, we assessed the distribution of cell complexity (total captured genes and UMIs) per cell sequencing depth across libraries and computed a linear model to compare the slope of the regression line for each different library. Ultimately, we computed the cumulative number of detected genes over multiple cells by averaging the total genes after 100 sampling of an increasing number of randomly sampled cells (from 1 to 100, using steps of 2), after running the "cellranger aggr" step described above.

To enhance our comprehension of the impact of DSP-fixation on gene recovery, we conducted an analysis comparing the categories of all genes captured in each library—either in all conditions or uniquely in a specific condition of each comparison. For the mouse data, we employed the complete list of gene annotations sourced from the Mouse Genome Informatics website (https://www.informatics.jax.org), while for the human data, we utilized the comprehensive list of gene annotations obtained from the HUGO Gene Nomenclature Committee (HGNC) (https://www.genenames.org/).

After ensuring that there were no remarkable differences in the main quality control (QC) metrics (library size, library complexity, percentage of mitochondrial and ribosomal expression) among the different samples, we performed an independent QC, normalization, and analysis for the libraries from different species and tissue, following the guidelines provided by Luecken et al. [32] and Heumos et al. [33]. We removed low-quality cells by filtering out barcodes with a very low number of UMIs and genes, or with a high percentage of mitochondrial expression, as it is indicative of lysed cells. Additionally, we considered removing barcodes with a large library size and complexity. We eliminated genes that were detected in very few cells. Notably, due to the inherent characteristics of colon biopsies (a higher number of epithelial cells, which are less resistant to sample processing), we followed a slightly different QC approach for mouse and human colon samples. In brief, we performed a first permissive QC filtering out cells with very high MT% (> 75% for mouse and > 85% for human) before proceeding to downstream analysis. We annotated cells to distinguish between the epithelial and non-epithelial fraction. Then, we repeated the QC step, using different thresholds for the epithelial fraction (> 60% MT) and for the non-epithelial cells (> 50% for mouse and > 25% for human). Finally, data was normalized and log-transformed.

Doublet was predicted with Scrublet [34] (version 0.2.3), and computed doublet scores were retained, with putative doublet cell barcodes flagged. However, no threshold was applied to filter them out at this stage, adopting a permissive approach. Consequently, during the clustering and annotation step, the clusters showing modified QC metrics, along with the co-expression of different lineage/population-specific gene markers and high doublet scores, were assessed to determine whether a specific cluster could be classified as a group of doublets and subsequently excluded.

Cell clustering and annotation

To achieve successful cell-type annotation combining data from the same tissue and species (mouse and human colon samples), we removed the batch effect with the *Harmony* integration method [35] using the library as a confounder variable. After integration, we created a k-nearest neighbors (KNN) graph with the "FindNeighbors" function using the first 20 Principal Components (PC), followed by the cell clustering with the Louvain clustering algorithm using the "FindClusters" function at different resolutions. To visualize our data in a two-dimensional embedding, we run the Uniform Manifold Approximation and Projection (UMAP) algorithm. Then, we performed a differential expression analysis (DEA) for all clusters to determine their marker genes using the normalized RNA counts. To annotate the clusters into specific cell types, we examined the expression of canonical gene markers, compared the results of the DE analysis, and referred to gene markers from published annotated datasets. We used the following datasets as references: human PBMCs based on Stuart et al. [36]; mouse lung based on Angelidis et al. [37], Zhang et al. [38], and Bain and MacDonald [39]; mouse colon based on Tsang et al. [40]; and human colon based on Garrido-Trigo et al. [41]. Furthermore, we performed specific cell-type sub-clustering when required a fine-grained resolution to capture a specific cell state of interest. Doublets and low-quality cells were automatically removed at this point.

Comparison of gene expression profiles

Gene expression correlation analysis

The overall similarity of pseudo-bulk gene expression profiles across experimental conditions was explored using the Pearson correlation (r^2). To generate the pseudo-bulk profiles, we computed the mean of the log-average gene counts across all cells per sample, as well as for each defined cell population independently. Due to the sparsity of scRNA-seq data, we only computed correlation for the cell types with over 100 cells and only if over 20 cells were present by condition to compare. Additionally, to avoid biases driven by cell populations with different sizes, we randomly downsampled the number of cells to the condition having fewer cells. To assess the strength and significance of the correlation, we used linear regression models; we considered a strong significant linear correlation when $r^2 > 0.9$ and p-value < 0.05.

Compositional analysis

To estimate the changes in the proportions of cell populations across experimental conditions, we applied the *scCODA* package (version 0.1.2) [42], a Bayesian approach that considers inherent biases in cell-type compositions and addresses the low-replicate issue, making it an appropriate tool for our experimental design. This compositional analysis depends on a reference cell type, and credible changes (nominal FDR < 0.05) for varying log-fold changes should be interpreted in relation to the selected reference cell type. For this reason, we used the "automatic" selection of reference cell type for single comparisons, but we set a manual "pre-selected reference" cell type when multiple comparisons were conducted against the same condition. In the latter case, we tested multiple reference cells to ensure consistent results.

Differential expression analysis (DEA)

To define condition-specific signatures, we performed differential expression analysis (DEA) using the Seurat function "*FindMarkers*" with the MAST test. We defined genes to be significantly differentially expressed (sDE) if the Log2 Fold Change (Log2FC) > |1|, with a false discovery rate (FDR) adjusted p-value < 0.05, and if they were present in at least 10% of cells.

Gene Set Enrichment Analysis (GSEA)

To unravel biological pathways affected by a specific condition, we performed a pre-ranked GSEA with the *fgsea* package [43] and employed multiple gene sets from the Human Molecular Signatures Database (MSigDB), specifically, the mouse and human Hallmarks (H), Canonical Curated Pathways (C2:CP) such as Reactome, and Gene Ontology (C5:GO) gene sets from Biological Process (BP). We excluded gene sets that did not satisfy our pathway size criteria (10–300 genes) and considered significant only results with an FDR adjusted p-value < 0.05, where > 5 genes overlapped with the gene set, and this accounted for > 15% of the gene set.

Condition-specific signatures

To evaluate the effectiveness of the FixNCut protocol in preventing cells from undergoing stress, we assessed condition-specific signatures and sDE genes associated with external stressors. For human PBMC samples, we included publicly available signatures, such as the ex vivo PBMC handling signature on microarrays by Baechler et al. [44] and the human PBMC sampling time-dependent signature on single-cell by Massoni-Badosa et al. [4]. Additionally, we studied multiple dissociation-induced gene expression signatures for the tissue samples, including dissociation on mouse muscle stem cells by van den Brink et al. [22], warm-dissociation on mouse kidney samples by Denisenko et al. [3], and warm collagenase on mouse primary tumors and patient-derived mouse xenografts by O'Flanagan et al. [2]. With this purpose, we downloaded the sDE genes from these studies and computed signature-specific scores using the *Ucell* package [45]. To account for biases arising from differences in cell population numbers across protocols, we downsampled at a maximum of 250 cells per cell type. Moreover, we also investigated the expression of specific genes across cell populations and conditions. Statistical analysis between sample protocols was performed using the Wilcoxon signed-rank test.

Abbreviations

ACME	ACetic-MEthanol
CM	Complete Media
DEA	Differential expression analysis
DSP	Dithiobis(succinimidyl propionate)
DSP-PE	DSP-fixed, paraffin-embedded
FFPE	Formalin-fixed, paraffin-embedded
GEM	Gel Beads-in-emulsion
GSEA	Gene Set Enrichment Analysis
H&E	Hematoxylin and eosin
HBSS	Hank's Balanced Salt Solution
HVG	Highly variable gene
IBD	Inflammatory bowel disease
LMO	Lipid-modified oligo
mAbs	Monoclonal antibodies
ME	Mean expression
MFI	Mean fluorescent intensity
NCP	Neural cell populations
NHS ester	N-Hydroxysuccinimide ester
PBMC	Peripheral blood mononuclear cell
PC	Principal components
QC	Quality control
ROS	Reactive oxygen species
RT	Room temperature
STEP	Spatial Tissue Exploration Program
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
UPR	Unfolded protein response

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03219-5.

Additional file 1: Supplementary Figures. *Supplementary Figures 1-6 and Supplementary figure legends.*

Additional file 2: Table 1. This table contains the differentially expressed (DE) genes between fixed and fresh human peripheral blood mononuclear cells (PBMCs), considering all cell populations. The differential expression analysis was performed using Seurat MAST, and genes with an FDR adjusted p-value < 0.05 and present in at least 10% of cells were considered DE (see Methods). Genes with positive Log2FC indicate higher expression in fixed PBMCs, while those with negative Log2FC are DE in fresh PBMCs.

Additional file 3: Table 2. This table contains differentially expressed (DE) genes between fixed and fresh human PBMCs, separated by cell population. The analysis was performed using Seurat MAST, with genes considered as DE with a FDR adjusted p-value < 0.05 and present in at least 10% of cells (see Methods). A positive Log2FC value indicates higher expression in fixed PBMCs, while a negative value indicates differential expression in fresh PBMCs.

Additional file 4: Table 3. This table contains differentially expressed (DE) genes for mouse tissue comparisons, considering all cell populations. The comparisons include: 1) fixed versus fresh lung samples, 2) fixed versus fresh colon samples, 3) fixed compared to cryopreserved (cryo), 4) fixed/cryo versus cryo, and 5) fixed/cryo compared with fixed on lung samples. The statistical analysis was performed using Seurat MAST with genes considered DE if they had an FDR adjusted p-value < 0.05 and were present in at least 10% of cells (see Methods). The results include positive Log2FC values indicating higher expression in the first condition mentioned, while negative values indicate higher expression in the second condition mentioned.

Additional file 5: Table 4. This table contains the list of differentially expressed (DE) genes between fixed and fresh mouse lung samples, categorized by cell population. The test was performed using Seurat MAST, where genes with a FDR adjusted p-value < 0.05 and present in at least 10% of cells were considered as DE (see Methods). The Log2FC values indicate the direction of expression change, where positive values indicate higher expression in fixed samples and negative values indicate higher expression in fresh samples.

Additional file 6: Table 5. This table contains the results of gene set enrichment analysis (GSEA) between fixed and fresh mouse lung samples by cell population. The analysis was performed using the *fgsea* package and multiple gene sets from MSigDB. Gene sets containing 10-300 genes were included, and only results with a FDR adjusted p-value < 0.05 were considered if they contained >5 overlapping genes, which represents >15% of the gene set (see Methods).

Additional file 7: Table 6. This table contains the differentially expressed (DE) genes between fixed and fresh mouse colon samples, analyzed by cell population using Seurat MAST. Genes were considered DE if they had an FDR adjusted p-value < 0.05 and were present in at least 10% of cells (see Methods). The Log2FC value indicates whether the gene is more highly expressed in fixed (positive values) or fresh (negative values) samples.

Additional file 8: Table 7. This table contains the results of gene set enrichment analysis (GSEA) between fixed and fresh mouse colon samples by cell population. The analysis was performed using the *fgsea* package and multiple gene sets from MSigDB. Gene sets containing 10-300 genes were included, and results were considered significant if they had an FDR adjusted p-value < 0.05 and if >5 genes from the gene set overlapped with the differentially expressed genes, representing more than 15% of the gene set (see Methods).

Additional file 9: Table 8. This table contains differentially expressed (DE) genes between fixed/cryo and cryo mouse lung samples, analyzed by cell population. The test was performed with Seurat MAST, and genes were considered DE with a FDR adjusted p-value < 0.05, and present in at least 10% of cells (see Methods). A positive Log2FC indicates higher expression in fixed or fixed/cryo samples, whereas negative values indicate differential expression in cryo or fixed samples.

Additional file 10: Table 9. This table contains the results of gene set enrichment analysis (GSEA) for fixed/cryo versus cryo mouse lung samples, by cell population. The analysis was performed using the *fgsea* package with multiple gene sets from MSigDB. Gene sets containing within 10-300 genes were included and the obtained results were considered significant if the FDR adjusted p-value was <0.05 and if >5 genes overlapped with the gene set, representing >15% of the gene set (see Methods).

Additional file 11: Table 10. This table contains differentially expressed (DE) genes for human colon biopsies comparisons, considering all cell populations. The comparisons include: 1) fixed and fresh, 2) fixed and cryo, 3) fixed/cryo versus cryo, and 4) fixed/cryo compared with fixed samples. The analysis was performed using Seurat MAST, and genes were considered DE with a FDR adjusted p-value < 0.05 and present in at least 10% of cells (see Methods). A positive Log2FC indicates higher expression in fixed or fixed/cryo samples, whereas negative values indicate differential expression in fresh, cryo or fixed samples.

Additional file 12: Table 11. This table contains the differentially expressed (DE) genes between fixed and fresh human colon biopsies, separated by cell population. The test was performed using Seurat MAST, with genes considered DE if they had a FDR adjusted p-value < 0.05 and were present in at least 10% of cells (see Methods). Positive Log2FC values indicate higher expression in fixed samples, while negative values indicate higher expression in fresh samples.

Additional file 13: Table 12. This table contains the differentially expressed (DE) genes between fixed/cryo versus cryo human colon biopsies, separated by cell population. The test was performed using Seurat MAST, with genes considered DE if they had a FDR adjusted p-value < 0.05 and were present in at least 10% of cells (see Methods). Positive Log2FC values indicate higher expression in fixed/cryo samples, while negative values indicate higher expression in cryo samples.

Additional file 14: Table 13. This table contains the results of gene set enrichment analysis (GSEA) for fixed versus fresh human colon biopsies comparison by cell population. The test was performed with the *fgsea* package using multiple gene sets from MSigDB. Gene sets were included only if they contained between 10-300 genes, and obtained results were considered with a FDR adjusted p-value < 0.05, only if more than 5 genes were overlapping with the gene set and represented more than 15% of the gene set (see Methods).

Additional file 15: Table 14. This table contains the results of gene set enrichment analysis (GSEA) for fixed/cryo versus cryo h uman colon biopsies comparison by cell population. The test was performed with the *fgsea* package using multiple gene sets from MSigDB. Gene sets were included only if they contained between 10-300 genes, and obtained results were considered with a FDR adjusted p-value < 0.05, only if more than 5 genes were overlapping with the gene set and represented more than 15% of the gene set (see Methods).

Additional file 16. Peer review history.

Acknowledgements
The LMO oligos used were kindly provided by Chris McGinnis from the Gartner lab (UCSF). The authors would like to thank the CRG Scientific IT Unit and the maintainers of the CNAG-CRG compute cluster for providing assistance with essential computing resources. Also, the authors would like to thank all the team members for their contributions during brainstorming and discussion sessions.

Review history
The review history is available as Additional file 16.

Disclaimer
Content of this publication reflects only the author's view and the JU is not responsible for any use that may be made of the information it contains.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
H.H. and L.G.M. designed the study. L.G.M. and J.P. devised the use of DSP for fixing tissues following dissociation. D.M. and G.C. performed the scRNA-seq experiments. L.J.-G. and J.C.N. annotated all datasets. L.J.-G. performed the computational analysis. D.M., J.C.N., S.R., E.M.-A., and V.G. performed and analyzed the FACS data. K.W. and K.B.J. performed the CODEX experiments. H.-M. performed all mice work. N.K.R., L.M.B., J.P.B., F.T., L.K.S., S.S., M.v.d.B., T.K, P.L.v.d.V., M.C.N., P.R., and A.S. provided clinical material. L.J.-G., D.M., and J.C.N. prepared the figures. L.J.-G, D.M, H.H., and L.G.M. wrote the manuscript with contributions from all the co-authors. All authors read and approved the current version of the manuscript.

Funding
This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (IMI 2 JU) under grant agreement No 831434 (3TR; Taxonomy, Targets, Treatment, and Remission). The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. Also, this project has received funding from the European Union's H2020 research and innovation program under grant agreement No 848028 (DoCTIS; Decision On Optimal Combinatorial Therapies In Imids Using Systems Approaches). L.J.-G. has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Universities. V.G. is supported by grant #2008-04050 from The Leona M. and Harry B. Helmsley Charitable Trust. E.M. is funded by grant RH042155 (RTI2018-096946-B-I00) from Ministerio de Ciencia e Innovacion. I.A.P holds a Victorian Cancer Agency Mid-Career Fellowship 2019. A.S. is funded by PID2021-123918OB-I00 from MCIN/AEI/51 10.13039/501100011033 and co-funded by "FEDER: A way to make Europe". Australian Prostate Cancer BioResource (APCB) collection in Adelaide is supported by funding from the South Australian Immunogenomics Cancer Institute and the South Australian Health and Medical Research Institute.

Availability of data and materials
The complete raw data (FASTQ files) and count matrices generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE229944 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE229944) [46]. The processed scRNA-seq datasets and metadata (R objects, Seurat) analyzed in the current study have been deposited at Zenodo (https://doi.org/10.5281/zenodo.10468192) [47].
The detailed DSP-fixation protocol can be accessed at Protocols.io (https://www.protocols.io/view/fixncut-v1-0-14egn3xjql5d/v1) [48].
The code to reproduce the full analysis is hosted at GitHub https://github.com/LJimenezGracia/FIXnCUT_benchmarking [49].

Declarations

Ethics approval and consent to participate
Experiments with mice were conducted in accordance with the guidelines of the Animal Care and Use Committee of Barcelona Science Park (protocol CEEA-PCB-14-000275). Healthy PBMC donors were collected with written informed consent at the Centro Nacional de Análisis Genómico (CNAG). Healthy human buffy coats were provided by the Australian Red Cross Blood Service, and the ethical approval was obtained by the Melbourne Human Ethics Committee (01/14). Colon biopsies from Ulcerative Colitis patients in remission were collected with written informed consent at the Hospital Clinic of Barcelona with ethical approval granted by the clinical research ethics committee of the Hospital Clínic of Barcelona (Reg. HCB/2022/0125). Prostate tissues were collected with written informed consent from previously untreated patients undergoing radical prostatectomy at St. Andrew's Hospital, Adelaide, Australia, through the Australian Prostate Cancer BioResource. Ethical approval for tissue collection and experimentation was obtained from St Andrew's (#80) and the University of Adelaide (H-2012-016) Human Research Ethics committees. This study was conducted in accordance with the Declaration of Helsinki principles.

Consent for publication
Not applicable.

Competing interests
H.H. is a co-founder and shareholder of Omniscope, a scientific advisory board member of MiRXES and Nanostring, and a consultant to Moderna and Singularity. L.G.M is an advisor and shareholder of Omniscope and advisor for ArgenTAG, Millenium Sciences, and Truckee Applied Genomics. J.C.N. is a consultant of Omniscope. M.C.N. has received grants (unrestricted) from GSK, the European Commission, and Lung Foundation Netherlands. M.v.d.B. has received grants (unrestricted) from AstraZeneca, Novartis, GlaxoSmithKline, Roche, Genentech, and Sanofi. F.T. received speaker's fees from Falk, Janssen, and AbbVie. S.S. has been a consultant for AbbVie, Bristol Myers Squibb, Boehringer Ingelheim, Ferring, Genentech/Roche, Janssen, Lilly, Novartis, Merck Sharp Dohme, Medimmune/AstraZeneca, Pfizer, Protagonist, Sanofi, Takeda, Theravance, and UCB and is a paid speaker for AbbVie, Ferring, Janssen, Merck Sharp Dohme, Novartis, Takeda, and UCB. P.R. has been a consultant for Takeda and Omass. I.A.P. holds current research grants with AstraZeneca and BMS. Omniscope has filed a patent related to the application of the FixNCut protocol. All other authors declare no competing interests.

Author details
[1]Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain. [2]Universitat de Barcelona (UB), Barcelona, Spain. [3]Inflammatory Bowel Disease Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [4]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Barcelona, Spain. [5]Cancer Immunology Program, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia. [6]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC, Australia. [7]Monash University Department of Surgery, Alfred Hospital, Melbourne, VIC, Australia. [8]Adelaide Centre for Epigenetics (ACE), University of Adelaide, Adelaide, South Australia, Australia. [9]South Australian immunoGENomics Cancer Institute (SAiGENCI), University of Adelaide, Adelaide, South Australia, Australia. [10]Australian Genomics Research Facility, Adelaide, South Australia, Australia. [11]Freemasons Foundation Centre for Men's Health, University of Adelaide, Adelaide, South Australia, Australia. [12]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia. [13]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [14]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain. [15]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany. [16]Department of Internal Medicine I, University Hospital Schleswig-Holstein (UKSH), Campus Kiel, Kiel, Germany. [17]Department of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, Groningen, Netherlands. [18]Groningen Research Institute for Asthma and COPD, University of Groningen, University Medical Center Groningen, Groningen, Netherlands. [19]Department of Pathology & Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands. [20]ICREA, Barcelona, Spain. [21]St Jude Children's Research Hospital, Memphis, TN 38105, USA. [22]Omniscope, Barcelona, Spain.

Received: 12 April 2023 Accepted: 18 March 2024
Published online: 29 March 2024

References

1. Ziegenhain C, Vieth B, Parekh S, Hellmann I, Enard W. Quantitative single-cell transcriptomics. Brief Funct Genomics. 2018;17:220–32.
2. O'Flanagan CH, Campbell KR, Zhang AW, Kabeer F, Lim JLP, Biele J, et al. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol. 2019;20:210.
3. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21:130.
4. Massoni-Badosa R, Iacono G, Moutinho C, Kulis M, Palau N, Marchese D, et al. Sampling time-dependent artifacts in single-cell genomics studies. Genome Biol. 2020;21:112.
5. Marsh SE, Walker AJ, Kamath T, Dissing-Olesen L, Hammond TR, de Soysa TY, et al. Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain. Nat Neurosci. 2022;25:306–16.
6. Boonlayangoor P, Telischi M, Boonlayangoor S, Sinclair TF, Millhouse EW. Cryopreservation of human granulocytes: study of granulocyte function and ultrastructure. Blood. 1980;56:237–45.
7. Kotsakis A, Harasymczuk M, Schilling B, Georgoulias V, Argiris A, Whiteside TL. Myeloid-derived suppressor cell measurements in fresh and cryopreserved blood samples. J Immunol Methods. 2012;381:14–22.
8. de Ruiter K, van Staveren S, Hilvering B, Knol E, Vrisekoop N, Koenderman L, et al. A field-applicable method for flow cytometric analysis of granulocyte activation: cryopreservation of fixed granulocytes. Cytom Part J Int Soc Anal Cytol. 2018;93:540–7.
9. Wang X, Yu L, Wu AR. The effect of methanol fixation on single-cell RNA sequencing data. BMC Genomics. 2021;22:420.
10. García-Castro H, Kenny NJ, Iglesias M, Álvarez-Campos P, Mason V, Elek A, et al. ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. Genome Biol. 2021;22:89.
11. Katzenelenbogen Y, Sheban F, Yalin A, Yofe I, Svetlichnyy D, Jaitin DA, et al. Coupled scRNA-Seq and intracellular protein activity reveal an immunosuppressive role of TREM2 in cancer. Cell. 2020;182:872-885.e19.
12. Phan HV, van Gent M, Drayman N, Basu A, Gack MU, Tay S. High-throughput RNA sequencing of paraformaldehyde-fixed single cells. Nat Commun. 2021;12:5636.
13. Vértesy Á, Eichmueller OL, Naas J, Novatchkova M, Esk C, Balmaña M, et al. Cellular stress in brain organoids is limited to a distinct and bioinformatically removable subpopulation. bioRxiv. 2022:2022.03.11.483643. https://doi.org/10.1101/2022.03.11.483643v1.

14. Machado L, Relaix F, Mourikis P. Stress relief: emerging methods to mitigate dissociation-induced artefacts. Trends Cell Biol. 2021;31:888–97.
15. Machado L, Geara P, Camps J, Santos MD, Teixeira-Clerc F, Herck JV, et al. Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation. Cell Stem Cell. 2021;28:1125-1135.e7.
16. Attar M, Sharma E, Li S, Bryer C, Cubitt L, Broxholme J, et al. A practical solution for preserving single cells for RNA sequencing. Sci Rep. 2018;8:2151.
17. Gerlach JP, van Buggenum JAG, Tanis SEJ, Hogeweg M, Heuts BMH, Muraro MJ, et al. Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. Sci Rep. 2019;9:1469.
18. Nesterenko PA, McLaughlin J, Cheng D, Bangayan NJ, Burton Sojo G, Seet CS, et al. Droplet-based mRNA sequencing of fixed and permeabilized cells by CLInt-seq allows for antigen-specific TCR cloning. Proc Natl Acad Sci. 2021;118:e2021190118.
19. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Moser K, Ortmann WA, et al. Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. Genes Immun. 2004;5:347–53.
20. Arceneaux D, Chen Z, Simmons AJ, Heiser CN, Southard-Smith AN, Brenan MJ, et al. A contamination focused approach for optimizing the single-cell RNA-seq experiment. iScience. 2023;26:107242.
21. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science. 2016;353:925–8.
22. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017;14:935–6.
23. Black S, Phillips D, Hickey JW, Kennedy-Darling J, Venkataraaman VG, Samusik N, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nat Protoc. 2021;16:3802–35.
24. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.
25. Uniken Venema WTC, Ramírez-Sánchez AD, Bigaeva E, Withoff S, Jonkers I, McIntyre RE, et al. Gut mucosa dissociation protocols influence cell type proportions and single-cell gene expression levels. Sci Rep. 2022;12:9897.
26. Deleersnijder D, Callemeyn J, Arijs I, Naesens M, Van Craenenbroeck AH, Lambrechts D, et al. Current methodological challenges of single-cell and single-nucleus RNA-sequencing in glomerular diseases. J Am Soc Nephrol JASN. 2021;32:1838–52.
27. Gutiérrez-Franco A, Hassan MN, Mularoni L, Plass M. Methanol fixation is the method of choice for droplet-based single-cell transcriptomics of neural cells. bioRxiv. 2022:2022.08.03.502652. https://doi.org/10.1101/2022.08.03.502652v1.
28. Fortmann SD, Frey BF, Hanumanthu VS, Liu S, Goldsborough A, Kilchrist KV, et al. Fixation before dissociation using a deep eutectic solvent preserves in vivo states and phospho-signaling in single-cell sequencing. bioRxiv. 2023:2023.02.13.528370. https://doi.org/10.1101/2023.02.13.528370v1.
29. Veny M, Garrido-Trigo A, Corraliza AM, Masamunt MC, Bassolas-Molina H, Esteller M, et al. Dissecting common and unique effects of anti-α4β7 and anti-tumor necrosis factor treatment in ulcerative colitis. J Crohns Colitis. 2021;15:441–52.
30. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods. 2019;16:619–26.
31. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29.
32. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology. 2023. https://doi.org/10.15252/msb.20188746.
33. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24:550–72.
34. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. 2019;8:281-291.e9.
35. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
36. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177:1888-1902.e21.
37. Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. Nat Commun. 2019;10:963.
38. Zhang H, Kang Z, Gong H, Xu D, Wang J, Li Z, et al. Digestive system is a potential route of COVID-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process. Gut. 2020;69:1010–8.
39. Bain CC, MacDonald AS. The impact of the lung environment on macrophage development, activation and function: diversity in the face of adversity. Mucosal Immunol. 2022;15:223–34.
40. Tsang DKL, Wang RJ, De Sa O, Ayyaz A, Foerster EG, Bayer G, et al. A single cell survey of the microbial impacts on the mouse small intestinal epithelium. Gut Microbes. 2022;14:2108281.
41. Garrido-Trigo A, Corraliza AM, Veny M, Dotti I, Melon-Ardanaz E, Rill A, et al. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in inflammatory bowel disease. bioRxiv. 2022:2022.11.28.518139. https://doi.org/10.1101/2022.11.28.518139v1.
42. Büttner M, Ostner J, Müller CL, Theis FJ, Schubert B. scCODA is a Bayesian model for compositional single-cell data analysis. Nat Commun. 2021;12:6876.
43. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv. 2021:060012. https://doi.org/10.1101/060012v3.
44. Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation | Genes & Immunity [Internet]. [cited 2023 Mar 23]. Available from: https://www.nature.com/articles/6364098.
45. Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796–8.

46. Jiménez-Gracia L, Heyn H, Martelotto L. FixNCut: Single-cell genomics through reversible tissue fixation and dissociation. GSE229944. Gene Expression Omnibus. 2024. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE229944.
47. Jiménez-Gracia L, Heyn H, Martelotto L. FixNCut: Single-cell genomics through reversible tissue fixation and dissociation. Zenodo. 2024. https://doi.org/10.5281/zenodo.10468192.
48. Jiménez-Gracia L, Heyn H, Martelotto L. FixNCut v1.0. Protocols.io. 2023. https://doi.org/10.17504/protocols.io.14egn3xjql5d/v1.
49. Jiménez-Gracia L, Heyn H, Martelotto L. FIXnCUT_benchmarking. GitHub; 2023. https://github.com/LJimenezGracia/FIXnCUT_benchmarking .

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Additional File 1: Supplementary Figures

Figure S1. FixNCut protocol tested in human PBMCs. (**a**) Bright-field microscopy images of fresh (**left**) and fixed (**right**) human PBMCs. (**b**) cDNA profiles of human PBMCs in fresh (**top**) and fixed (**bottom**) samples. (**c**) Final GEX library BioA profiles of human PBMCs in fresh (**left**) and fixed (**right**) samples. (**d**) Comparative analysis of the number of detected UMIs (**left**) and genes (**right**) for each cell barcode based on sequencing reads using a linear model. (**e**) Distribution of main QC metrics (number of UMIs, genes, and percentage of mitochondrial genes) by protocol. (**f**) Distribution of main QC metrics by cell type and protocol. (**g**) Overlap of highly variable genes (HVGs) shared between fresh and fixed human PBMCs, considering 3000 HVGs computed for each sample independently. (**h**) Linear regression model comparing the average gene expression levels of expressed genes by cell type. The coefficient of determination (R2) computed with Pearson correlation and the corresponding p-value are indicated. (**i**) Dot plot showing the average expression for top genes (*x-axis*) across all 19 cell types (*y-axis*). Dot size represents the percentage of cells in a cluster expressing each gene, and the color indicates the average expression level. (**j**) Score of sampling-time gene signature [4] for fresh and fixed human PBMCs across cell populations. Statistical analysis between fixed and fresh cells was performed using Wilcoxon signed-rank test; significance results are indicated (ns, p>0.05, * p<=0.05, ** p<=0.01, *** p<=0.001, **** p<=0.0001). (**k**) Overlap between significant DE genes enriched in fresh PBMCs with genes from the sampling-time signature [4].

Figure S2. FixNCut protocol tested in mouse lung samples. (**a**) Bright-field (**left**), DAPI fluorescent (**middle**), and two microscopy images overlay (**right**) of fresh (**top**) and fixed (**bottom**) mouse lung cells. (**b**) Comparative analysis of the number of detected UMIs (**left**) and genes (**right**) for each cell barcode based on sequencing reads using a linear model. (**c**) Gene annotation categories for all captured genes in both conditions or uniquely in fresh or fixed mouse lung. (**d**) Distribution of the main QC metrics (number of UMIs, genes, and percentage of mitochondrial genes) by protocol, considering all cell types. (**e**) Distribution of the main QC metrics by cell-type and protocol. (**f**) Overlap of highly variable genes (HVGs) shared between fresh and fixed mouse lung samples, considering 3000 HVGs computed by protocol. (**g**) Linear regression model comparing average gene expression levels of expressed genes between fresh and fixed samples by cell-type. The coefficient of determination (R2) computed with Pearson correlation and the corresponding p-value are indicated. (**h**) Dotplot showing average expression for the top genes (*x-axis*) for all 20 cell types (*y-axis*). The dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level.

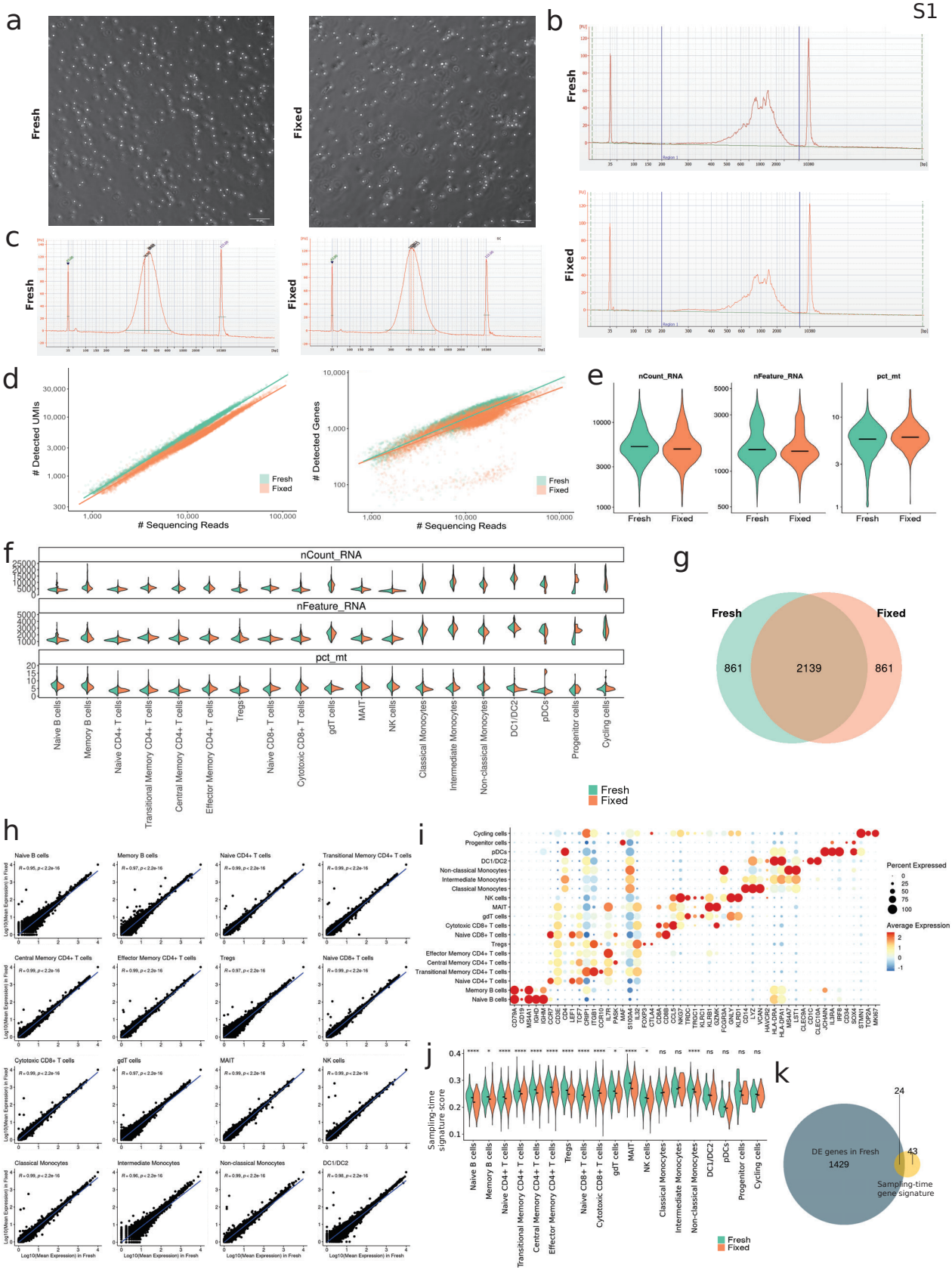
Figure S3. FixNCut protocol tested in mouse colon samples. (**a**) Comparative analysis of the number of detected UMIs (**left**) and genes (**right**) for each cell barcode based on sequencing reads using a linear model. (**b**) Gene annotation categories for all captured genes in both conditions or uniquely in fresh or fixed mouse colon. (**c**) Distribution of main QC metrics (number of UMIs, genes, and percentage of mitochondrial genes) by protocol. (**d**) Distribution of main QC metrics by cell type and protocol. (**e**) Overlap of highly variable genes (HVGs) shared between fresh and fixed mouse colon samples, considering 3000 HVGs computed for each sample independently. (**f**) Linear regression model comparing the average gene expression levels of expressed genes by cell type. The coefficient of determination (R2) computed with Pearson correlation and the corresponding p-value are indicated. (**g**) Dotplot showing average expression for the top genes (*y-axis*) for all 16 cell types

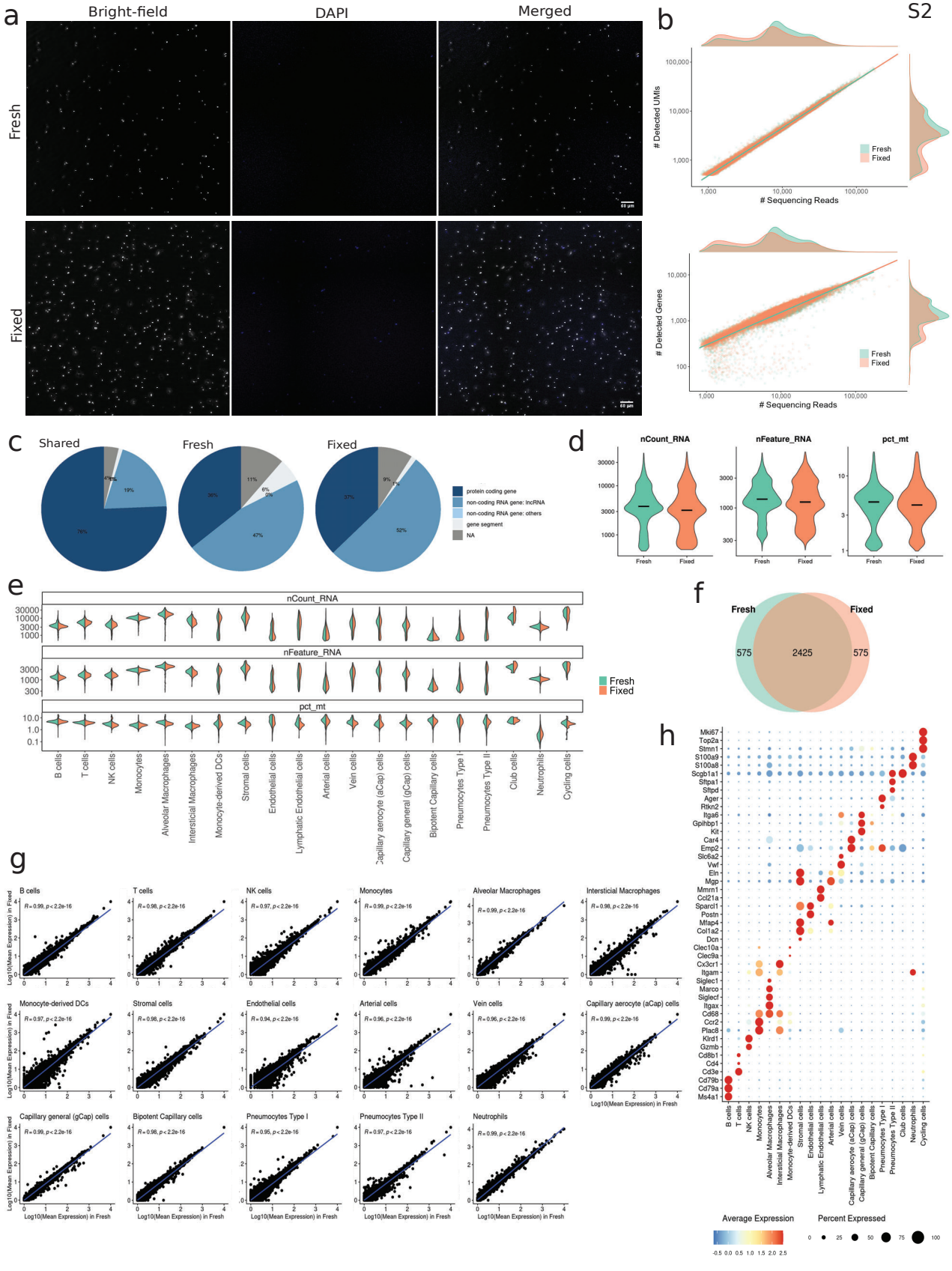
(*x-axis*). The dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level.

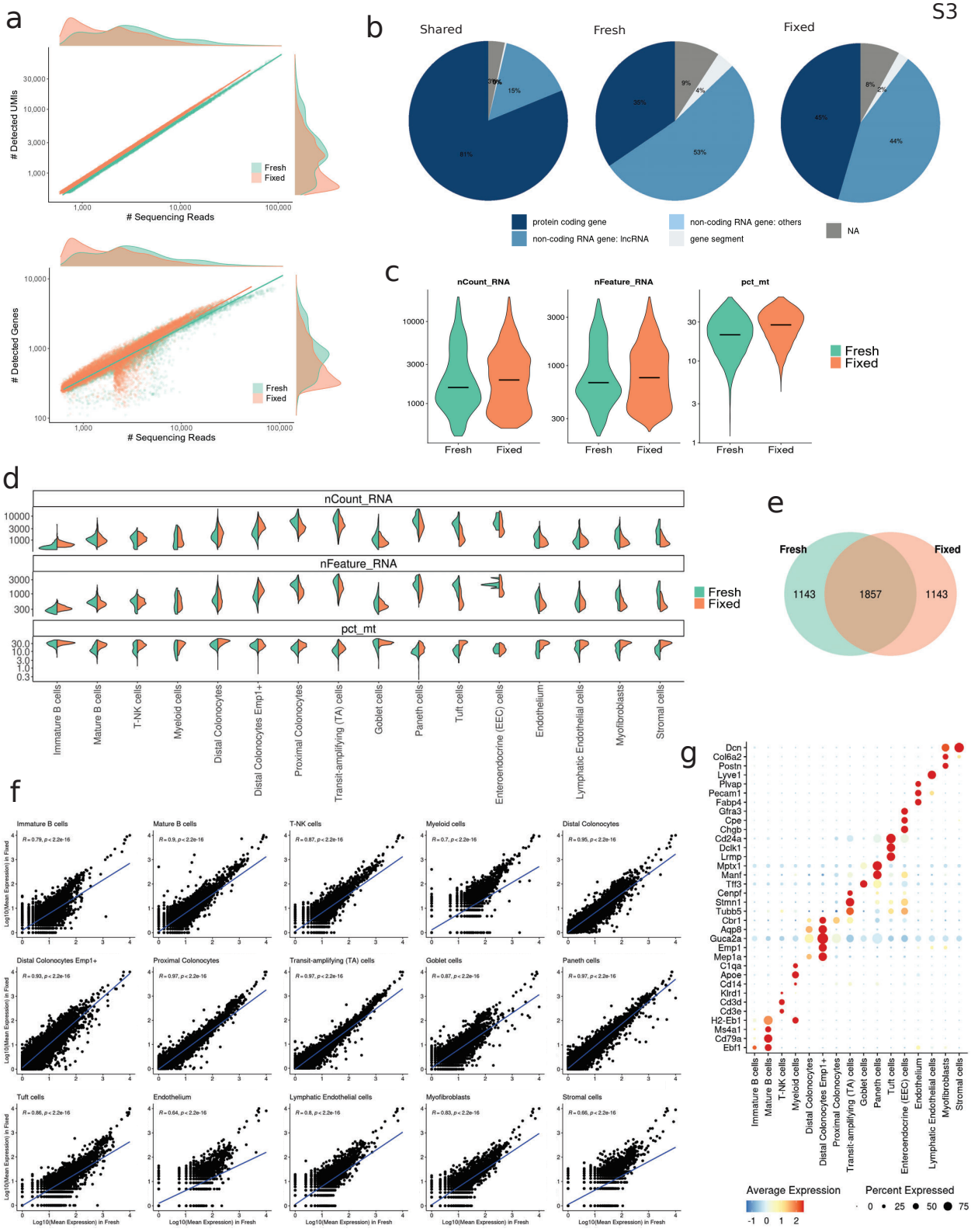
Figure S4. Long-term storage of fixed mouse lung samples. (**a**) Comparative analysis of the number of detected UMIs (**left**) and genes (**right**) for each cell barcode based on sequencing reads using a linear model. (**b**) Gene annotation categories for all captured genes, either in all three conditions or uniquely in cryo or fixed/cryo mouse lungs. (**c**) Distribution of main QC metrics (number of UMIs, genes and percentage of mitochondrial genes) by protocol, considering all cell types. (**d**) Distribution of main QC metrics by cell-type and protocol. (**e**) Overlap of highly variable genes (HVGs) shared between protocols, considering the 3000 HVGs computed for each protocol independently. (**f**) Linear regression model comparing the average gene expression levels of expressed genes by cell-type and protocol; fixed/cryo vs fixed (**top**) and fixed/cryo vs cryo (**bottom**). The coefficient of determination (R2) computed with Pearson correlation and the corresponding p-value are indicated. (**g**) Dotplot showing average expression for the top genes (x-axis) for all 20 cell types (y-axis). Dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level.

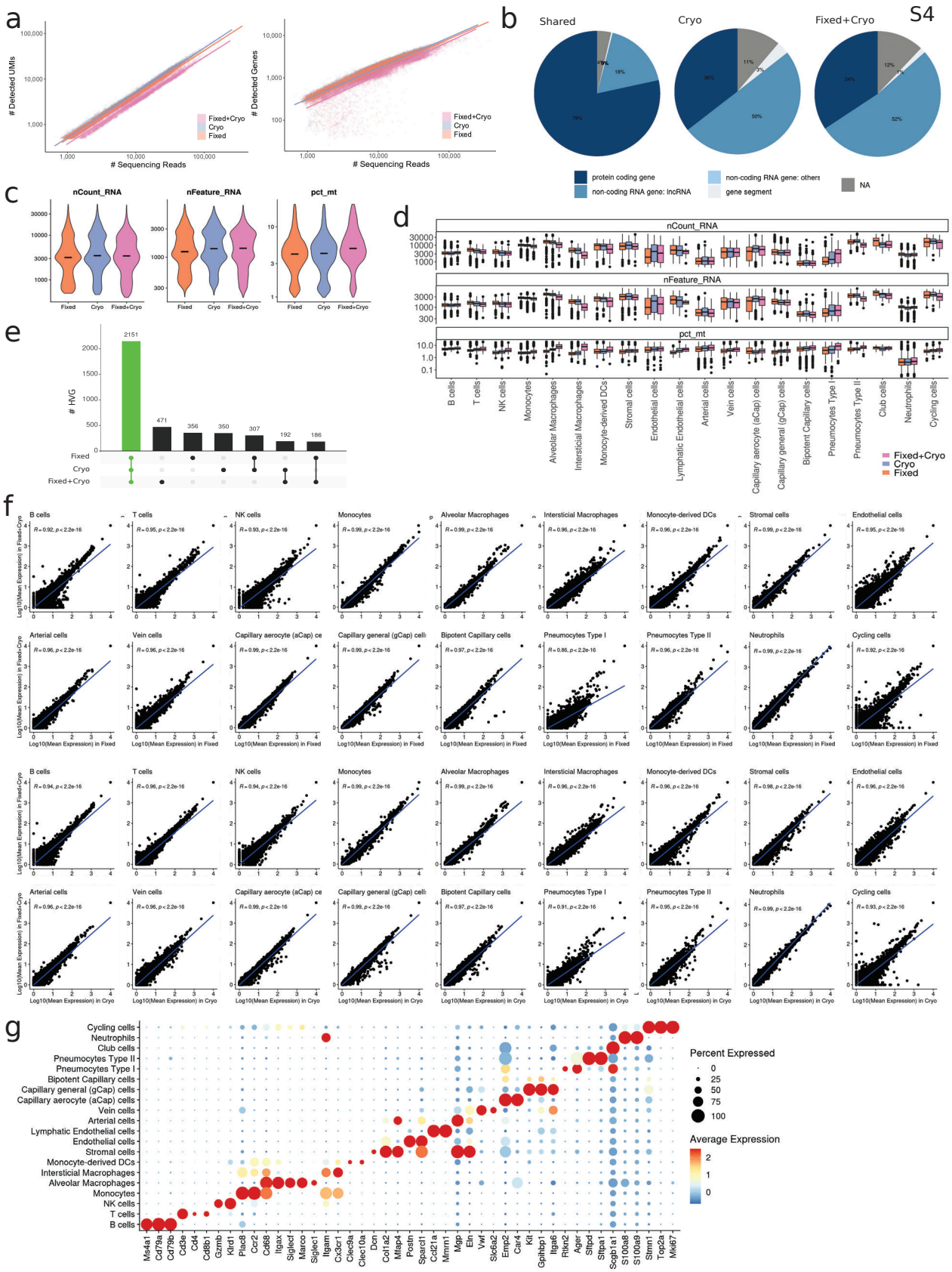
Figure S5. FixNCut protocol tested in human colon biopsies. (**a**) Comparative analysis of the number of detected UMIs (**left**) and genes (**right**) for each cell barcode based on sequencing reads using a linear model. (**b**) Gene annotation categories for all captured genes, either in all conditions or uniquely in fresh, fixed, cryo or fixed/cryo human colon biopsies. (**c**) Distribution of main QC metrics (number of UMIs, genes and percentage of mitochondrial genes) by protocol. (**d**) Distribution of main QC metrics by cell-type and protocol. (**e**) Overlap of highly variable genes (HVGs) shared between protocols, considering the 3000 HVGs computed for each condition independently. (**f**) Linear regression model comparing average gene expression levels of expressed genes by cell-type and protocol; fixed vs fresh (**top**) and fixed/cryo vs cryo (**bottom**). (**g**) Dotplot showing average expression for the top genes (*x-axis*) for all 21 cell types (*y-axis*). Dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level. (**h**) Score of warm collagenase gene signature [2] for human mouse colon by cell population across protocols.

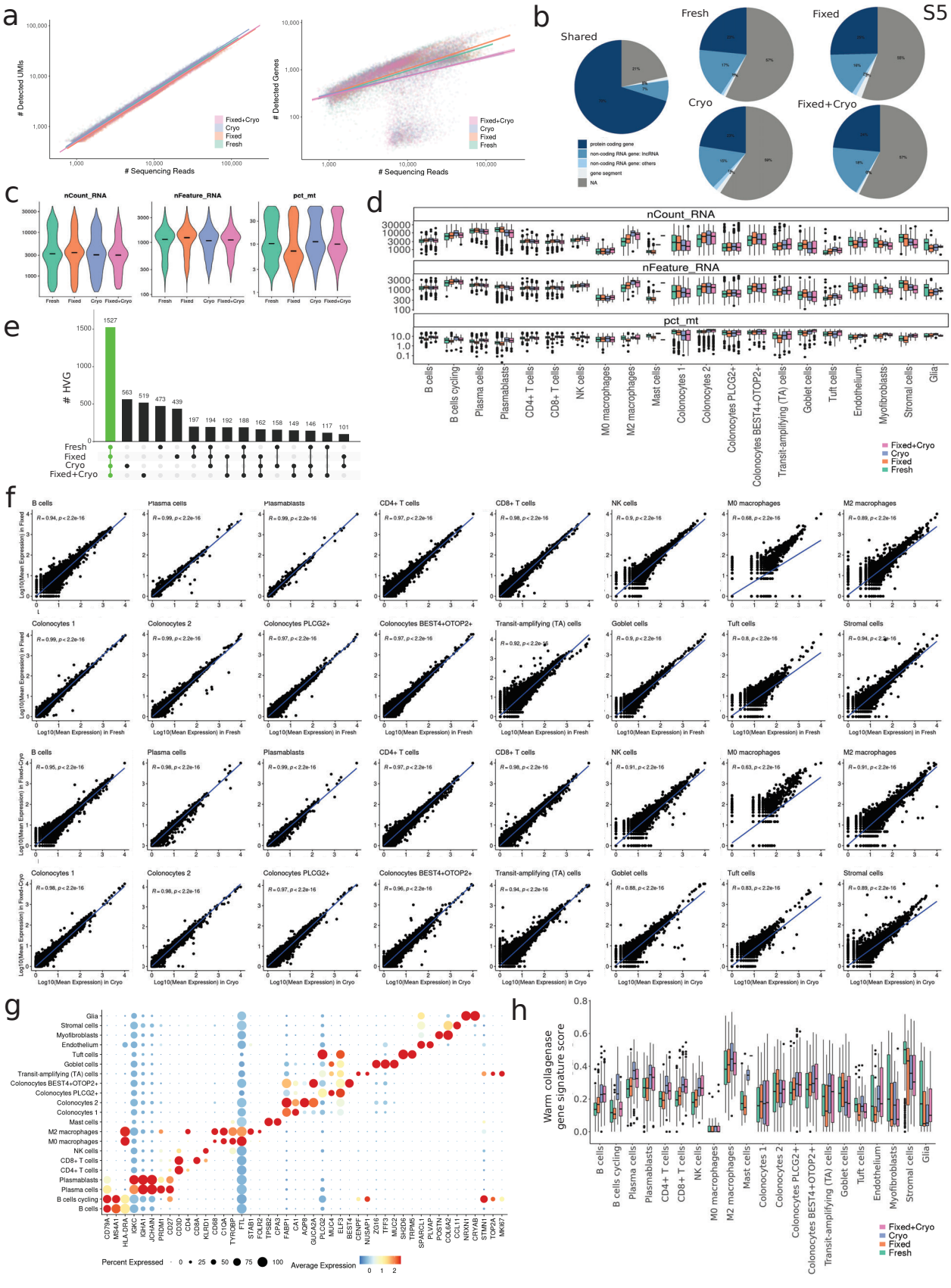
Figure S6. Flow cytometry analysis in fixed cells and tissues in mouse and human. (**a**) Representative flow cytometry experiment of forward scatter (FSC) and side scatter (SSC) in cryopreserved (**left**) and cryo+fixed (**right**) human PBMCs from healthy donors (n=3). (**b**) Representative flow cytometry experiment of cell viability based on DAPI expression in cryopreserved (**left**) and cryo+fixed (**right**) human PBMCs from healthy donors (n=3). (**c**) Representative flow cytometry experiment of cell apoptosis based on the expression of Annexin V FITC and PI in fresh (**left**) and fixed (**right**) freshly isolated PBMCs. Cells were analyzed after isolation and fixation at day 0 (**top**) and 2-days after fixation (**bottom**) (n=4). (**d**) Representative gating strategy of one experiment analyzed by flow cytometry of freshly isolated PBMCs from healthy donors (n=4). PBMCs were stained with anti-human CD3, CD19, CD4 and CD8 monoclonal antibodies (mAbs). T cells were selected by the positive expression of CD3. CD4 positive and CD8 positive T cells were selected from CD3 positive. (**e**) Representative flow cytometry experiment of forward scatter (FSC) and side scatter (SSC) (**top**) and cell viability based on DAPI expression (**bottom**) of human colon biopsies from IBD donors.

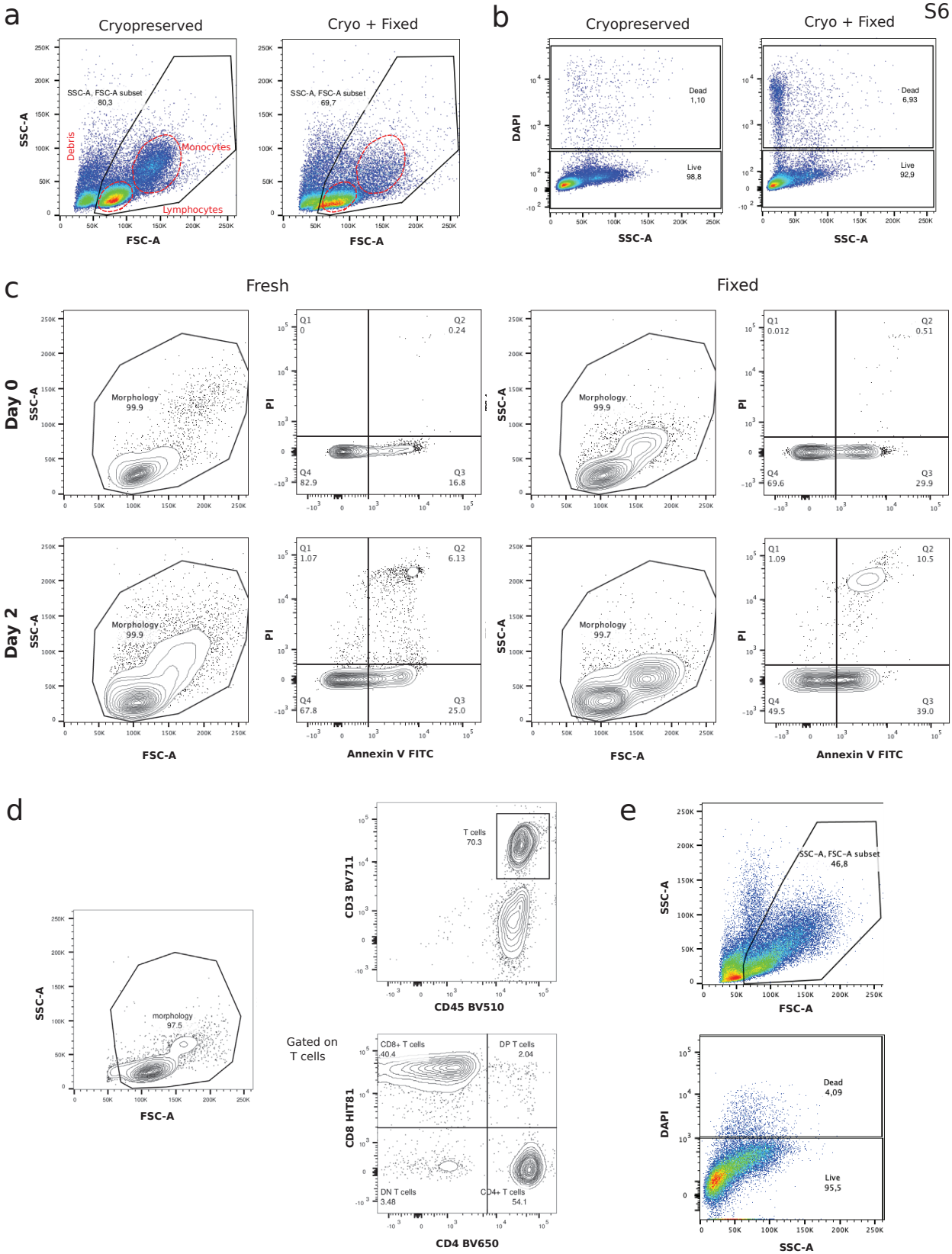

a
Fresh
Fixed
b
Fresh
Fixed
c
Fresh
Fixed
d
Detected UMIs
Sequencing Reads
Detected Genes
Sequencing Reads
Fresh
Fixed
e
nCount_RNA
nFeature_RNA
pct_mt
Fresh
Fixed
f
nCount_RNA
nFeature_RNA
pct_mt
Naive B cells
Memory B cells
Naive CD4+ T cells
Transitional Memory CD4+ T cells
Central Memory CD4+ T cells
Effector Memory CD4+ T cells
Tregs
Naive CD8+ T cells
Cytotoxic CD8+ T cells
gdT cells
MAIT
NK cells
Classical Monocytes
Intermediate Monocytes
Non-classical Monocytes
DC1/DC2
pDCs
Progenitor cells
Cycling cells
Fresh
Fixed
g
Fresh
Fixed
861
2139
861
h
Naive B cells
Memory B cells
Naive CD4+ T cells
Transitional Memory CD4+ T cells
Central Memory CD4+ T cells
Effector Memory CD4+ T cells
Tregs
Naive CD8+ T cells
Cytotoxic CD8+ T cells
gdT cells
MAIT
NK cells
Classical Monocytes
Intermediate Monocytes
Non-classical Monocytes
DC1/DC2
Log10(Mean Expression) in Fresh
Log10(Mean Expression) in Fixed
i
Percent Expressed
Average Expression
j
Sampling-time signature score
Fresh
Fixed
k
DE genes in Fresh
1429
24
43
Sampling-time gene signature


a
Bright-field
DAPI
Merged
Fresh
Fixed
b
S2
Detected UMIs
Sequencing Reads
Fresh
Fixed
Detected Genes
c
Shared
Fresh
Fixed
protein coding gene
non-coding RNA gene: lncRNA
non-coding RNA gene: others
gene segment
NA
d
nCount_RNA
nFeature_RNA
pct_mt
e
B cells
T cells
NK cells
Monocytes
Alveolar Macrophages
Interstitial Macrophages
Monocyte-derived DCs
Stromal cells
Endothelial cells
Lymphatic Endothelial cells
Arterial cells
Vein cells
Capillary aerocyte (aCap) cells
Capillary general (gCap) cells
Bipotent Capillary cells
Pneumocytes Type I
Pneumocytes Type II
Club cells
Neutrophils
Cycling cells
f
575
2425
575
g
Log10(Mean Expression) in Fresh
Log10(Mean Expression) in Fixed
h
Average Expression
Percent Expressed


a
Detected UMIs
Sequencing Reads
Fresh
Fixed
Detected Genes
b
Shared
Fresh
Fixed
protein coding gene
non-coding RNA gene: lncRNA
non-coding RNA gene: others
gene segment
NA
c
nCount_RNA
nFeature_RNA
pct_mt
d
e
1143
1857
1143
f
Immature B cells
Mature B cells
T-NK cells
Myeloid cells
Distal Colonocytes
Distal Colonocytes Emp1+
Proximal Colonocytes
Transit-amplifying (TA) cells
Goblet cells
Paneth cells
Tuft cells
Endothelium
Lymphatic Endothelial cells
Myofibroblasts
Stromal cells
Log10(Mean Expression) in Fresh
Log10(Mean Expression) in Fixed
g
Average Expression
Percent Expressed


a
Detected UMIs
Sequencing Reads
Detected Genes
Fixed+Cryo
Cryo
Fixed
b
Shared
Cryo
Fixed+Cryo
S4
protein coding gene
non-coding RNA gene: lncRNA
non-coding RNA gene: others
gene segment
NA
c
nCount_RNA
nFeature_RNA
pct_mt
Fixed
Cryo
Fixed+Cryo
d
nCount_RNA
nFeature_RNA
pct_mt
B cells
T cells
NK cells
Monocytes
Alveolar Macrophages
Interstitial Macrophages
Monocyte-derived DCs
Stromal cells
Endothelial cells
Lymphatic Endothelial cells
Arterial cells
Vein cells
Capillary aerocyte (aCap) cells
Capillary general (gCap) cells
Bipotent Capillary cells
Pneumocytes Type I
Pneumocytes Type II
Club cells
Neutrophils
Cycling cells
e
HVG
2151
471
356
350
307
192
186
Fixed
Cryo
Fixed+Cryo
f
Log10(Mean Expression) in Fixed+Cryo
Log10(Mean Expression) in Fixed
Log10(Mean Expression) in Cryo
g
Percent Expressed
0
25
50
75
100
Average Expression
2
1
0
Ms4a1
Cd79a
Cd79b
Cd3e
Cd4
Cd8b1
Gzmb
Klrd1
Plac8
Ccr2
Cd68
Itgax
Siglecf
Marco
Siglec1
Itgam
Cx3cr1
Clec9a
Clec10a
Dcn
Col1a2
Mfap4
Postn
Sparcl1
Cd21a
Mmrn1
Mgp
Eln
Vwf
Slc6a2
Emp2
Car4
Kit
Gpihbp1
Itga6
Rtkn2
Ager
Sftpd
Sftpa1
Scgb1a1
S100a8
S100a9
Stmn1
Top2a
Mki67


a
Detected UMIs
Sequencing Reads
Detected Genes
Fixed+Cryo
Cryo
Fixed
Fresh
b
Shared
Fresh
Fixed
Cryo
Fixed+Cryo
S5
protein coding gene
non-coding RNA gene: lncRNA
non-coding RNA gene: others
gene segment
NA
c
nCount_RNA
nFeature_RNA
pct_mt
d
nCount_RNA
nFeature_RNA
pct_mt
B cells
B cells cycling
Plasma cells
Plasmablasts
CD4+ T cells
CD8+ T cells
NK cells
M0 macrophages
M2 macrophages
Mast cells
Colonocytes 1
Colonocytes 2
Colonocytes PLCG2+
Colonocytes BEST4+OTOP2+
Transit-amplifying (TA) cells
Goblet cells
Tuft cells
Endothelium
Myofibroblasts
Stromal cells
Glia
e
HVG
1527
563
519
473
439
197
194
192
188
162
158
149
146
117
101
Fresh
Fixed
Cryo
Fixed+Cryo
f
Log10(Mean Expression) in Fixed
Log10(Mean Expression) in Fresh
Log10(Mean Expression) in Fixed+Cryo
Log10(Mean Expression) in Cryo
g
Percent Expressed
Average Expression
h
Warm collagenase gene signature score


a
Cryopreserved
Cryo + Fixed
SSC-A, FSC-A subset
80,3
69,7
Debris
Monocytes
Lymphocytes
SSC-A
FSC-A
b
Cryopreserved
Cryo + Fixed
DAPI
Dead
1,10
6,93
Live
98,8
92,9
SSC-A
c
Fresh
Fixed
Day 0
Day 2
Morphology
99.9
99.7
PI
Annexin V FITC
Q1
0
0.24
0.012
0.51
1.07
6.13
1.09
10.5
Q4
82.9
69.6
67.8
49.5
Q3
16.8
29.9
25.0
39.0
d
morphology
97.5
CD3 BV711
CD45 BV510
T cells
70.3
Gated on
T cells
CD8 HIT81
CD4 BV650
CD8+ T cells
40.4
DP T cells
2.04
DN T cells
3.48
CD4+ T cells
54.1
e
SSC-A, FSC-A subset
46,8
Dead
4,09
Live
95,5

Publication 2. Interpretable Inflammation Landscape of Circulating Immune cells[2]

Laura Jiménez-Gracia*, Davide Maspero*, Sergio Aguilar-Fernández*, Francesco Craighero*, Maria Boulougouri, Sara Ruiz, Domenica Marchese, Ginevra Caratù, Jose Liñares-Blanco, Miren Berasategi, Ricardo O Ramirez Flores, Angela Sanzo-Machuca, Ana M. Corraliza, Hoang A. Tran, Rachelly Normand, Jacquelyn Nestor, Yourae Hong, Tessa Kole, Petra van der Velde, Frederique Alleblas, Flaminia Pedretti, Adrià Aterido, Martin Banchero, German Soriano, Eva Román, Maarten van den Berge, Azucena Salas, Jose Manuel Carrascosa, Antonio Fernández Nebro, Eugeni Domènech, Juan Cañete, Jesús Tornero, Javier Pérez-Gisbert, Ernest Choy, Giampiero Girolomoni, Britta Siegmund, Antonio Julià, Violeta Serra, Roberto Elosua, Sabine Tejpar, Silvia Vidal, Martijn C. Nawijn, Ivo Gut, Julio Saez-Rodriguez, Sara Marsal, Pierre Vandergheynst, Alexandra-Chloé Villani, Juan C. Nieto†, Holger Heyn†. **Interpretable Inflammation Landscape of Circulating Immune cells**

* These authors contributed equally

† Joint senior authors

DOI: https://doi.org/10.1101/2023.11.28.568839

The manuscript is currently under revision and has undergone substantial improvements compared to the version published on BioRxiv. In this thesis, I am including the revised version, which will be used for the discussion.

Raw data: Accessible via the Sequence Read Archive (SRA), NCBI Gene Expression Omnibus (GEO) and European Genome Archive (EGA) databases. Further details, including specific accession numbers, are provided in *Supplementary Table 1 - Sheet 1* of the article.

Processed data: https://zenodo.org/records/13374575

Code: https://github.com/Single-Cell-Genomics-Group-CNAG-CRG/Inflammation-PBMCs-Atlas

Data and code will be available upon publication.

[2] In this paper, I collected and preprocessed all the data included in the study, contributed to data integration and cell identity annotation to generate the inflammation atlas, computed the inflammation-related signature activity scores and performed a comprehensive analysis across cell-types and diseases, and validated findings on external datasets. However, I did not focus on the implementation of machine learning models for biomarker discovery or the patient classifier. Therefore, I will extensively discuss the results from which I can claim direct credit.

Interpretable Inflammation Landscape of Circulating Immune cells

Laura Jiménez-Gracia[1,2*], Davide Maspero[1*], Sergio Aguilar-Fernández[1*], Francesco Craighero[3,*], Maria Boulougouri[1,3], Sara Ruiz[1], Domenica Marchese[1], Ginevra Caratù[1], Jose Liñares-Blanco[4,5], Miren Berasategi[1], Ricardo O Ramirez Flores[4], Angela Sanzo-Machuca[6,7], Ana M. Corraliza[6,7], Hoang A. Tran[8,9,10], Rachelly Normand[8,9,10], Jacquelyn Nestor[8,9,10], Yourae Hong[11], Tessa Kole[12,13], Petra van der Velde[12,14], Frederique Alleblas[12,14], Flaminia Pedretti[15], Adrià Aterido[16,17], Martin Banchero[12,14], German Soriano[18,19], Eva Román[18,19], Maarten van den Berge[12,13], Azucena Salas[6,7], Jose Manuel Carrascosa[20], Antonio Fernández Nebro[21], Eugeni Domènech[22], Juan Cañete[23], Jesús Tornero[24], Javier Pérez-Gisbert[25], Ernest Choy[26], Giampiero Girolomoni[27], Britta Siegmund[28], Antonio Julià[16,17], Violeta Serra[15], Roberto Elosua[29,30,31], Sabine Tejpar[11], Silvia Vidal[32], Martijn C. Nawijn[12,14], Ivo Gut[1,2], Julio Saez-Rodriguez[4,33], Sara Marsal[16,17,34], Pierre Vandergheynst[3], Alexandra-Chloé Villani[8,9,10], Juan C. Nieto[1†], Holger Heyn[1,2†]

1 Centro Nacional de Análisis Genómico, C/Baldiri Reixac 4, 08028 Barcelona, Spain.
2 Universitat de Barcelona (UB), Barcelona, Spain.
3 Signal Processing Laboratory 2 (LTS2), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
4 Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany
5 Department of Statistics. University of Granada, 18071, Granada, Spain.
6 Inflammatory Bowel Disease Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
7 Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Spain.
8 Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, 02129, Massachusetts, USA.
9 Broad Institute of MIT and Harvard, Cambridge, 02142, Massachusetts, USA.
10 Harvard Medical School, Boston, Massachusetts, USA.
11 Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium.
12 Groningen Research Institute for Asthma and COPD (GRIAC), University Medical Center Groningen, Groningen, Netherlands.
13 Department of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.
14 Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.
15 Experimental Therapeutics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain.
16 Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain.
17 IMIDomics, Inc.
18 Department of Gastroenterology, Biomedical Research Institut Sant Pau (IIB Sant Pau), Barcelona, Spain.
19 Biomedical Research Network on Hepatic and Digestive Diseases (CIBEREHD), Instituto de Salud Carlos III. Madrid, Spain.
20 Dermatology Department, Hospital Universitari Germans Trias i Pujol. Badalona, Spain.
21 Rheumatology Department, Hospital Regional Universitario Carlos Haya. Málaga, Spain.
22 Gastroenterology Department, Hospital Universitari Germans Trias i Pujol, Badalona, Spain.
23 Rheumatology Department, Fundació Clínic per a la Recerca Biomèdica. Barcelona, Spain.
24 Rheumatology Department, Hospital Universitario Guadalajara. Guadalajara, Spain.
25 Gastroenterology Department, Hospital Universitario de la Princesa. Madrid, Spain.
26 Section of Rheumatology, Cardiff University, Cardiff, United Kingdom
27 Section of Dermatology and Venereology, University of Verona, 37129 Verona, Italy.
28 Department of Gastroenterology, Rheumatology and Infectious Diseases, Charité-Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany.
29 Hospital del Mar Research Institute (IMIM). Barcelona, Catalonia, Spain.
30 CIBERCV, Instituto de Salud Carlos III. Madrid, Spain.
31 Faculty of Medicine, University of Vic-Central University of Catalonia. Vic, Catalonia, Spain
32 Group of Immunology-Inflammatory Diseases, Biomedical Research Institut Sant Pau (IIB Sant Pau), Barcelona, Spain.
33 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, U.K.
34 On behalf of IMID-Consortium.

* These authors contributed equally
† Joint senior authors.
To whom correspondence should be addressed.
Juan C. Nieto (juan.nieto@cnag.eu);
Holger Heyn (holger.heyn@cnag.eu)

Abstract

Inflammation is a biological phenomenon involved in a wide variety of physiological and pathological processes. Although a controlled inflammatory response is beneficial for restoring homeostasis, it can become unfavorable if dysregulated. In recent years, major progress has been made in characterizing acute and chronic inflammation in specific diseases. However, a global, holistic understanding of inflammation is still elusive. This is particularly intriguing, considering the crucial function of inflammation for human health and its potential for modern medicine if fully deciphered. Here, we leverage advances in the field of single-cell genomics to delineate the full spectrum of circulating immune cell activation underlying inflammatory processes during infection, immune-mediated inflammatory diseases and cancer. Our single-cell atlas of >6.5 million peripheral blood mononuclear cells from 1047 patients and 19 diseases allowed us to learn a comprehensive model of inflammation in circulating immune cells. The atlas expanded our current knowledge of the biology of inflammation of acute (e.g. inflammatory bowel disease, sepsis) and chronic (e.g. cirrhosis, asthma, and chronic obstructive pulmonary disease) disease processes and laid the foundation to develop a precision medicine framework using unsupervised as well as explainable machine learning. Beyond a disease-centered classification, we charted altered activity of inflammatory molecules in peripheral blood cells, depicting functional biomarkers to further understand mechanisms of inflammation. Finally, we have laid the groundwork for developing precision medicine diagnostic tools for patients experiencing severe acute or chronic inflammation by learning a classifier for inflammatory diseases, presenting cells in circulation as a powerful resource for patient stratification.

Introduction

Inflammation is a biological response or state of the immune system that serves to protect the human body from environmental challenges, thereby preserving homeostasis and structural integrity of tissues and organs[1]. Inflammatory processes are activated in response to various triggers, such as infection or injury, and involve a multistep defensive mechanism aimed at eliminating the source of perturbation[2–4]. Thus, inflammation represents an altered state within the immune system, which can manifest as either a protective or pathological response[5]. The cellular and molecular mediators of inflammation play pivotal roles in nearly every human disease, encompassing a wide array of biological processes, including the complex interplay of cytokines, myeloid and lymphoid cells[6].

The initiation of inflammatory processes is driven by cellular stimulation, triggered by the release of proinflammatory cytokines[7,8]. These cytokines exert autocrine and paracrine effects, activating endothelial cells, subsequently increasing vascular permeability. This allows immune cells to infiltrate tissues at the site of infection, facilitated by chemokines. Chemokines are essential for recruiting additional immune cells, playing a crucial role in phagocytosis and pathogen eradication[9]. In the bloodstream, activated immune cells release cytokines and travel to various tissues. Inflammation is a central driver in cardio-vascular[10], autoimmune[11,12], infectious diseases[13,14] and even cancer[15]. The success of therapies targeting inflammation underscores the importance of understanding the underlying pathways[16–18]. Thus, categorizing patients based on their specific inflammatory cell states in the bloodstream has significant potential for advancing disease management[19].

Single-cell RNA sequencing (scRNA-seq) is becoming a conventional method for detecting altered cell states in blood, enabling the comparison of transcriptional profiles during perturbations, including inflammation[20]. Previous works revealed cellular profiles across diverse conditions, creating a shared phenotypic space that facilitates comparisons among patients and conditions, and generating a comprehensive view of inflammation[21]. Consequently, differential analysis of cell states and gene expression programs can now guide a holistic understanding of inflammation in acute and chronic diseases to form the basis for future precision medicine tools in diagnostics and novel treatments. In this regard, interpretable machine learning will play a pivotal role to extract disease-driving features from large healthy and disease single-cell references. Eventually, comprehensive models will allow the classification of patients for precise diagnostics and the patient stratification for tailored treatments.

Our study initially defined common immune cell types in peripheral blood, before capturing disease-specific inflammatory cell states that exhibit functional specialization within the inflammatory landscape. Beyond a disease-centered classification, we modeled the expression profiles of inflammatory molecules to define interpretable biomarkers driving immune cell activation, migration, cytotoxic responses, and antigen presentation activities. Ultimately, we developed a classifier based on the peripheral blood mononuclear cell (PBMC) reference, establishing inflammatory immune cell features as a precision medicine diagnostic tool for patients suffering from severe acute or chronic inflammation.

Main

An inflammation landscape of circulating immune cells

To chart a comprehensive landscape of immune cells in circulation of healthy individuals and patients suffering from inflammatory diseases, we analyzed the transcriptomic profiles of >6.5 million (6,340,934 million after filtering) PBMCs, representing 1047 patients and 19 diseases, split into a main Inflammation atlas and two validation datasets (**Fig. 1a,b**). Diseases broadly classified into five distinct groups: 1) Immune-mediated inflammatory diseases (IMIDs), 2) acute and 3) chronic inflammation, 4) infection and 5) cancers, which were profiled along with healthy donor samples (**Fig. 1a**). We completed our dataset with additional studies to generate a comprehensive resource of immune cell states across inflammatory diseases and beyond (**Fig. 1a**; **Extended Data Fig. 1a; Supplementary Table 1**). Our cohort included various scRNA-seq chemistries (10x Genomics 3' and 5' mRNA) and experimental designs (CellPlex and genotype multiplexing), as well as individuals of both sexes and across age groups, to comprehensively capture technical and biological variability (see **Methods**). To learn a generative model of circulating immune cells of inflammatory diseases, we applied probabilistic modeling of the single-cell data using scVI[22] and scANVI[23], considering clinical characteristics (disease, sex and age). scANVI generates a lower-dimensional cell embedding space, before reconstructing gene expression data. Batch effects are removed based on gene-specific parameters, learned during the integration. Its generative probabilistic models proved superior performances in integrating complex datasets compared to other approaches[24], particularly if cell annotations are available (**Extended Data Fig. 1b-c**). Applied here, the resulting gene expression profiles and the cell embedding space were batch effect corrected, while preserving biological heterogeneity (i.e. previously annotated cell types and states; **Supplementary Table 2**). From the joint embedding space, we initially assigned cells to major cell lineages and 15 subpopulations (*Level 1*; **Fig. 1c; Extended Data Fig. 1d**). Then, following a recursive, top-down clustering approach, we obtained a total of 64 cell populations (*Level 2*), comprehensively resembling immune cell states of the innate and adaptive compartments and allowing a fine-grained description of cellular states (immune cell types with distinct activation-related transcriptomes; **Extended Data Fig. 2**; **Supplementary Table 2,3**). High-level compositional analysis (*Level 1*) across diseases revealed significant changes of cell type distributions (scCODA[25]; **Extended Data Fig. 1d**) and validated previously described alterations in peripheral blood cells of patients suffering from inflammatory diseases. For example, we confirmed low levels of UTCs, ILCs and Naive CD4 T cells, together with high proportions of B cells and Monocytes in SLE[26–28]. IBD patients showed lower levels of UTCs and ILCs[29] and we observed lower proportions of UTC accompanied by a larger fraction of Monocytes and B cells in RA[30–32], line with previous observations. Lymphopenia, a common event during the development of Sepsis[33], and lymphocytosis, typical of HIV infection[34], were also confirmed.

Diving deeper into genes and gene programs to characterize inflammatory diseases, our subsequent analysis followed three complementary strategies 1) to identify disease-driving mechanisms (gene signature and Gene Regulatory Network (GRN) activity), 2) to define biomarkers for inflammatory responses (feature extraction) and 3) to classify patients based on their disease-specific signatures

(projection). Therefore, we looked at gene expression profiles holistically, but also delineated the inflammatory process by focusing on molecules that trigger immune cell activation, cellular migration and extravasation, antigen presentation and cytotoxic responses (**Supplementary Table 4**)[35–41]. These strategies jointly allowed us to enlarge our understanding of inflammatory processes and their contribution to inflammatory diseases, but also form the basis for precision medicine tools by establishing cells in circulation as potent biomarkers for disease diagnostics.

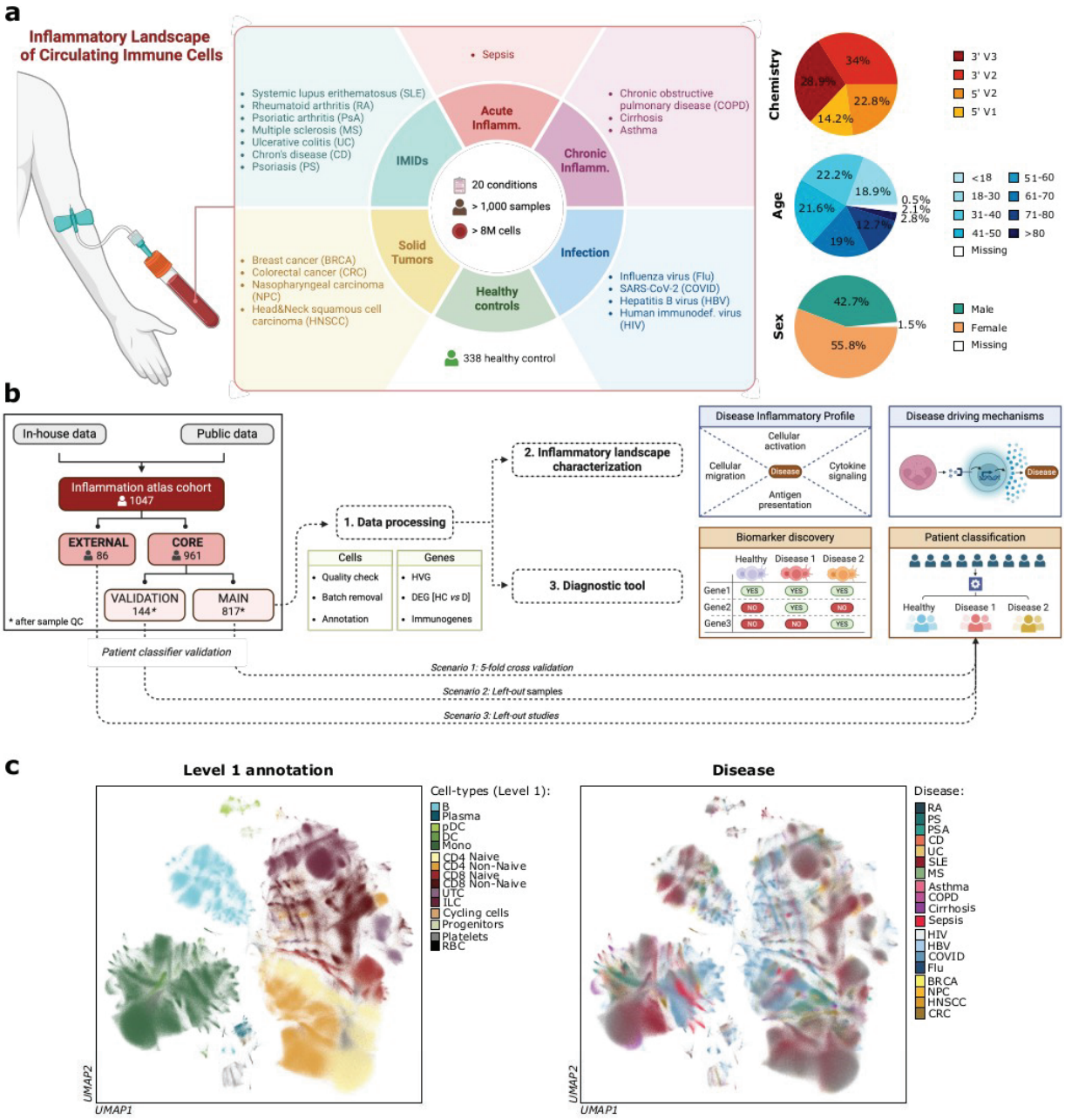


Figure 1. Inflammation Landscape of Circulating Immune Cells. (**a**) Left: Schematic overview illustrating the number of cells, samples, and conditions (diseases and disease groups) analyzed. Right: Pie charts displaying metadata related to the scRNA-seq chemistry (10x Genomics assay and version) and patient demographics (age and gender). (**b**) Schematic overview of the analysis workflow followed, detailing the division of the overall dataset into Main, Validation, and External. The figure illustrates the specific tasks and analyses performed with each dataset. (**c**) Uniform manifold approximation and projection (UMAP) embedding for the scANVI-corrected latent space considering the Main dataset (4,435,922 cells) across patients and diseases colored by the major cell lineages (left, *Level 1*) and diseases (right).

Inflammation-related signatures across diseases and cell types

To identify gene expression programs and regulatory mechanisms across cell types and diseases, we first grouped inflammatory molecules into 21 signatures that delineate multiple processes, including immune cell activation, cellular migration, extravasation, antigen presentation, and cytokine-related signaling (**Supplementary Table 4**). To tailor these signatures to reflect the inflammation landscape of circulating immune cells, we refined these using Spectra, yielding a comprehensive set of 119 cell type-specific factors (**Supplementary Table 4**). We then ran a Univariate Linear Model (ULM) analysis using DecoupleR[42] on the scANVI-corrected gene expression data, providing an inflammation signature activity score for each group. Finally, we ran a Linear Mixed Effect Model (LMEM) between diseased and healthy samples to highlight disease-specific alterations (**Supplementary Table 5**).

We observed a general trend of increased activity in immune-relevant signatures as compared to healthy donors (>50% increased average signature scores; **Fig. 2a**). For IMIDs, we observed the characteristic upregulation of adhesion molecule signatures, TNF via NFkB signaling, antigen cross-presentation and antigen presenting signatures[43,44]. IFN type 1 and 2 signatures were significantly downregulated in most IMIDs and cell types, except for Non-Naive CD8 T cells that showed an upregulation, pointing to a common cell type-specific mechanism[45]. Remarkably, SLE showed a uniquely strong upregulation of the IFN-induced signature in all immune cell types, accompanied by an upregulation of chemokines and chemokine receptors. MS showed a decreased IFN-induced signature and increased chemokine receptor activity, in line with the migratory capacity of blood cells to infiltrate the brain during the course of the disease[46]. As previously reported, we captured the upregulation of the TNF receptor/ligand signature mainly in Non-Naive CD8 T cells for Sepsis (together with an increase in IFNG response in Monocytes), with a decrease in the other inflammatory signals (adhesion molecules and cytokines)[47]. In contrast, all chronic inflammatory diseases upregulated the activity of antigen presenting molecules and increased IFN-induced signaling. This IFN-induced signature was also increased in viral infections, such as flu and COVID-19, while we found a decreased activity in HIV and HBV. Finally, within solid tumors, CRC and NPC presented a strong upregulation of TNF via NFkB signaling. Intriguingly, only RA, PSA, UC and CD showed an enrichment in the T follicular helper (Tfh) signature in Non-Naive CD4 T cells, highlighting the role of circulating Tfh cells in these diseases. In IMIDs more generally, both Naive and Non-Naive CD4 T cell populations were enriched in T helper signatures, pointing to an early priming of Naive T towards helper T cells driven inflammation[48,49]. Finally, to assess the similarity of the inflammatory profiles among diseases, we performed hierarchical clustering of the inflammation signature activity score across all cell types (*Level 1;* **Extended Data Fig. 3a**).

Expanding previous discoveries of a pan-immune IFN-response in SLE patients[27], especially in the myeloid compartment[50], we identified a pronounced IFN activity in SLE across all cell types (**Figure 2a**). To elucidate the regulatory mechanisms and, specifically, transcription factors (TF) driving the IFN activity, we conducted a GRN analysis of IFN-related factors (see **Methods**). This identified STAT1 and SP1 as the primary regulators of the IFN-induced signature, each TF with distinct cell type specificities, suggesting distinct GRNs to regulate the IFN activity in a cell type-specific manner (**Supplementary Table 6**). Monocytes and Non-Naive CD8 T cells showed enrichment for both STAT1 and SP1. In

contrast, Naive CD8 T cells, pDC, DC, and Plasma populations were uniquely enriched for STAT1 activity, whereas Naive CD4 T and ILC populations showed specificity for SP1 activity (**Fig. 2b**). Further examination of STAT1 and SP1 target genes revealed their pronounced upregulation across various cell types and most pronounced in SLE samples (**Fig. 2b; Extended Data Fig. 3b**; **Supplementary Table 6**). Further dissection of the regulatory dynamics in SLE patients at cell state level (*Level 2*) confirmed an increased activity of STAT1 and decreased activity of SP1 in IFN-response Monocytes and IFN-response CD8 T cells compared to other myeloid and CD8 T cell populations (**Extended Data Fig. 3c**)[51]. Interestingly, SP1 was upregulated in classical and regulatory Monocytes, as well as central-memory (CM) and Activated CD8 T cells, which have been described to play a particular role in the development of SLE (p-value Adj <0.05**; Supplementary Table 6**). Finally, STAT1 and SP1 showed distinct activity during disease progression (flare and non-flare patients; **Fig. 2c**; **Supplementary Table 6**). STAT1 activity increased during flares, most notably in CD8 T cell populations, while SP1 activity increased mostly in myeloid populations in the absence of flares. These results provide insights into the distinct regulatory mechanisms driving disease activity in SLE, further highlighting the cell type-specific roles and disease progression dynamics of STAT1 and SP1.

Considering distinct cell types as unique contributors to the inflammatory immune landscape, IFN signatures have been used as a biomarker to define disease activity in autoimmune diseases[52]. However, it remains elusive which immune subpopulations contribute to these signatures to further suggest specific interventions. Observing an enriched IFN type 1 and 2 activity in Non-Naive CD8 T cells in IMIDs, we next seeked to discover subpopulations as the signature driver. Here, we observed a significant upregulation across almost all CD8 T cell populations, however, with a differential pattern across diseases. We then decomposed the signal to gene-level, based on the gene weights, to identify the most relevant contributors (**Extended Data Fig. 3d,e)**. Intriguingly, *FGFBP2 and GZMB* showed increased expression levels, with restriction to specific effector-memory (EM) CD8 T cell subtypes (EM CX3CR1 high, EM CX3CR1 int, Eff HOBIT and Activated; **Fig. 2e**). Of note, *FGFBP2* and *GZMB* were recently described as markers of CD8 T cells localized to areas of epithelial damage, where they potently express IFN, driving the IFN response in the DCs within this niche[45]. To validate that CD8 T cells with intra-epithelial markers were indeed present in peripheral blood, we confirmed their expression in an independent PBMC cohort[53], observing an overexpression of *FGFBP2* and *GZMB* in CD8 T Memory cells of UC patients (p-value Adj < 0.05; **Fig. 2f**). Confirming their overexpression also in intestinal tissue from UC patients (**Fig. 2g**)[54], further supported peripheral blood profiles as biomarker resources for inflammatory events in tissues. Given that CD8 T memory cells in the blood also contain immune memory unrelated to IBD, which can dilute the signal, disease-specific validation at patient-level is still necessary to confirm *FGFBP2* and *GZMB* as biomarkers for IBD.

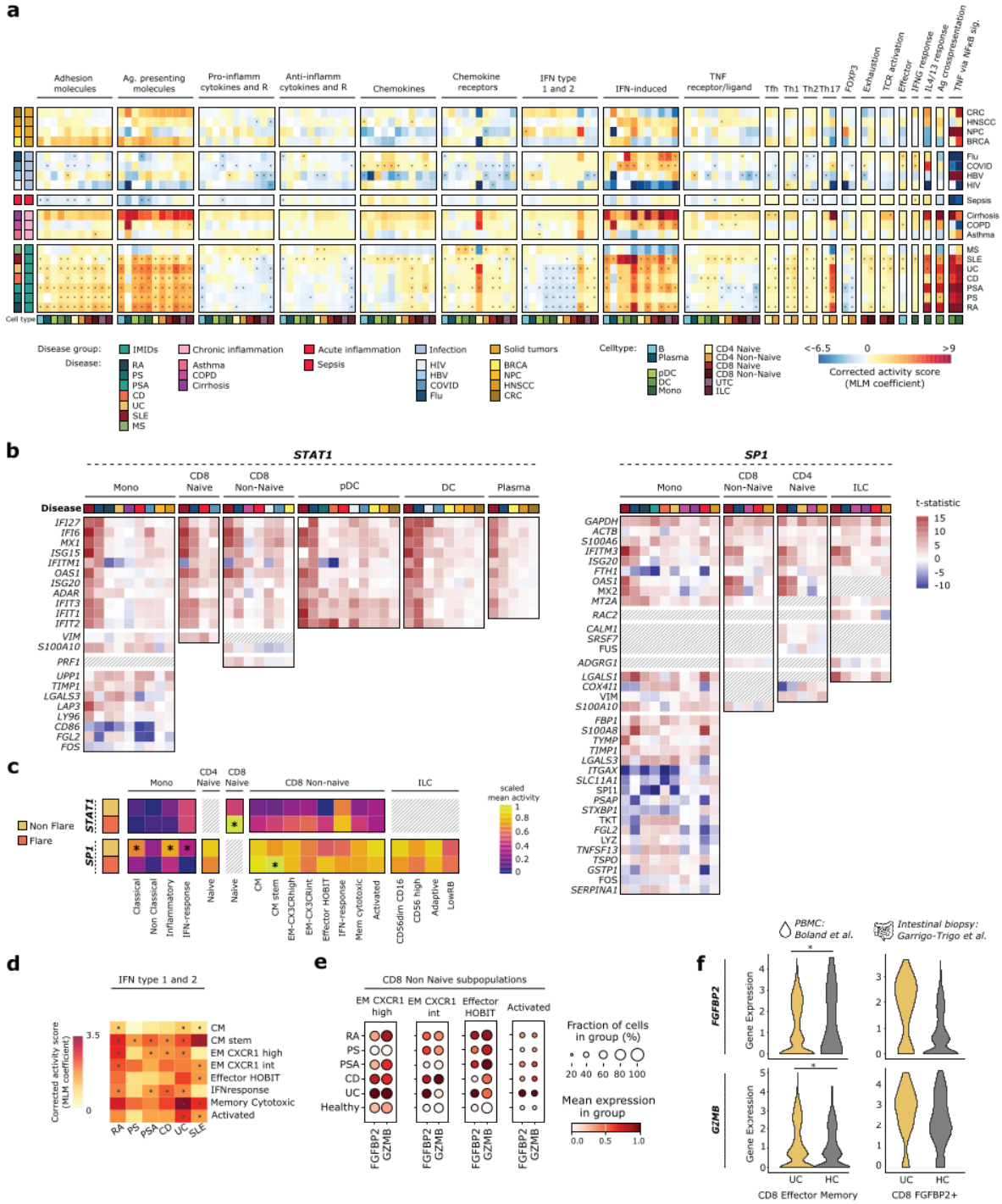


Figure 2. Inflammation-related signatures across cell types and diseases. (**a**) Heatmap displaying the corrected signature activity score of the 119 cell type-specific Spectra factors across diseases and cell types (*Level 1*). Here, the corrected signature activity score represents the coefficient value after running a Linear Mixed Effect Model (LMEM) comparing diseases versus Healthy Control (HC) on the Univariate Linear Model (ULM) estimates computed using the cell-type (*Level1*) and patient pseudobulk on the corrected count matrix. The *x-axis* represents the Spectra cell type-specific factor associated with a given function (*top annotation*). The *y-axis* represents the diseases grouped by disease group. (**b**) Heatmap displaying the transcription factor (TF) specificity of STAT1 and SP1 across different cell types and diseases. The t-statistic represents the differential expression of genes between diseased and healthy samples, highlighting shared genes between TF target genes and IFN-induced cell-type signatures. (**c**) Heatmap representing the average scaled TF activity of STAT1 and SP1 across *level 2* population for Flare and non flare patients from *Perez et al*[27]. (**d**) Heatmap displaying the corrected IFN Type 1 and 2 signature activity score across Non-Naive CD8 T cells (*Level 2*) across IMID diseases. Here, the corrected signature activity score represents the coefficient value after running a LMEM comparing diseases versus Healthy Control (HC) on the ULM estimates computed using the cell-type (*Level2*) and patient pseudobulk on the corrected count matrix. (**e**) Dotplot showing the uncorrected average expression of the *FGFBP2* and *GZMB* genes from IFN Type 1 and 2

signature (*x-axis*) across different subpopulations of Non-Naive CD8 T cells (*Level 2*) on IMID diseases and health (SCGT00 study). The dot size reflects the percentage of cells of each disease expressing each gene, and the color represents the average expression level. (**f**) *FGFBP2* and *GZMB* expression levels in CD8 EM T cells from an independent scRNA-seq PBMC validation cohort of UC (n=7) and healthy (n=8) donors from *Boland et al.*[53] (**g**) *FGFBP2* and *GZMB* expression levels in CD8 FGFBP2+ T cells from scRNA-seq tissue biopsies from UC (n=6) and healthy (n=6) donors from *Garrigo-Trigo et al.*[54] For panels (**a**) and (**d**), significant signature activity differences between disease and HC are marked with a dot (·) (p-value Adj < 0.05). For panels (**c**) and (**f**), asterisk (*) indicates statistical significant changes using the Wilcoxon signed-rank test and corrected for multiple testing (p-value Adj < 0.05).

Functional biomarker selection through interpretable modeling

Biomarker discovery using linear models (such as the above applied ULM) or standard differential expression analysis suffers from the limitation that genes are considered independently. Thus, we considered the possibility of categorizing cells to their respective disease origin through an interpretable machine learning pipeline, to guide the selection of functional disease discriminatory genes and biomarkers (starting from 935 genes, see **Methods; Supplementary Table 4**). Therefore, we next applied a supervised classification approach, together with a post-hoc interpretability method, to allow the inference of the gene-wise importance, stratified by disease and cell type (*Level1)*. We based our strategy on Gradient Boosted Decision Trees (GBDTs), a state-of-the-art machine learning technique proven to be effective in complex tasks with noisy data and non-linear feature dependencies[55]. GBDTs iteratively build an ensemble of decision trees, by trading the complexity of the model (i.e. the number of trees) with its generalization power. Here, we used the XGBoost library[56] and tuned its hyperparameters through the Optuna library[57] (see **Methods**). As GBDTs require post-hoc interpretability tools in order to infer explanations, we computed SHAP (SHapley Additive exPlanation) values[58], shown to provide attributions that are locally consistent and that can be aggregated into global explanations. SHAP values explained the output of the classifier, in our case the predicted conditions, as a sum of contributions of each feature, that is, the gene expression profiles, while also considering its interactions with other genes.

Combining the two approaches, we were able to rank genes based on their importance, stratified by cell type and disease. We executed the analysis considering each cell type (*Level 1*) independently, resulting in gene rankings based on their discriminative power to distinguish a certain disease over the others. Such strategy mitigated the impact of cell-specific expression profiles and allowed the interrogation of differential gene importance to distinguish diseases in distinct cell types. We applied the pipeline on the corrected gene expression profiles after scANVI integration, but also tested uncorrected log-normalized data as an input. Overall, we achieved high performance to assign each cell to the correct disease label with a Balanced Accuracy Score (BAS) of 0.87 and a Weighted F1 score (WF1) of 0.90, computed on a test set of 20% cells, when starting from scANVI-corrected data (**Fig. 3a)**. Instead, log-normalized counts resulted in decreased BAS and WF1, highlighting the improvement provided after batch correction (0.65 and 0.78; **Fig. 3a**). Performances were consistent among cell types, with less abundant cell populations obtaining generally lower scores (e.g., Plasma, BAS = 0.78, WF1 = 0.80; **Extended Data Fig. 4**). Noteworthy, flu obtained the lowest recall (0.03), mainly being classified as COVID-19 (74% of cells), likely due to the low cell numbers (flu accounts for ~0.004% of cells). When

validating our approach by predicting the inflammatory disease label of unseen studies, the SHAP gene rankings provided a better feature selection than random gene sets (**Extended Data Fig. 10**).

Leveraging the interpretable machine learning pipeline and the gene sets for biomarker discovery, resulted in a rich resource of prioritized discriminative genes associated with disease and cell types. Ordering the genes by their importance within immune cell types (*Level1*) and diseases identified previously described biomarkers, such as *STAT3* in CD4 T cells for RA samples[59] and IFN genes in multiple cell types for SLE samples[50] (**Extended Data Fig. 5a**). The SHAP values of *CYBA*, stood out as a strong candidate marker to classify diseases affecting barrier tissue, namely Psoriatic Arthritis (PSA), Psoriasis (PS), Ulcerative colitis (UC), Crohn's Disease (CD), Chronic Obstructive Pulmonary Disease (COPD) and Asthma (**Fig. 3b; Extended Data Fig. 5b, 9**). *CYBA* encodes the primary component of the microbicidal oxidase system of phagocytes. In line, the importance of the gene across diseases was seen mainly in Monocytes (**Extended Data Fig. 5c, 9**). Interestingly, high expression of CYBA drove the model to classify intestinal and pulmonary inflammatory diseases (UC, CD, COPD and Asthma), whereas reduced levels of expression were relevant to classify skin-related diseases (PS and PSA; **Fig. 3b; Extended Data Fig. 5b**). Mutations in *CYBA* cause an autosomal recessive chronic granulomatous disease and patients show an impaired phagocyte activation and fail to generate superoxide. Consequently, patients show recurrent bacterial and fungal infections in barrier tissues, including the skin[60]. Thus, we hypothesize that reduction of *CYBA* in skin-related IMIDs leads to an impaired immune barrier function and frequent recurrent infections causing localized, symptomatic flares of PS and PSA. Reactive Oxidative Species (ROS) produced by mucosa-resident cells or by newly recruited innate immune cells are essential for antimicrobial mucosal immune responses and defense against pathogenic attack[60]. In UC, CD, COPD and Asthma, an upregulation of *CYBA* may result in the accumulation of ROS, a hallmark of the diseases. In an independent scRNA-seq validation cohort, we confirmed the overexpression of *CYBA* in peripheral Monocytes from UC patients compared to healthy donors (p-value Adj <0.05; **Fig. 3c**), further validated with an upregulation of *CYBA* in intestinal tissue macrophages from UC and CD patients (p-value Adj <0.05; **Fig. 3c**)[61]. Specifically, in inflammatory bowel's disease tissue biopsies, multiple monocyte-derived populations that infiltrate tissues from circulation and tissue-resident phagocytic macrophages showed a significant upregulation of *CYBA* expression compared to healthy controls (p<0.05; **Extended Data Fig. 5d**). Considering that SHAP values rank gene importance based on interactions with other genes, validation in a larger patient cohort is required to confirm *CYBA* as a reliable biomarker for IBD.

Further exploring SHAP value ranks, highlighted the importance of *IFITM1* for all chronic diseases (COPD, Asthma and Cirrhosis; **Extended Data Fig. 5e, 9**). *IFITM1* inhibits viral entry into host cells by preventing the fusion of the viral membrane with the host cell membrane[62]. The importance of *IFITM1* was mainly observed in lymphoid cells, with the high expression of *IFITM1* specifically in Non-Naive CD4 T cells and ILCs driving the model to classify COPD (CD4/CD8 T cells and ILCs; **Fig. 3d; Extended Data Fig. 5f, 9**). In line, ILC and T cell accumulation is associated with the decline of lung function and severity in COPD patients. T cells also mediate autoimmune responses by inducing the production of IgG autoantibodies in B cells of COPD patients[63]. We hypothesize chronic inflammation to trigger higher expression of *IFITM1* in lymphoid cells, which further facilitates their accumulation[64]. In

Asthma, a lower expression of *IFITM1* in Non-Naive CD4 T cells and ILCs drove the classification of the disease, the former potentially being associated with Th2-mediated allergic asthma **(Extended Data Fig. 5f, 9)**[65]. Also Cirrhosis classification was driven by lower *IFITM1* expression in Non-Naive CD4 and CD8 T cells (**Fig. 3d**), an event validated in liver tissue from cirrhotic patients[67] that presented significant lower expression of *IFITM1* specifically in Non-Naive CD8 T cells (p-value Adj <0.01; **Fig. 3e**). Downregulation of *IFITM1* could be a consequence of a prolonged inflammation and exposure to stimuli leading to a state of dysfunction of CD8 T cells in Cirrhosis[67].

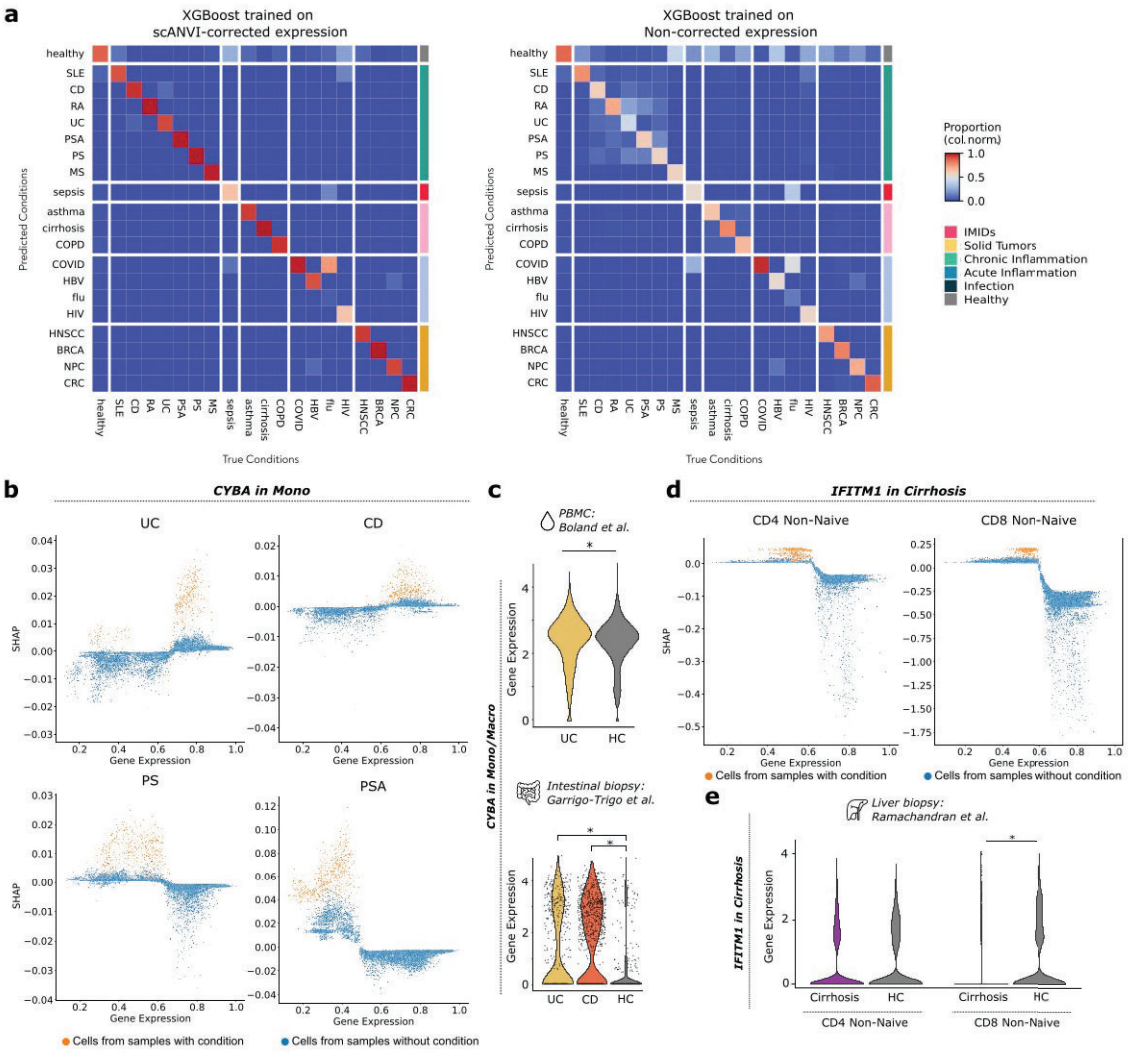


Figure 3. Functional biomarker discovery using interpretable machine learning. (**a**) Normalized confusion matrices displaying proportion of predictions belonging to each true condition. Diagonal values correspond to the Recall metric. XGBoost was trained on the scANVI batch corrected **(left)** or batch-uncorrected **(right)** log-scaled cell expression profiles. (**b**) Scatter plot of max-normalized gene expression against SHAP values computed for *CYBA* gene on Monocyte population (annotation *Level1*) and considering the output of XGBoost for a given disease (UC, CD, PS and PSA, from left to right, top to bottom). (**c**) *CYBA* expression levels in an independent single-cell PBMC validation cohort[53] on Monocytes from UC (n=7) and healthy (n=8) donors **(top)** and on intestinal biopsies[54] from CD (n=6), UC (n=6) and healthy (n=6) patient biopsies, focusing only on Monocyte-derived and macrophage cells **(bottom)**. **(d)** Scatter plot of max-normalized gene expression against SHAP values computed for *IFITM1* gene on Non-Naive CD4 (left) CD8 (right) T populations (annotation *Level1*) considering the output of the XGBooost for the cirrhosis disease. (**e**) *IFITM1* expression levels in an independent scRNA-seq dataset from liver tissue biopsies of cirrhotic (n=5) and healthy (n=5) donors[67], focusing on Non-Naive CD4 **(left)** and CD8 **(right)** T populations. In panels (**c**) and (**e**), asterisks (*) indicates statistical significant changes using the Wilcoxon signed-rank test. In Panels (**b**) and (**d**), we limited the visualization to up to 60,000 cells, sampling an equal percentage from each patient corresponding to 5%, 7.5% and 7.5% of Monocytes, Non-Naive CD4 T and Non-Naive CD8 T cells, respectively. Cells belonging to samples with or without the given condition (disease) are marked in orange or blue respectively.

Classifying Unseen Patients by Reference-Mapping

The ability to accurately classify cells belonging to different cell types according to their respective diseases prompted us to classify patients based on their disease of origin, creating the basis for a universal classifier as a precision medicine tool for inflammatory diseases. Single-cell information layed a foundation for better understanding the diversity and traits within the populations, but classifying new patients remains a challenge due to data sparsity, noise, and batch-effects across studies. By considering each patient as an ensemble of expression profiles across all cells, we learned a generative model while integrating the single-cell reference as a basis to project new patients into the same embedding space. Such strategy allowed us to map unseen and unlabeled query patient data into our reference embedding space, providing a common ground for downstream classification.

Projecting expression data into a lower dimensional space is a common strategy to reduce noise[68] and to map query data into a reference atlas[69]. Here, we propose a novel computational framework to exploit the cell embedding for classification of patients into conditions (e.g. inflammatory diseases), thus, turning the single-cell reference into a diagnostic tool (**Fig. 4a and Extended Data Fig. 8**). Therefore, we first generated the embeddings with scANVI of both the reference and the unseen query datasets, while also transferring the cell labels to the latter. Then, we defined a cell type pseudobulk profile per patient by averaging the embedded features of the corresponding cells (*Level1*; see **Methods**). Next, we trained an independent classifier to assign correct disease labels, considering one cell type at a time. We handled uncertainty at cell type level via majority-voting system to determine most frequent conditions. To assess the performance of our framework, we proposed three scenarios: 1) a 5-fold cross-validation splitting the full reference atlas into five balanced sets, 2) a Validation dataset with unseen patients, and 3) an External dataset with unseen studies (**Fig. 4b**). The External dataset has been excluded from our Data Processing pipeline before quality control, and the Validation dataset immediately after, allowing us to evaluate different levels of reference mapping complexity.

Strikingly, our classification strategy achieved high performance in the cross-validation scenario, resulting in 0.90±0.03 WF1 (minimum 0.87; **Fig. 4c**) and 0.85±0.07 BAS (minimum 0.79; **Extended Data Fig. 6e**), averaged across five independent runs. Such high performances strongly suggest gene expression profiles of inflammatory diseases to be separated in the embedded space and, more importantly, immune cells in circulation to be capable of serving as a source for patient classification. Consistently with results obtained from functional biomarker selection analysis, Flu was the only disease that failed to be classified (Recall: 0.18). Training a classifier for each cell type separately allowed us to assess their relevance in distinguishing inflammatory diseases, particularly for diseases with lower overall performance scores (**Fig. 4d,e**). While certain diseases (COVID, COPD and Asthma) were particularly well classified by lymphoid and myeloid cell types, HIV was best classified by naive lymphoid cells (i.e., Naive CD4 and CD8 T cells and B with F1 of 0.83), while dendritic cell types (i.e. DC, and pDC) did not allow correct disease assignment (0.29). Overall, Plasma and UTC showed the lowest BAS (0.53 and 0.67) and WF1 (0.64 and 0.78), highlighting the strength of our majority voting approach. All in all, these results show the strength of an integrated atlas and comprehensive model of inflammatory diseases to holistically model predictive features across cell types and diseases.

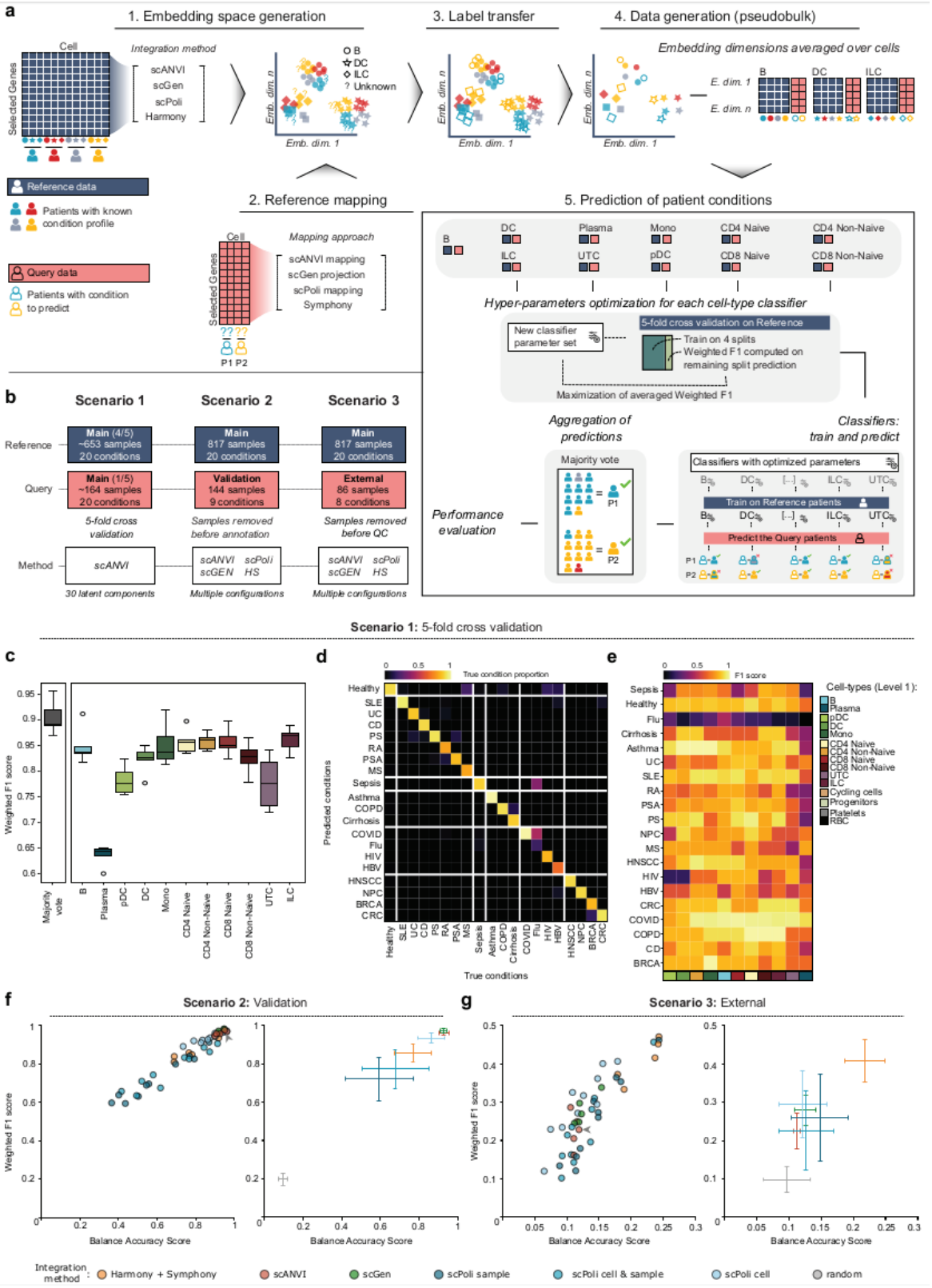


Figure 4. Schematic representation of the patient classifier pipeline and performance evaluation. (**a**). Schematic representation of the patient classifier pipeline. (**b**) Description of the three performance evaluation scenarios. In our datasets we always have only one sample for each patient. (**c**) Distribution of Weighted F1-scores for each left-out split in Scenario 1's cross-validation. (**d**) Normalized confusion matrices displaying proportion of predictions belonging to each true condition after aggregating all the predictions of the left-out folds. Diagonal values correspond to the Recall metric. **(e)** F1-score for each combination of cell-type and disease, after aggregating all the predictions of the left-out folds. (**f-g**) Performance evaluation from Scenario 2 and 3, respectively, showing (**left**) the distribution of Weighted F1 and Balanced Accuracy Score for all the configurations of each data integration approach (Arrows highlight scANVI configuration applied in Scenario 1), and (**right**) the mean and standard-deviation of each data integration method.

Increasing the complexity by classifying unseen patient samples of the Validation cohort, the performance remained very high, with a BAS of 0.95 and a WF1 of 0.98. However, the classification of External cohorts resulted in strongly decreased BAS of 0.12 and a WF1 of 0.23 (see **Extended Data Fig. 7,11**). To improve the mapping of unseen studies, we next investigated whether alternative integration approaches and hyperparameter tuning were able to learn a more robust embedding. Therefore, we applied scGen[70] (post-hoc batch-effect correction after training), scPoli[71] (directly learning also patient-wise embeddings), and Harmony/Symphony[72,73] (linear approach not based on Variational Autoencoders; VAEs). All newly introduced methods achieve high performance in the Validation dataset, with scANVI, scPoli and scGen achieving similar high BAS (>0.94) and WF1 (>0.97), followed by Harmony (BAS: 0.92 and WF1: 0.94; **Fig. 4f, Extended Data Fig. 7, 11**), considering the best performing configuration. While all the approaches lose predictive power on the External datasets, Harmony performed best with a BAS of 0.24 and a WF1 of 0.47. Interestingly, scPoli obtained the best result on External data when the learned patient-wise embedding was used to classify patients (BAS: 0.24 and WF1: 0.46), but the same model underperforms on the Validation data (BAS: 0.37 and WF1: 0.60). Since we evaluated different hyperparameters for each approach, we were able to estimate a method's sensitivity to the chosen configuration. Here, Harmony excels as the best performing and tunable method on the External dataset, with BAS of 0.22±0.03 and WF1 of 0.41±0.06), among the considered hyperparameters.

Discussion

Mapping the plasticity of the immune cells in circulation is now possible by using sequencing technologies that allow an unbiased immuno-phenotyping of single cells[75,76]. Importantly, such transcriptome-wide methods do not require previous knowledge, but leverage the random sampling of cells, and transcripts within cells, to derive comprehensive cellular landscapes. Recent technologies enable the sampling of thousands of cells per sample and hundred-thousands per patient cohort, pushing the resolution towards fine-grained cellular maps and increasing the power to identify disease-specific states and programs[76]. To date, single-cell sequencing has been applied to a multitude of inflammatory diseases to determine alterations in cell type composition and to pinpoint disease-driving mechanisms as potential therapeutic targets[20]. However, a complete map of immune cell states across diseases, holistically charting immune plasticity in inflammatory diseases, has been elusive. We reasoned that integrating single-cell transcriptome maps of cells in circulation across a variety of diseases and millions of cells would allow us to extract the full spectrum of inflammatory processes and to generate a comprehensive model of inflammation in circulating immune cells.

Our strategy split the analysis into three phases, starting with the supervised extraction of inflammation-related patterns (gene signature and GRN activity), followed by the discovery of interpretable biomarkers (features) and finally patient classification (projection). Especially the latter presents an intriguing avenue towards personalized medicine strategies to manage and monitor inflammatory diseases and to move beyond current clinical practice using single biomarker strategies (e.g. CD4 T lymphocyte count in HIV)[77]. Here, the size of the datasets and the depth of cellular resolution, combined

with an advanced machine learning framework, allowed the separation of diseases in the latent space and eventually the correct classification of unseen patient samples across disease categories and entities. Hence, we provide strong evidence for the potential of using immune cells in circulation as liquid biopsy for diagnostics, when combined with machine learning models that holistically learned the full spectrum of immune cell variability across patients and diseases[71]. Such models represent the cornerstone to build classifiers for inflammatory diseases as precision diagnostics tools for the clinic[78]. We note though that the presented evidence provides the basis for establishing blood-based patient classification and stratification with the following limitations.

The results obtained with 5-fold cross-validation and with Validation dataset proved that our computational framework is able to learn a cell embedding that can be exploited to classify patients with inflammatory conditions. However, scANVI, the main integration method applied in this work, was not able to correctly map External studies on the reference atlas. We suspected that the poor performances can be related to a suboptimal choice of integration parameters. Therefore, we tested scANVI, together with scGen, scPoli and Harmony/Symphony with different parameter configurations. Our results pointed to Harmony/Symphony as the most robust method to integrate and query reference and external datasets, respectively. On the other hand, scPoli was both the best and poorest performer on External data and also showed a vast variability in the Validation dataset. Indeed, scPoli is one of the most complex methods that require to select the number of latent dimensions for cell-wise and patient-wise embedding space, with highly affected performances. While linear approaches have less representation power than VAEs, they are also less prone to overfitting and more robust to the hyperparameter choice. Hence, in settings where hyperparameter tuning and validation is not possible due to the lack of condition labels, as in clinical diagnostics, tools like Harmony/Symphony might be preferable to more complex VAEs.

Our results showed patients' gene expression profiles to be separated in the embedded space and immune cells in circulation to provide sufficient information to enable liquid biopsy diagnostic tool development. Nevertheless, none of the methods was able to classify external samples with high performances, likely related to uncorrected batch effects between chemistries and studies. Although we considered covariates, such as gender and age, including additional differences between patients and studies could further improve the classification performance (e.g., treated status and type, comorbidities, disease subtype and duration). Also, incorporating additional data modalities, such as cell composition or epigenetic states, could further enlarge the feature space and improve patient classification. In the same line, advanced single-cell technologies that capture additional blood cell types, such as neutrophils, could contribute further disease biomarkers. Towards leveraging single-cell technologies in the diagnostic context, we call for the definition of best practices and quality control standards to reduce batch effects and the generation of large training datasets in highly controlled conditions. While we focus here on optimizing the outcome of classifying external datasets, extensive benchmarks of different integration methods and parameter configurations should be performed on additional datasets, including multiple tissues, to systematically evaluate their performance in classifying samples into specific disease conditions. Such investigation could complement similar studies[24] that already assessed performances of integration approaches at the cell level.

However, the concept of using immune cells as a sensor for diseases is highly intriguing and opens the door for the development of future universal diagnostic tools[79]. For some of the tested diseases, such as sepsis, current biomarker strategies may already provide sufficient sensitivity considering the pathology and clinical manifestation of the disease (i.e. lymphopenia as a hallmark of sepsis)[47]. Nevertheless, for diseases such as in rheumatology and IBD, many patients are undiagnosed or diagnosed as False Positive, and more accurate universal tools are needed[80,81]. The here established classification framework was capable of classifying most diseases correctly using a majority-voting strategy across all cell types. However, for certain disease types, such as HIV, a tailored, cell type-centric approach may lead to improved results. HIV showed an increased classification score using lymphoid cell types alone, in line with the tropism of the virus infecting mainly CD4 T cells[82,83].

While the machine learning strategy developed for patient stratification was not interpretable, our complementary approach using GBDT, together with SHAP, and a tailored list of immune cell molecules with defined function, provided explainable results and a rich resource for disease-discriminative gene and biomarker discovery[55,58]. High absolute SHAP values highlight genes that are important to characterize the inflammatory landscape and top ranked genes have a high discriminative power between diseases. Here, *CYBA*, the light, alpha subunit of microbicidal oxidases in phagocytes showed the strongest importance mainly through high SHAP score in Monocytes. Barrier site diseases scored the highest, with intriguing opposite directions in intestinal (high; UC and CD) and lung (high; COPD and Asthma) as compared to skin (low; PS and PSA) diseases. We went on to validate *CYBA* overexpression in a PBMC validation cohort[53] and patient biopsies[61] using scRNA-seq. The increased expression in Monocytes as well as tissue-resident macrophages in UC and CD patients, suggests a role in the disease pathology, potentially through its oxidase function to produce superperoxide and ROS[84]. Importantly, UC mouse models treated with superoxide dismutase showed significantly attenuated UC disease burden in a dose-dependent manner and reduced lipid peroxidation in colonic tissue. Simultaneously, leukocyte rolling and adhesion in colonic venules of colitis rats were significantly reduced, contributing to strongly reduced inflammatory phenotypes[85]. Next, *IFITM1*, a gene involved in inhibiting viral entry into host cells, emerged as an important gene in Cirrhosis, COPD, and Asthma. We validated *IFITM1* downregulation in Non-Naive CD8 T cells in liver tissue from cirrhotic patients[67], suggesting its potential role as a disease biomarker. In the case of COPD and Asthma, although *IFITM1* showed a discriminative power in distinguishing the two diseases, no significant expression differences were observed compared to healthy patients. Hence, disease-specific validation is needed to confirm a gene with high discriminative power between diseases to also be considered as a disease biomarker for clinical diagnostics.

Bringing reference atlases and data resources into the clinics is challenging without providing clear examples and strategies for their implementation. We generated a comprehensive landscape of inflammation in circulating immune cells from acute and chronic inflammatory diseases. Using advanced machine learning pipelines, we developed interpretable models for biomarker identification that can be further validated as a stand-alone or combinatory diagnostic test. On the other hand, we classified diseases based on generative models that learned the full inflammatory feature space across cell types and disease, laying the foundation for a universal diagnostic tool for inflammatory diseases.

Declarations

Ethics approval and consent to participate

Human blood processed in-house for this project was pre-selected and included within other ongoing studies. All the studies included were conducted in accordance with ethical guidelines and all patients provided written informed consent. Ethical committees and research project approvals for the different studies included in this manuscript are detailed in the following text.

SCGT00 was approved by Hospital Universitari Vall d'Hebron Research Ethics Committee (PR(AG)144/201). **SCGT01** received the IRB approval by the Parc de Salut Mar Ethics Committee (2016/7075/I). **SCGT02** received the ethics approval by the Medisch-Etische Toetsingscommissie (METc) committee; for asthma patients (ARMS and ORIENT projects – NL53173.042.15 and NL69765.042.19 respectively), for COPD patients (SHERLOck project, NL57656.042.16), and finally, healthy controls (NORM project, NL26187.042.09). **SCGT03** was approved by the Comité Ético de Investigación con Medicamentos del Hospital Universitario Vall d'Hebron (654/C/2019). **SCGT04** and **SCGT06** was approved by the Comitè d'Ètica d'Investigació amb medicaments (CEim) del Hospital de la Santa Creu i Sant Pau (EC/21/373/6616 and EC/23/258/7364). **SCGT05** was approved by the institutional review boards of the Commissie Medische Ethiek UZ KU Leuven/Onderzoek (S66460 and S62294).

Data availability

Single-cell RNA-sequencing (scRNA-seq) in-house generated data and associated count matrices are accessible at Sequence Read Archive (SRA), NCBI Gene Expression Omnibus (GEO) and European Genome Archive (EGA) databases under the following accession numbers: PRJNA1045296 (SCGT00 and SCGT00val), GSE248689 (SCGT01), GSE248695 (SCGT02), GSE248685 (SCGT03), GSE248693 (SCGT04), EGAS50000000590 (SCGT05) and GSE270165 (SCGT06). Previously published scRNA-seq data included in this project, either FASTQ files or processed count matrices, were obtained from GEO, BioStudies Array Expresse, Broad Institute DUOS, Synapse, Genome Sequence Analysis (GSA), CellXGene Data Portal, and 10X Genomics. Further details specified on the **Supplementary Table 1 - Sheet 1**.

The processed scRNA-seq datasets and metadata analyzed in the current study have been deposited at Zenodo: https://zenodo.org/records/13374575.

Code availability

The code to reproduce the full analysis presented in this article is hosted in the Github repository: https://github.com/Single-Cell-Genomics-Group-CNAG-CRG/Inflammation-PBMCs-Atlas

Author information

LJG, JCN and HH conceived the project. JCN and HH supervised the project. LJG, DM, SAF, FC, MBoulougouri and JLB performed the computational and statistical analysis. RORF, HAT, RN and JN provide strategic advice on computational tasks and data interpretation. AA and MBanchero generated VCF files for data demultiplexing. ASM and AMC provided processed objects for validation analysis. SR, DM, GC, MB, and YH generated datasets. TK, PvdV, FA, FP, GS, ER, MvdB, AS, JMC, AFN, ED, JC, JT, JPG, EC, GG, BS, AJ, VS, RE, ST, SV, MCN and SM provided patient samples. LJG, DM, SAF, FC, JSR, PV, ACV, JCN and HH interpreted the results. LJG, DM, SAF, FC, JCN and HH wrote the manuscript with input from all the authors. All authors read and approved the current version of the manuscript.

Acknowledgements

The authors would like to thank the helpful support received by the authors from publicly available data[87,88] used in the current study by providing processed data in the optimal format. Additionally, we appreciate the great effort put in creating the DISCO database[88] with multiple-tissue atlases using single-cell datasets. The authors would like to thank the CNAG Scientific IT Unit and the maintainers of the CNAG compute cluster for providing assistance with essential computing resources. Also, the authors would like to thank all the team members, in particular Paula Nieto, Helena L. Crowell, Marc Elosua-Bayes, Mohamed Abdalfatah, and Ramon Massoni-Badosa, for their contributions during brainstorming and discussion sessions.

Funding

This project has received funding from the European Union's H2020 research and innovation program under grant agreement No. 848028 (DoCTIS; Decision On Optimal Combinatorial Therapies In Imids Using Systems Approaches). L.J.-G. has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Universities. D.M. is supported by the Juan de la Cierva Fellowship (JDC2022-049637-I) from the Spanish Ministry of Science and Innovation and the European Union "NextGenerationEU"/PRTR. F.C. is funded by the Swiss National Science Foundation (SNSF) grant No CRSII5_205884/1. M.Boulougouri is funded by the Graph Neural Networks for Explainable Artificial Intelligence ERA-NET + EJP (20CH21_195579) grant. J.L.B. is supported by the Spanish Ministry of Universities through Margarita Salas fellow (RSUC.UDC.MS06). R.O.R.F. is supported by DFG through CRC/SFB 1550 "Molecular Circuits of Heart Disease" Y.H. is supported by a Junior Postdoctoral fellowship from the Research Foundation Flanders (FWO 12D5823N). S.T. is supported by a BOF-Fundamental Clinical Research mandate (FKO) from KULeuven and by the Belgian Foundation Against Cancer (FAF-C/2018/1301). V.S. is funded by Asociación Española Contra el Cáncer (AECC). M.C.N. acknowledges funding from GSK, the Netherlands Lung Foundation (project No. 4.1.18.226) and the European Union's H2020 Research and Innovation Program under grant agreement No. 874656

(discovAIR). This collaboration project is co-financed by the Ministry of Economic Affairs and Climate Policy by means of the PPP-allowance made available by the Top Sector Life Sciences & Health to stimulate public-private partnerships. A.S. is funded by PID2021-123918OB-I00 from MCIN/AEI/51 10.13039/501100011033 and co-funded by “FEDER: A way to make Europe”. Part of the computational analyses were supported by the Google Cloud Research Credits program with the award GCP19980904.

Competing interests

H.H. is co-founder and shareholder of Omniscope, scientific advisory board member of Nanostring and MiRXES and consultant to Moderna and Singularity. J.C.N. is scientific consultant to Omniscope. V.S. has received research grants from AstraZeneca and honoraria from GSK unrelated to this study. M.v.d.B has received research grants (unrestricted) from AstraZeneca, Novartis, GlaxoSmithKline, Roche, Genentech, Chiesi and Sanofi. M.N. has been awarded with research grants (unrestricted) from AstraZeneca and GSK. A.S. is the recipient of research grants from Roche-Genentech, Abbvie, GSK, Scipher Medicine, Pfizer, Alimentiv, Inc, Boehringer Ingelheim and Agomab; receives consulting fees from Genentech, GSK, Pfizer, HotSpot Therapeutics, Alimentiv, Origo Biopharma, Deep Track Capital, Great Point Partners and Boxer Capital; and is on the advisory boards of BioMAdvanced Diagnostics, Goodgut and Orikine. A.A. is a computational biologist at IMIDomics, Inc. A.J. is the chief data scientist at IMIDomics, Inc. S.M. is the co-founder and CMO at IMIDomics, Inc. J.S.R. reports funding from GSK, Pfizer and Sanofi and fees/honoraria from Travere Therapeutics, Stadapharm, Astex, Pfizer, Grunenthal and Owkin.

References

1. Medzhitov, R. The spectrum of inflammatory responses. *Science* **374**, 1070–1075 (2021).
2. Newton, K., Dixit, V. M. & Kayagaki, N. Dying cells fan the flames of inflammation. *Science* **374**, 1076–1080 (2021).
3. Casanova, J.-L. & Abel, L. Mechanisms of viral inflammation and disease in humans. *Science* **374**, 1080–1086 (2021).
4. Agirman, G., Yu, K. B. & Hsiao, E. Y. Signaling inflammation across the gut-brain axis. *Science* **374**, 1087–1092 (2021).
5. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454**, 428–435 (2008).
6. Netea, M. G. *et al.* A guiding map for inflammation. *Nat. Immunol.* **18**, 826–831 (2017).
7. Roe, K. An inflammation classification system using cytokine parameters. *Scand. J. Immunol.* **93**, e12970 (2021).
8. Eltzschig, H. K. & Carmeliet, P. Hypoxia and Inflammation. *N. Engl. J. Med.* **364**, 656–665 (2011).
9. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. *FEBS J.* **285**, 2944–2971 (2018).
10. Soehnlein, O. & Libby, P. Targeting inflammation in atherosclerosis — from experimental insights to the clinic. *Nat. Rev. Drug Discov.* **20**, 589–610 (2021).
11. Psarras, A., Wittmann, M. & Vital, E. M. Emerging concepts of type I interferons in SLE pathogenesis and therapy. *Nat. Rev. Rheumatol.* **18**, 575–590 (2022).
12. Penkava, F. *et al.* Single-cell sequencing reveals clonal expansions of pro-inflammatory synovial CD8 T cells expressing tissue-homing receptors in psoriatic arthritis. *Nat. Commun.* **11**, 4767 (2020).
13. Manthiram, K., Zhou, Q., Aksentijevich, I. & Kastner, D. L. The monogenic autoinflammatory diseases define new pathways in human innate immunity and inflammation. *Nat. Immunol.* **18**, 832–842 (2017).
14. Zaiss, D. M. W., Pearce, E. J., Artis, D., McKenzie, A. N. J. & Klose, C. S. N. Cooperation of ILC2s and TH2 cells in the expulsion of intestinal helminth parasites. *Nat. Rev. Immunol.* 1–9 (2023) doi:10.1038/s41577-023-00942-1.
15. Cao, L. L. & Kagan, J. C. Targeting innate immune pathways for cancer immunotherapy. *Immunity* **56**, 2206–2217 (2023).
16. Dinarello, C. A., Simon, A. & van der Meer, J. W. M. Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. *Nat. Rev. Drug Discov.* **11**, 633–652 (2012).
17. Koenig, L. M. *et al.* Blocking inflammation on the way: Rationale for CXCR2 antagonists for the treatment of COVID-19. *J. Exp. Med.* **217**, e20201342 (2020).
18. Țiburcă, L. *et al.* The Treatment with Interleukin 17 Inhibitors and Immune-Mediated Inflammatory Diseases. *Curr. Issues Mol. Biol.* **44**, 1851–1866 (2022).
19. Taquet, M. *et al.* Acute blood biomarker profiles predict cognitive deficits 6 and 12 months after COVID-19 hospitalization. *Nat. Med.* **29**, 2498–2508 (2023).
20. Dann, E. *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* 1–11 (2023) doi:10.1038/s41588-023-01523-7.
21. Rauber, S. *et al.* Resolution of inflammation by interleukin-9-producing type 2 innate lymphoid cells. *Nat. Med.* **23**, 938–944 (2017).
22. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
23. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
24. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
25. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* 1–12 (2023) doi:10.1038/s41587-023-01940-3.
26. Cho, Y.-N. *et al.* Mucosal-Associated Invariant T Cell Deficiency in Systemic Lupus Erythematosus. *J. Immunol.* **193**, 3891–3901 (2014).
27. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
28. Thangjam, N., Dey, B., Khonglah, Y., Tiewsoh, I. & Ksoo, R. Natural Killer Cell Count in

Systemic Lupus Erythematosus Patients: A Flow Cytometry-Based Study. *Cureus* **15**, e46885 (2023).
29. Ju, J. K. *et al.* Activation, Deficiency, and Reduced IFN-γ Production of Mucosal-Associated Invariant T Cells in Patients with Inflammatory Bowel Disease. *J. Innate Immun.* **12**, 422–434 (2020).
30. Adlowitz, D. G. *et al.* Expansion of Activated Peripheral Blood Memory B Cells in Rheumatoid Arthritis, Impact of B Cell Depletion Therapy, and Biomarkers of Response. *PLOS ONE* **10**, e0128269 (2015).
31. Cairns, A. P., Crockard, A. D. & Bell, A. L. The CD14+ CD16+ monocyte subset in rheumatoid arthritis and systemic lupus erythematosus. *Rheumatol. Int.* **21**, 189–192 (2002).
32. IJMS | Free Full-Text | Rheumatoid Arthritis: Pathogenic Roles of Diverse Immune Cells. https://www.mdpi.com/1422-0067/23/2/905.
33. Drewry, A. M. *et al.* Persistent Lymphopenia After Diagnosis of Sepsis Predicts Mortality. *Shock* **42**, 383 (2014).
34. Moir, S. & Fauci, A. S. B cells in HIV infection and disease. *Nat. Rev. Immunol.* **9**, 235–245 (2009).
35. Turner, M. D., Nedjai, B., Hurst, T. & Pennington, D. J. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochim. Biophys. Acta BBA - Mol. Cell Res.* **1843**, 2563–2582 (2014).
36. Pishesha, N., Harmand, T. J. & Ploegh, H. L. A guide to antigen processing and presentation. *Nat. Rev. Immunol.* **22**, 751–764 (2022).
37. Blum, J. S., Wearsch, P. A. & Cresswell, P. Pathways of Antigen Processing. *Annu. Rev. Immunol.* **31**, 443–473 (2013).
38. Bhat, M. Y. *et al.* Comprehensive network map of interferon gamma signaling. *J. Cell Commun. Signal.* **12**, 745–751 (2018).
39. Murayama, M. A., Shimizu, J., Miyabe, C., Yudo, K. & Miyabe, Y. Chemokines and chemokine receptors as promising targets in rheumatoid arthritis. *Front. Immunol.* **14**, (2023).
40. Lee, B.-W. & Moon, S.-J. Inflammatory Cytokines in Psoriatic Arthritis: Understanding Pathogenesis and Implications for Treatment. *Int. J. Mol. Sci.* **24**, 11662 (2023).
41. Kany, S., Vollrath, J. T. & Relja, B. Cytokines in Inflammatory Disease. *Int. J. Mol. Sci.* **20**, 6008 (2019).
42. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).
43. Sun, S.-C., Chang, J.-H. & Jin, J. Regulation of nuclear factor-κB in autoimmunity. *Trends Immunol.* **34**, 282–289 (2013).
44. Ishina, I. A. *et al.* MHC Class II Presentation in Autoimmunity. *Cells* **12**, 314 (2023).
45. Thomas, T. *et al.* A longitudinal single-cell therapeutic atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. 2023.05.05.539635 Preprint at https://doi.org/10.1101/2023.05.05.539635 (2023).
46. Feng, X. *et al.* Interferon-β corrects massive gene dysregulation in multiple sclerosis: Short-term and long-term effects on immune regulation and neuroprotection. *eBioMedicine* **49**, 269–283 (2019).
47. Reyes, M. *et al.* An immune-cell signature of bacterial sepsis. *Nat. Med.* **26**, 333–340 (2020).
48. Raphael, I., Joern, R. R. & Forsthuber, T. G. Memory CD4+ T Cells in Immunity and Autoimmune Diseases. *Cells* **9**, 531 (2020).
49. Huang, Y. *et al.* T peripheral helper cells in autoimmune diseases: What do we know? *Front. Immunol.* **14**, (2023).
50. Rönnblom, L. & Leonard, D. Interferon pathway in SLE: one key to unlocking the mystery of the disease. *Lupus Sci. Med.* **6**, e000270 (2019).
51. Aue, A. *et al.* Elevated STAT1 expression but not phosphorylation in lupus B cells correlates with disease activity and increased plasmablast susceptibility. *Rheumatology* **59**, 3435–3442 (2020).
52. Andreou, N.-P., Legaki, E. & Gazouli, M. Inflammatory bowel disease pathobiology: the role of the interferon signature. *Ann. Gastroenterol.* **33**, 125–133 (2020).
53. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).
54. Garrido-Trigo, A. *et al.* Macrophage and neutrophil heterogeneity at single-cell spatial resolution in inflammatory bowel disease. 2022.11.28.518139 Preprint at https://doi.org/10.1101/2022.11.28.518139 (2022).

55. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).
56. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://dl.acm.org/doi/10.1145/2939672.2939785.
57. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Preprint at https://doi.org/10.48550/arXiv.1907.10902 (2019).
58. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
59. Anderson, A. E. *et al.* Expression of STAT3-regulated genes in circulating CD4+ T cells discriminates rheumatoid arthritis independently of clinical parameters in early arthritis. *Rheumatology* **58**, 1250–1258 (2019).
60. Zhang, L., Yu, L., Li, J., Li, Z. & Zhao, X. Novel Compound Heterozygous CYBA Mutations Causing Neonatal-Onset Chronic Granulomatous Disease. *J. Clin. Immunol.* **43**, 1131–1133 (2023).
61. Denson, L. A. *et al.* Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* **154**, 2097–2110 (2018).
62. Garrido-Trigo, A. *et al.* Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat. Commun.* **14**, 4506 (2023).
63. Diamond, M. S. & Farzan, M. The broad-spectrum antiviral functions of IFIT and IFITM proteins. *Nat. Rev. Immunol.* **13**, 46–57 (2013).
64. Wen, L., Krauss-Etschmann, S., Petersen, F. & Yu, X. Autoantibodies in Chronic Obstructive Pulmonary Disease. *Front. Immunol.* **9**, (2018).
65. Hsu, A. T., Gottschalk, T. A., Tsantikos, E. & Hibbs, M. L. The Role of Innate Lymphoid Cells in Chronic Respiratory Diseases. *Front. Immunol.* **12**, (2021).
66. Cui, J. *et al.* TH1-mediated airway hyperresponsiveness independent of neutrophilic inflammation. *J. Allergy Clin. Immunol.* **115**, 309–315 (2005).
67. Ramachandran, P. *et al.* Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512–518 (2019).
68. Lebossé, F. *et al.* CD8+ T cells from patients with cirrhosis display a phenotype that may contribute to cirrhosis-associated immune dysfunction. *eBioMedicine* **49**, 258–268 (2019).
69. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
70. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
71. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
72. De Donno, C. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* 1–10 (2023) doi:10.1038/s41592-023-02035-2.
73. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
74. Kang, J. B. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
75. Edahiro, R. *et al.* Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nat. Genet.* **55**, 753–767 (2023).
76. DeMartino, J. *et al.* Single-cell transcriptomics reveals immune suppression and cell states predictive of patient outcomes in rhabdomyosarcoma. *Nat. Commun.* **14**, 3074 (2023).
77. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
78. Maina, E. K. *et al.* CD4+ T cell counts in initiation of antiretroviral therapy in HIV infected asymptomatic individuals; controversies and inconsistencies. *Immunol. Lett.* **168**, 279–284 (2015).
79. Youssef, A. *et al.* External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* 1–2 (2023) doi:10.1038/s41591-023-02540-z.
80. Sanyal, A. J. *et al.* Diagnostic performance of circulating biomarkers for non-alcoholic steatohepatitis. *Nat. Med.* **29**, 2656–2664 (2023).
81. van Steenbergen, H. W., Cope, A. P. & van der Helm-van Mil, A. H. M. Rheumatoid arthritis

prevention in arthralgia: fantasy or reality? *Nat. Rev. Rheumatol.* 1–11 (2023) doi:10.1038/s41584-023-01035-y.
82. Feld, L., Glick, L. R. & Cifu, A. S. Diagnosis and Management of Crohn Disease. *JAMA* **321**, 1822–1823 (2019).
83. Clark, I. C. *et al.* HIV silencing and cell survival signatures in infected T cell reservoirs. *Nature* **614**, 318–325 (2023).
84. Zuroff, L. *et al.* Immune aging in multiple sclerosis is characterized by abnormal CD4 T cell activation and increased frequencies of cytotoxic CD4 T cells with advancing age. *EBioMedicine* **82**, 104179 (2022).
85. Jarmakiewicz-Czaja, S., Ferenc, K. & Filip, R. Antioxidants as Protection against Reactive Oxidative Stress in Inflammatory Bowel Disease. *Metabolites* **13**, 573 (2023).
86. Seguí, J. *et al.* Superoxide dismutase ameliorates TNBS-induced colitis by reducing oxidative stress, adhesion molecule expression, and leukocyte recruitment into the inflamed intestine. *J. Leukoc. Biol.* **76**, 537–544 (2004).
87. Zhang, C. *et al.* Single-cell RNA sequencing reveals intrahepatic and peripheral immune characteristics related to disease phases in HBV-infected patients. *Gut* **72**, 153–167 (2023).
88. Palshikar, M. G. *et al.* Executable models of immune signaling pathways in HIV-associated atherosclerosis. *Npj Syst. Biol. Appl.* **8**, 1–15 (2022).
89. Li, M. *et al.* DISCO: a database of Deeply Integrated human Single-Cell Omics data. *Nucleic Acids Res.* **50**, D596–D602 (2022).
90. Jiang, Y. *et al.* Single cell RNA sequencing identifies an early monocyte gene signature in acute respiratory distress syndrome. *JCI Insight* **5**, (2020).
91. Cillo, A. R. *et al.* Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **52**, 183-199.e9 (2020).
92. Schafflick, D. *et al.* Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* **11**, 247 (2020).
93. Liu, Y. *et al.* Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nat. Commun.* **12**, 741 (2021).
94. Wang, S. *et al.* An atlas of immune cell exhaustion in HIV-infected individuals revealed by single-cell transcriptomics. *Emerg. Microbes Infect.* **9**, 2333–2347 (2020).
95. Savage, A. K. *et al.* Multimodal analysis for human ex vivo studies shows extensive molecular changes from delays in blood processing. *iScience* **24**, (2021).
96. Mistry, P. *et al.* Transcriptomic, epigenetic, and functional analyses implicate neutrophil diversity in the pathogenesis of systemic lupus erythematosus. *Proc. Natl. Acad. Sci.* **116**, 25222–25228 (2019).
97. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493-1508.e20 (2019).
98. Ahern, D. J. *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916-938.e58 (2022).
99. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895-1913.e19 (2021).
100. Terekhova, M. *et al.* Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB−CD8+ memory T cells and accumulation of type 2 memory T cells. *Immunity* **56**, 2836-2854.e9 (2023).
101. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
102. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
103. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
104. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Annotated data. 2021.12.16.473007 Preprint at https://doi.org/10.1101/2021.12.16.473007 (2021).
105. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
106. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON*.
107. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).

108. Wu, C., MacLeod, I. & Su, A. I. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.* **41**, D561–D565 (2013).
109. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
110. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).
111. La Manno, G. *et al.* Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
112. Massoni-Badosa, R. *et al.* An Atlas of Cells in the Human Tonsil. 2022.06.24.497299 Preprint at https://doi.org/10.1101/2022.06.24.497299 (2022).
113. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
114. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
115. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
116. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
117. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
118. Walle, T. wallet-maker/cytopus: Cytopus v1.30. Zenodo https://doi.org/10.5281/zenodo.8366975 (2023).
119. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. Preprint at https://doi.org/10.12688/f1000research.8987.2 (2016).
120. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
121. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
122. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
123. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).
124. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
125. Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res.* **51**, 10934–10949 (2023).
126. Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B. & Saez-Rodriguez, J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *eLife* **12**, e93161 (2023).
127. Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M. & Onishi, M. Multiobjective Tree-Structured Parzen Estimator. *J. Artif. Intell. Res.* **73**, 1209–1250 (2022).
128. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
129. Chen, H., Janizek, J. D., Lundberg, S. & Lee, S.-I. True to the Model or True to the Data? Preprint at https://doi.org/10.48550/arXiv.2006.16234 (2020).
130. Welford, B. P. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics* **4**, 419–420 (1962).
131. Dong, W., Moses, C. & Li, K. Efficient k-nearest neighbor graph construction for generic similarity measures. in *Proceedings of the 20th international conference on World wide web* 577–586 (Association for Computing Machinery, New York, NY, USA, 2011). doi:10.1145/1963405.1963487.
132. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
133. Godfrey, D. I., Uldrich, A. P., McCluskey, J., Rossjohn, J. & Moody, D. B. The burgeoning family of unconventional T cells. *Nat. Immunol.* **16**, 1114–1123 (2015).
134. Ebbo, M., Crinier, A., Vély, F. & Vivier, E. Innate lymphoid cells: major players in inflammatory diseases. *Nat. Rev. Immunol.* **17**, 665–678 (2017).

Methods

Atlas of Circulating Immune Cells

The Inflammation Landscape of Circulating Immune Cells atlas has been conceived as a comprehensive resource to expand the current knowledge of physiological and pathological inflammation through the study of circulating immune cells. With this aim, we have included data representing both acute and chronic inflammatory processes, as well as healthy donors. Further details about the included datasets are available (**Supplementary Table 1**).

The project includes in-house single-cell RNA-sequencing data generation from samples shared by our collaborators from several research institutions. Samples were collected with written informed consent obtained from all participants and comply with the ethical guidelines for human samples. Specifically, we generated data from patients suffering Rheumatoid Arthritis (RA), Psoriatic Arthritis (PSA), Crohn's Disease (CD), Ulcerative Colitis (UC), Psoriasis (PS), Systemic Lupus Erythematosus (SLE) and healthy controls in collaboration with the Vall d'Hebron Research Institute within the DoCTIS consortia [https://doctis.eu/] (**SCGT00**). Additionally, we processed and obtained data from healthy controls in collaboration with the Institut Hospital del Mar d'Investigacions Mèdiques (**SCGT01**); Asthma, Chronic Obstructive Pulmonary Disease (COPD) and healthy control samples in collaboration with the University Medical Center Groningen (**SCGT02**); Breast Cancer (BRCA) samples in collaboration with the Vall d'Hebron Institute of Oncology (**SCGT03**); cirrhosis samples in collaboration with the Biomedical Research Institut Sant Pau (**SCGT04**); samples of patients suffering Colorectal Cancer (CRC) in collaboration with the Katholieke Universiteit Leuven (**SCGT05**) and, finally, COVID and healthy control samples also in collaboration with Biomedical Research Institut Sant Pau (**SCGT06**).

Moreover, we also included publicly available datasets to complete our cohort. Specifically, we considered data from patients suffering sepsis from **Reyes *et al.***[47] and **Jiang *et al.***[90], Head and Neck Squamous Cell Carcinoma (HNSCC) from **Cillo *et al.***[90], Hepatitis B Virus (HBV) from **Zhang *et al.***[87], Multiple Sclerosis (MS) from **Schafflick *et al.***[91], NasoPharyngeal Cancer (NPC) from **Liu *et al.***[93], Human Immunodeficiency Virus (HIV) from **Palshikar *et al.***[87] and **Wang *et al.***[93], SLE from **Perez *et al.***[27], **Savage *et al.***[94] and **Mistry *et al.***[95], cirrhosis from **Ramachandran *et al.***[67], CD from **Martin *et al.***[96], COVID-Flu-Sepsis from **COMBAT** (Ahren *et al.*)[97] as well as COVID from **Ren *et al.***[98] and healthy controls from **Terekhova *et al.***[99] and **10X Genomics** together with the available healthy samples from all the cited studies. The data access identifiers for each project can be found in the **Supplementary Table 1 - Sheet 1**. When raw data was available, we downloaded FASTQ files; otherwise, we retrieved the raw count matrices from the NCBI Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo/] or Sequence Read Archieve (SRA) [https://submit.ncbi.nlm.nih.gov/about/sra/], the BioStudies Array Express [https://www.ebi.ac.uk/biostudies/arrayexpress], Broad Institute DUOS [https://duos.broadinstitute.org/], Synapse [https://www.synapse.org], Genome Sequence Analysis (GSA) [https://ngdc.cncb.ac.cn/gsa-human/], CellXGene Data Portal [https://cellxgene.cziscience.com/datasets] and 10X Genomics [https://www.10xgenomics.com/datasets] resources. For all studies, we also collected clinical metadata.

Cell barcodes. The *'cellID'* barcodes assigned were inspired by the TCGA project [https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/]. Each barcode unequivocally identifies a cell, and it is composed by the *studyID* (project), *libraryID* (10X GEM channel), *patientID*, *chemistry* (only when 3' and 5' GEX available for the same sample), *timepoint* (if multiple observations for a patient) and the 10X Genomics cell barcode, respectively [e.g., SCGT00_L046_P006.3P_T0_AAACCCAAGGTGAGAA].

Sample collection

Human blood samples were collected in EDTA tubes (BD Biosciences). Peripheral blood mononuclear cells (PBMCs) from the **SCGT00, SCGT02, SCGT04, SCGT05**, and **SCGT06** datasets were isolated using Ficoll density gradient centrifugation (Lymphoprep™, Stem Cell Technologies; Ficoll-Plus, GE Healthcare Biosciences AB). PBMCs belonging to the **SCGT01** and **SCGT03** datasets were isolated using Vacutainer® CPT tube (BD Biosciences). Subsequently, all aliquots were centrifuged following the manufacturer's protocol. After centrifugation, PBMCs were washed and resuspended in freezing media. Aliquots were gradually frozen using a commercial freezing box (Mr. Frosty, Nalgene, Thermo Fisher Scientific) at -80 °C for 24 h before being transferred to liquid nitrogen for long-term storage.

Cell thawing and preprocessing

Cryopreserved PBMCs were thawed in a water bath at 37ºC and transferred to a 15 ml Falcon tube containing 10 ml of pre-warmed RPMI media supplemented with 10% FBS (Thermo Fisher Scientific). Samples were centrifuged at 350 x g for 8 min at RT, supernatant was removed and pellets resuspended with 1 ml of cold 1X PBS (Thermo Fisher Scientific) supplemented with 0.05% BSA (PN 130-091-376, Miltenyi Biotec). Samples were incubated during 10 min at RT with 0.1 mg/ml of DNAse I (PN LS002007, Worthington-Biochem) in order to eliminate ambient DNA and favor the resuspension of the pellet. Cells were filtered with a 40 µm strainer (PN 43-10040-70, Cell Strainer) to remove eventual clumps and washed by adding 10 ml of cold PBS+0.05% BSA. Samples were centrifuged at 350 x g for 8 min at 4ºC and resuspended in an adequate volume of PBS+0.05% BSA in order to reach the desired concentration. Cells concentration and viability were verified with a TC20™ Automated Cell Counter (Bio Rad) upon staining of the cells with Trypan Blue.

Sample multiplexing by genotyping

PBMCs samples were evenly mixed in pools of 8 donors per library following a multiplexing approach based on donor's genotype, as done by Kang *et al.*[100] for a more cost and time-efficient strategy. Importantly, in the case of **SCGT00** libraries were designed to pool samples together from the same disease with different response to treatment (not relevant in this study) whereas in the case of the **SCGT02 Asthma+HC** cohort, 6 samples belonging to patients were pooled with 2 samples derived from non-smoking healthy control individuals. With this approach, we aimed to avoid technical artifacts that could mask subtle biological differences.

3' Cell Plex

PBMCs samples belonging to the **SCGT01**, **SCGT02 COPD+HC**, **SCGT04** and **SCGT06** cohort were multiplexed with 10X Genomics Cell Plex kit following the Cell Multiplexing Oligo Labeling for Single Cell RNA Sequencing Protocol (10x Genomics). While for **SCGT02 COPD+HC** and **SCGT06** projects we pooled 8 samples from patients with healthy controls together, for **SCGT01** and **SCGT04** we only included samples from the condition of interest. Briefly, 0.2-1 million cells were centrifuged at 350x g at RT with a swinging-bucket rotor, resuspended in 100 µl of Cell Multiplexing Oligo (3' CellPlex Kit Set A PN-1000261, 10x Genomics) and incubated at RT for 5 mins. Cells were washed 3 times with cold 1X PBS (Thermo Fisher Scientific) supplemented with 1% BSA (MACS Miltenyi), all centrifugations being performed at 350x g at 4C. Cells were finally resuspended in an appropriate volume of 1X PBS-1% BSA in order to obtain a final cell concentration of approximately 1600 cells/ul and counted using a TC20™ Automated Cell Counter (Bio-Rad Laboratories, S.A). An equal number of cells of each sample was pooled and filtered with a 40 µm strainer to remove eventual clumps, final cell concentration and viability of the pools were verified before loading onto the Chromium for cell partitioning.

Cell encapsulation and single cell RNA-sequencing library preparation

Multiplexed samples were loaded for a Target Cell Recovery between 20000 and 60000 cells (corresponding to 5000-7500 cells per sample within each plex). More specifically, samples belonging to **SCGT00** and **SCGT01** cohorts were encapsulated using standard throughput Chromium Next GEM Single Cell 3' Reagent Kit v3.1, while multiplexes samples belonging to **SCGT02 Asthma+HC and COPD+HC, SCGT04** and **SCGT06** were encapsulated using the high throughput Chromium Next GEM Single Cell 3' HT Reagent Kit v3.1 in combination with the Chromium X instrument. On the other hand, some samples from the **SCGT00, as well as SCGT03** and **SCGT05** cohort were loaded in a standard assay with a target recovery of 6-8000 cells per sample using the Chromium Next GEM Single Cell 5' Reagent Kit v2 (10X Genomics, PN-1000263).

Libraries were prepared following manufacturer's instructions of protocols CG000315 or CG000390, for the standard assay without and with sample multiplexing, and protocols CG000416 and CG000419 for the high throughput assay without and with sample multiplexing. Protocol CG000331 was instead followed for the **SCGT03** and **SCGT05** cohort. Between 20-200 ng of cDNA were used for preparing libraries and final library size distribution and concentration were determined using a Bioanalyzer High Sensitivity chip (Agilent Technologies). Sequencing was carried out on a NovaSeq6000 system (Illumina) and NextSeq500 using the following sequencing conditions: 28 bp (Read 1) + 10 bp (i7 index) + 10 bp (i5 index) + 90 bp (Read 2), to obtain approximately 40.000 read pairs per cell for the Gene Expression (GEX) library and 2000-4000 read pairs per cell for the CellPlex library.

Data processing

To profile the cellular transcriptome, we processed the sequencing reads with the 10X Genomics Inc. software package CellRanger (v6.1) [https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome] and mapped them against the human GRCh38 reference genome (GENCODE v32/Ensembl 98). This step was applied to the sequencing reads obtained from in-house processed samples and from published projects, when available.

Genotype processing

Genome-wide genotyping data for patients from the **SCGT00** and **SCGT02 Asthma+HC** studies were generated from PBMC samples. For **SCGT00,** 184 patients were distributed in four genotyping batches (N1=64 samples, N2=32 samples, N3=40 samples and N4=48) and an additional set of 32 patients were processed with the Illumina Omni2.5-8 and Illumina GSA MG v3-24 arrays (Illumina, USA), respectively. For **SCGT02 Asthma+HC**, 16 patients were distributed in two genotyping batches (N1=8 and N2=8 samples) using Infinium Global Screening Array-24 v3.0 (GSAMD-24v3.0) with the A1 array. Genotyping was done using GRCh37 human genome reference. Data pre-processing and quality control analysis have been separately performed for each genotyping batch of samples at IMIDomics, Inc. (Barcelona, Spain) and Erasmus MC (Rotterdam, The Netherlands). Quality control (QC) analysis was performed using PLINK software (v1.9 and v2). In the **SCGT00** QC analysis, we have identified autosomal SNPs and, using those SNPs from chromosome X, we have confirmed the consistency between SNP-estimated and clinically-reported genders. Then, we have quantified the percentage of SNPs with a minor allele frequency (MAF) higher than 5%. Next, we have computed the percentage of missingness both at the SNP-wise and sample-wise levels. Finally, we have assessed the heterozygosity rate (F) of each sample in order to evaluate if any of the genotyped samples could be contaminated. In the **SCGT02 Asthma+HC** QC analysis, we excluded samples with SNP calling rate lower than 98%.

Patient genotypes (VCF format) were simplified by removing Single Nucleotide Variants (SNVs) that were unannotated (chr 0), located in the sexual Y (chr 24), pseudo-autosomal XY (chr 25), or mitochondrial chromosomes (chr 26). As genotypes were obtained using the human hg19 reference genome, we converted their coordinates to the same reference genome used to mapped the sequencing reads (GRCh38), via the UCSC LiftOver tool [https://genome.ucsc.edu/cgi-bin/hgLiftOver]. LiftOver requires an input file in BED format. Thus, we used a python script [https://github.com/single-cell-genetics/cellsnp-lite/blob/master/scripts/liftOver/liftOver_vcf.py] to convert our VCF file accordingly.

Library demultiplexing

Multiplexed libraries from **SCGT00** and **SCGT02 Asthma+HC** cohorts were demultiplexed with cellsnp-lite (v1.2.2) in Mode 1a[102], which allows us to genotype single-cell GEX libraries by piling-up the expressed alleles based on a list of given SNPs. To do so, we used a list of 7.4 million common SNPs in the human population (MAF > 5%) published by the 1000 Genome Project consortium, and compiled by the authors [https://sourceforge.net/projects/cellsnp/files/SNPlist/]. Then, we performed the donor deconvolution with vireo (v0.5.6)[102], which assigns the deconvoluted samples to its donor identity using known genotypes, while detecting doublets and unassigned cells. Finally, we discarded detected doublets and unassigned cells before moving on to the downstream processing steps. For CellPlex libraries, we followed a joint deconvolution strategy combining CMO-hashing and genotype-based deconvolution; we generated pools of cells belonging to different samples based on the individual SNPs, and traced back to its donor of origin based on the CMO-hashing. When no genotype is available, the use of this dual approach minimizes the discarded cells.

Data analysis

All analyses presented in this manuscript were carried out using mainly Python, unless specified otherwise. In particular, we structured our data in Anndata objects[103] compatible with Scanpy suite[104], which allowed us to apply single-cell data processing and visualization best practices. The downstream analysis was performed with widely adopted machine learning and single-cell libraries, including scikit-learn[105] and scvi-tools[106]. All experiments and panels are reproducible with the code released in the project's GitHub repository.

Data standardization

Considering the diversity of the datasets included in the reference of circulating immune cells, a standardization step was needed.

Gene name harmonization. All datasets were mapped using human GRCh38 genome reference, but the annotation file version might differ, resulting in gene names with multiple aliases or deprecated symbols. To avoid gene redundancy or mismatching, we used ENSEMBL symbols instead of gene names. Then, for datasets without the ENSEMBL symbols, we compared all gene names with the HUGO Gene Nomenclature Committee (HGNC) database (latest version, February 2024) [https://www.genenames.org/], in order to convert them to the latest official HUGO name, merging possible duplicates, and retrieving the corresponding ENSEMBL symbol. For non-official genes, we used the MyGene python interface [https://mygene.info/][107] to query the ENSEMBL symbol. Finally, we removed 16 genes categorized as "*artifact*" or "*TEC (To be Experimentally Confirmed)"*.

Metadata harmonization. Patient metadata was unified across datasets, using common variable names and values for those present in multiple sources; specifically, we homogenized these variables of interest such as sex, age, disease, diseaseStatus, smokingStatus, ethnicity, or institute. For instance, '*M*', *'Male'*, and '*Hombre'* entries were replaced with '*male'.* Additionally, we created a new variable *'binned_age'* to group patients within a range of 10 years, considering that for the **SCGT01**, **SCGT04**,

and **SCGT11** datasets the specific age information was not available. As detailed below, the datasets missing sex and age information were considered as **External** data and used to evaluate the patient classifier.

Data splitting

Some studies in our cohort included patients with samples collected in multiple replicates, different time points, or using different chemistry protocols. In studies with multiple replicates, i.e., **Zhang2022** and **Terekhova2023**, we select samples with the largest number of cells. When multiple timepoints or disease status for the same patient are available, i.e., **Perez2022**, **COMBAT2022** and **Ren2021**, we kept only the samples associated with higher disease severity. For **SCGT00**, where samples were sequenced with different chemistry protocols, we randomly select them to balance the number of patients across protocols.

The filtered inflammation atlas cohort has been then split in two datasets, namely: **CORE** and **External**. The **External** dataset includes 86 samples and is used as an independent validation of our patient classifier pipeline. For this dataset, we selected studies that either involve diseases with a large support in our full cohort or lack metadata on sex and age. These chosen studies are: **SCGT00val**, **SCGT06**, **Palshikar2022**, **Ramachandran2019**, **Martin2019**, **Savage2021**, **Jiang2020**, **Mistry2019**, and **10XGenomics**.

After performing data quality control (removing low quality libraries and cells, see section below), the **CORE** dataset includes 961 samples, and was further split into **Main** and **Validation**, with 817 and 144 samples each, respectively. We first stratified samples based on the following metadata: *studyID*, *chemistry,* and *disease*. From each of those groups, we randomly selected 20% of samples to be part of the **Validation**, provided that they amounted to at least 5 samples. In the patient classifier pipeline, the **Main** dataset is used as a reference, while **Validation** and **External** are used as query dataset in two independent scenarios.

Further information about the samples classified in each group are detailed in the **Supplementary Table 1 - Sheet 3**.

Data quality control

We performed data Quality Control (QC) on the **CORE** dataset by computing the main metrics (i.e. library size, library complexity, and percentage of mitochondrial, ribosomal, hemoglobin and platelet-related gene expression) on the count matrix. Metric distributions were visualized grouping cells by library (10X Genomics) and by considering their chemistry (3' or 5' prime, and their version). Consequently, we removed low quality observation using permissive thresholds, while the robust cleaning process was performed during cell annotation tasks. In particular, we initially excluded the low quality libraries across datasets (<500 cells or <500 median genes recovered). Next, we removed low quality cells with a very low number of UMIs (<500) and genes (<250), or with a high percentage of mitochondrial expression (>20% for 3' V3 and 25% for 5' and 3' V2), as it is indicative of lysed cells. Then, we removed barcodes with a high library size (>50000 UMIs for 3' V3 and 5' V1, > 40000 UMIs for 5' V2 or >25000 UMIs for 3' V2 chemistry) or with a high complexity (>6000 genes for 3' V3, 5' V1 and 5' V2 or >4000 genes for 3' V2 chemistry). After cell QC, we also removed low quality libraries (<250 cells), low quality samples (<500 cells or <500 median genes recovered) as well as cells from a library if this patient recovered a low total number of cells (<50 cells). In addition, we eliminated genes that were detected in less than 20 cells in less than 5 patients, keeping a total of 22838 genes. Lastly, we computed the cell cycle score using the gene list provided by the function *cc.genes.updated.2019()* from the Seurat library[109] (v4.3.0.1), and defined the different cell cycle 'phases' (G1, G2M, and S). Before the dataset clean-up, we predicted doublets with Scrublet[109] using the function *scanpy.external.pp.scrublet()* from the *scanpy* library (v1.9.8), which provides a score to flag putative

doublets, but without filtering them out at this stage. Consequently, during the clustering and annotation step, the clusters co-expressing gene markers from different lineage/population and high doublet score were assessed to determine whether a specific cluster could be classified as a group of doublets and subsequently excluded. After this step, the **CORE** dataset was split into **Main** and **Validation**, as explained above.

Quality control on the **External** dataset was performed independently. We applied the same approach described above but, we filtered only bad quality libraries and cells.

Data processing for annotation

Annotation strategy. To identify all the immune cell types and states present in the human blood, we employed a recursive top-down approach inspired by previous work done by La Manno et al.[110] and Massoni-Badosa et al.[111]. Starting with 4,918,140 cells and 817 patients from the **Main** dataset, we divided the annotation into several stages. Briefly, we first grouped all cells into the primary compartments within our study. Subsequently, each compartment was processed aiming to detect potential doublets, low quality cells and cells resembling platelets or erythrocytes (cells with high expression of hemoglobin genes). Additionally, we also placed back some clusters of cells into their corresponding cell lineages, when wrongly clustered due to similar profiles (e.g. T cells found into the NK cell group, or vice versa). Then, we identified the clusters resembling specific biological cell profiles (cell subtypes) obtaining a final number of 64 different subpopulations, excluding *Doublets* and *LowQuality_cells*, that we defined as annotation *Level2.* Those cell subtypes were grouped into 15 cell populations that we defined as annotation *Level1*. At the end, our inflammation atlas contains 4,435,922 cells. For each group identified in the initial stage (cell lineages), we applied the following tasks, namely: normalization, feature selection, integration, clustering and annotation. In the following, we will always refer to the parameters of the initial stage, while the specifics of the subsequent steps (from lineages to cell types), along with the annotation labels and the marker genes used to define them, can be found **Supplementary Table 3**.

Data normalization. Following standard practices, filtered cells were normalized by total counts over all genes and multiplied by a scaling factor of 10^4 (*scanpy.pp.normalize_total(target_sum = 10^4)*). Then, the normalized count matrix X was log-transformed as $\log_e(X + 1)$ (*scanpy.pp.log1p()).*

Feature selection. Gene selection was performed by identifying the Highly Variable Genes (HVG). Before doing so, we excluded genes related to mitochondrial and ribosomal organules. Also, we skipped T and B cell receptor (TCR/BCR) genes, including joining and variable regions, since they are not useful to describe cell identities, but rather to capture patient specific clonally expanded cell populations within an inflammatory-related condition. Lastly, we excluded Major Histocompatibility Complex (MHC) genes. In order to reduce the influence of a study's specific composition, and prevent biases in the gene selection task, we preferred genes that are highly variable in as many studies as possible. Therefore, similarly to Sikkema et al.[112], we first considered each study independently and computed the HVGs using the Seurat implementation[113] (*scanpy.pp.highly_variable_genes(min_disp=0.3, min_mean=0.01, max_mean=4)*). Then, we ranked genes based on the number of studies in which they are among the HV. Finally, for the initial stage, we determined the minimum number of studies required to compose a HVGs list of over 3000 genes. Applying this strategy, we selected a total of 3236 genes being highly variable in at least 6 studies. In the following steps, we requested more than 2000 HVGs; the minimum number of studies required and the total of selected genes depends on Step and the cells under study. In order to identify Red Blood Cell (RBC) and Platelets, we kept genes associated with erythrocytes, such as hemoglobin subunits, and PPBP (platelet related) in the HVG list. Since such genes are known to be related with ambient RNA when found in other cell types, we subsequently removed them after having annotated the above cell types.

Data integration. Our dataset includes single-cell data obtained from multiple studies including different chemistry protocols, inflammatory status samples and a broad range of other clinical features (e.g., age and sex). While this is a strength point of our atlas, such high levels of heterogeneity induced by technical confounding factors and unwanted biological variability resulted in challenging integration tasks before clustering and annotating cell populations. Therefore, we employed scVI[114], a Variational AutoEncoder (VAE) approach that proves to be one of the most effective integration methods in complex scenarios, particularly when the annotation information is missing[24]. scVI takes as input the raw count matrix to generate an integrated, low-dimensional, embedding space, where the cell states are preserved and the batch effects are reduced. Moreover, scVI's embedding space can be exploited to cluster and annotate cells based on either known or cluster specific marker genes. Details on the scVI parameters used in each annotation step can be found in Supplementary Table 7.

Cell clustering. In order to cluster cells into cell types with the Leiden algorithm[115], we first built the K-Nearest Neighbors (KNN) graph using scVI's latent embeddings and *k=20* as the number of neighbors (*scanpy.pp.neighbors(n_neighbors=k)*). We then applied the Leiden algorithm using a resolution of *r=0.1* (*scanpy.tl.leiden(resolution=r)*). The *k* and *r* used in every other step for every lineage can be found in **Supplementary Table 3.**

Cell annotation. Cell clusters were manually annotated by immunology experts by comparing the expression levels of canonical gene markers. Moreover, the final step of annotation was performed using the clusters markers obtained performing a Differential Expression Analysis (DEA) among clusters (see **Supplementary Table 3)**. First, we ranked genes to characterize each cluster (*scanpy.tl.rank_genes_groups()*), by considering normalized RNA counts with the Wilcoxon sum rank test. Then, we selected those genes with Log2 Fold Change (log_2FC) > 0.25, False Discovery Rate (FDR) adjusted p-value < 0.05, and if they were present in at least 25% of cells. Importantly, cells belonging to RBC and Platelet populations were excluded from all the downstream analyses, except for label transfer performed as a step during patient classifier tasks (as explained below).

External annotation validation. We compared our independent annotations with the ones available in the largest public datasets (i.e. SLE from **Perez *et al.***[27], COVID-Flu-Sepsis from **COMBAT** (Ahren *et al.*)[97], COVID from **Ren *et al.***[98] and healthy controls from **Terekhova *et al.***[99]). To quantify the overlap of cells among groups, we computed the Adjusted Rand index (ARI)[116] to measure the similarity between our label assignments and the ones performed by the original authors. Further details are available in the **Supplementary Table 2**.

Feature selection post annotation

Gene selection. To improve the quality of downstream analysis to characterize the inflammation landscape it is necessary to perform a gene selection in order to remove dataset specific genes and reduce the batch effect. First, we performed data normalization (as described above) and then removed all the genes that are not expressed (raw count > 0) in at least 1 cell in each study, along with the genes associated to mitochondrial, ribosomal, TCR/BCR, MHC, hemoglobin and platelet cell types. This step retained a total of 14127 genes. Then, we identified three sets of genes: (i) the HVGs, (ii) the Differentially Expressed Genes (DEGs) between healthy and each inflammatory status, and (iii) Cytopus[117], a manually curated immune-specific gene list.

Highly variable genes (HVGs). Similarly to the feature selection approach described in the annotation section, we selected a total of 3283 HVGs, by using a threshold of at least 3000 genes. In practice, we first ranked the genes based on the number of studies in which they are concurrently highly variable (*scanpy.pp.highly_variable_genes(min_disp=0.3, min_mean=0.01, max_mean=4, batch_key='libraryID'))*, and then chose a minimum number of studies of 5.

Differentially Expressed Genes (DEGs) between healthy and each disease. We obtained a list of DEGs after grouping single-cell gene expression profiles into pseudobulks. Therefore, we first combined the expression profiles of individual cells to produce pseudobulks for every patient and cell-type (annotation level 1), removing groups with no more than 20 cells, using the Python implementation of decoupleR[42] (v.1.6.0) (*decoupler.get_pseudobulk(min_cells=20, sample_col='sampleID', groups_col='Level1', layer='counts', mode='sum')*). Then, we applied the edgeR's[118] (v.4.0.16) quasi-likelihood functions to search for DEGs between healthy patients and each other's inflammatory conditions, by considering one cell-type at a time. Since not all the cell-types were detected in each patient, we didn't perform the pairwise comparison if one disease had less than 3 pseudobulks. More in detail, for each pairwise comparison we first removed genes with a low expression value (*filterByExpr(y, group = disease)*). Second, we normalized by library size the aggregated raw counts (*calcNormFactors(y, logratioTrim = 0.3)*). Third, we corrected for the main confounding factors, i.e., chemistry protocol, sex and binned age, considering an additive model. One patient was excluded from the analysis due to missing age information. We defined the design of our comparison using the following patsy-style [https://patsy.readthedocs.io/en/latest/formulas.html] formula: '*~0 + C(disease) + C(chemistry) + C(sex) + C(binned_age)*''. Fourth, we estimated a Negative Binomial (NB) dispersion for each gene using (*estimateDisp()*) which we feed into a gene-wise NB Generalized Linear Model (*glmQLFit(robust=TRUE)*) to test for differentially expressed genes with a quasi-likelihood F-test (*glmQLFTest()*). Lastly, results obtained from each comparison were merged together and the F-test p-values were corrected using the Benjamini-Hochberg FDR procedure implemented in R (*p.adjust(method = 'BH')*). Given the corrected p-values and the $\log_2$FC we selected 6868 DEGs with p-value < 0.01 and absolute $\log_2$FC > 1.5.

Curated immune-specific genes. To be able to track the full spectrum of inflammatory processes, including immune activation and progression, we curated 9 inflammation related functions defined in the literature (1364 genes present in our dataset; **Supplementary Table 4**)[35–41] and complemented them with a published list of cell type-specific signatures derived from immunological knowledge based on single-cell studies (Cytopus[117]). Specifically, we retrieved all global gene sets for the leukocyte category and the following inflammatory-related cell-type specific factors: Naive and Non-Naive CD4 T cells (*CD4T_TFH_UP*, *CD4T_TH1_UP*, *CD4T_TH2_UP*, *CD4T_TH17_UP*, *Tregs_FOXP3_stabilization*); Naive and NonNaive CD8 T cells (*CD8T_exhaustion*, *CD8T_tcr_activation*); B cells (*B_effector*); Monocytes (IFNG *response*, *IL4-IL13 response*); and DC (DC antigen-crosspresentation).

Aggregation of gene sets. We generate the relevant gene set by doing the union of HVGs, DEGs, and the manually curated list. The final number of unique genes is 8253.

Datasets integration and gene expression correction via scANVI

Atlas-level analysis requires a careful preprocessing of the gene expression profiles to deal with the heterogeneity of the studies, the batch effect and the missing or noisy observations[119]. scANVI[23] is one of the existing methods capable of addressing these challenges and has been proven effective on atlas-level benchmarks compared to other integration methods. We validated its performance on our data by using the metrics from the *scib-metrics* package[24] (v.0.5) (see **Extended Data Fig. 1**).

scANVI integration. scANVI is an extension of the scVI model, employed previously for data integration, that also leverages the information of the cell type annotation. We first trained an scVI VAE (*scvi.model.SCVI*), and then trained scANVI (*scvi.model.SCANVI*) starting from the pre-trained scVI model (see parameters in **Supplementary Table 7**). Both models corrected for the *chemistry* batch, while also considering *libraryID*, *studyID*, *sex,* and *binned age* as covariates. After training, we generated the normalized corrected counts by sampling from scANVI's negative binomial posterior (*SCANVI.get_normalized_expression*). The batch effect was mitigated by sampling and averaging each cell's expression as if it originated from each chemistry protocol by setting the *transform_batch*

parameter to the list of chemistry protocols present in our atlas. To parallelize the gene expression sampling, we implemented a Nextflow (v.24.04) pipeline[120] that processed each sample independently and aggregated the obtained anndata objects afterwards.

Comparison of cell-type composition

Compositional cell-type analysis. To estimate the changes in the proportions of cell populations across conditions, we applied the scCODA Python package[121] (v0.1.9), a Bayesian modeling tool that takes into account the compositional nature of the data to reduce the risk of false discoveries. scCODA allows us to infer changes between conditions while considering other covariates, corresponding to the disease status in our setting. scCODA searches for changes between a reference cell type, assumed to be constant among different conditions, and the other cell types. We selected as the reference population the one that showed the lower variance across conditions, excluding rare cell populations (i.e., *Progenitors, pDCs* and *Cycling cells*). This resulted in the selection of DC as the reference cell type for all diseases. scCODA takes as input the count of cells belonging to each cell type in each patient and returns the list of cell types proportion changes with the corresponding corrected p-values (through the False Discovery Rate procedure; FDR). A *patsy-style* formula was used to build the covariate matrix, specified with 'healthy' as baseline and sex and binned age as covariates (*C(disease, Treatment('healthy')) + C(sex) + C(binned_age)*), since we are interested in detecting changes between a normal and diseased status. We reported only changes with a corrected p-value < 0.05 and a $\log_2$FC > 0.2.

Comparison of gene expression profiles

Gene factor inference. In order to expand the list of curated immune-related genes following a data-driven approach, we employed Spectra[25] (v.0.2.0), a matrix factorization algorithm that enables us to identify a minimal set of genes related to specific functions in the data, i.e., factors. Spectra takes as input cell-type labels to infer global and cell-type specific factors that decompose the overall gene expression matrix and each cell-type submatrices, respectively. Given our list of curated gene sets, we considered the Cytopus ones as global factors, while we regarded all the remaining as cell-types specific factors. The Spectra model was fitted with default parameters with the exception of *lambda*, which was set equal to 0.001. Considering the prohibitive computational resources required for applying Spectra on our single-cell data, we fed the algorithm with the metacell aggregated expression matrix, as described in the paragraph below. Spectra returned a list of 135 factors that are a linear combination of the gene expression from the original matrix. The coefficients included in the matrix can be then used as a proxy of the gene relevance in a given factor.

Metacell generation. We generated metacells using SEACells[123] (v.0.3.3), which aggregates cells by exploiting their distances in a low-dimensional embedding space. Starting from the normalizing data, we selected the top 3000 highly variable genes using the *highly_variable_genes* function in scanpy, with the *seurat* flavor. To define SEAcell's input embedding space, we calculated the first 50 Principal Components (PCs) and selected those PCs that explain 90% of the total variance observed across all 50 PCs. To avoid biases due to batch effect and other confounding factors, we executed SEAcell for each sample independently. In particular, we generated a number of metacells equal to the number of cells of each patient divided by 50. SEAcell was executed in parallel with Nextflow[120]. We further filtered the obtained metacells by computing the proportion of the most abundant annotation label (*Level1*) in each SEAcell group, and then removed the ones with a purity lower than 0.75. Overall, we defined 71108 metacells. Given the assignment of cells to each metacell, we generated each metacell's gene expression profile by averaging the corresponding cells' scANVI normalized and corrected expression profiles. Since scANVI returns counts sampled from a negative-binomial distribution, we also log-scaled the obtained metacell profiles.

Inflammation-related signature definition. Spectra provided a total of 135 factors, that include a refined gene list for each gene set we used as input. Thus, we need to assign those factors to our original gene sets for retrieving the corresponding biological function. For doing so, we performed enrichment analysis with Univariate Linear Models (UML) available in the Python implementation of decoupleR[42], to estimate the factors associated with each gene set. We considered the gene coefficients returned by Spectra as response variable and as explanatory variables a vector of weights equal to 1 or 0 if the gene were included in the gene set or not respectively. UML returns an *estimate* and a *p-value* for each enrichment. We correct those p-values for multiple comparison by computing the False Discovery Control with Benjamini-Hochberg procedure, implemented in *scipy* (v.1.12.0) Python library[123]. We kept 125 factors with a positive estimate and an adjusted p-value lower than 0.05. Finally, we assign to each factor the biological function that corresponds to the gene set that provided the highest estimated score.

Inflammation-related signature scores. To compare immune-relevant activation profiles across diseases and cell types, we applied an enrichment signature scoring procedure, considering the factors obtained with Spectra[25]. In particular, we applied an Univariate Linear Model (ULM) from decoupleR[42]. First, we generated pseudobulks stratified by cell type (annotations *Level1* or *Level2*) and patients, discarding groups with less than 10 cells. We averaged the scANVI corrected gene expression matrix of each cell belonging to a given pseudobulk, and then log-transformed and scaled the expression values to zero mean and unit variance, to reduce the impact of highly expressed genes. We fitted decoupleR's ULM by considering pseudobulk expression profiles as the response variable and the gene coefficient returned by Spectra as the explanatory one. We assessed the scores for the 119 cell-type-specific factors only in their corresponding cell-type. The output of the model is a t-student statistic for each combination of pseudobulk and factor, which is used as a proxy for the corresponding biological function activity: positive values are associated with more active functions in a given sample, and vice-versa. To identify the up- or down- regulated biological functions across inflammatory conditions, we compared the activation score between *healthy* and each disease, considering only comparisons that include at least 3 observations in both conditions. To take into account the batch effect induced by studies and chemistry protocols, that still affects the data (see **Extended Data Fig. 1b,c** *scbi* metrics and PCA), we applied a Linear Mixed Effect Model (LMEM). In particular, we fitted the function *mixedlm()* from *statsmodels* Python Library (v.0.14.1) with the formula*: "'Q("{factor}") ~ C(disease, Treatment(reference="healthy")) + 'f'C(chemistry)',* grouping by *"studyID".* We corrected the p-value obtained for multiple testing using FDR considering all the comparisons when tested at *Level1*, and within each *Level1* population when tested at *Level2*.

Gene Regulatory Network (GRN) Analysis. Pseudobulk matrices were calculated by averaging the corrected and standardized count matrices by cell type and sample. We compute differential expression analysis for each cell type in each disease using healthy individuals as reference. Linear Mixed Effect Model (LMEM) were used to model the expression levels of each gene independently, considering the disease as a fixed effect while modeling the ID of the study as a random effect. We used the mixedlm() function of statsmodels (v.0.14.0) Python package to run the analysis. To associate each cell-type specific "*IFN-induced"* factor with a given Transcription Factor (TFs) regulator, we integrated these signatures with the CollecTRI Gene Regulatory Network[125] by matching target genes to identify common genes between TF regulons and Spectra signatures. Therefore, each "*IFN-induced"* signature was thus linked to a subset of TF regulons. The activity of each TF was calculated using only the common genes between each TF and each signature, employing the UML from decoupleR[42] and the z-values obtained from the differential expression analysis. To ensure robustness, only regulons with at least 10 gene targets were considered. This pipeline was applied across "*IFN-induced"* factors and diseases, focusing on the activity in the cell type where the Spectra signature was identified. Negative activities (t-stat < 0) and non-significant results (p-value > 0.05) were filtered out. This analysis identified STAT1 and SP1 as the sole transcription factor regulators of the defined cell types.

To explore disease distribution according to "*IFN_induced*" Spectra signatures, the MOFA[126] (version 0.3.2) framework was utilized. We used diseases as samples, cell types as views (representing the Spectra signatures), unique gene targets from SP1 and STAT1 as features and z-values from DEG of these targets as values. The convergence mode was set to fast, and the dropR2 argument was set to 0.0001. For the comparison of Flare and Non Flare patients from Systemic Lupus Erythematosus (SLE), non-corrected log1p normalized single-cell expression matrix from *Perez et al.*[27] was used to further investigate SP1 and STAT1 regulon activities across both categories. Pseudobulk profiles were calculated by averaging by cell type, considering only cell types (annotation *Level 2)* with a minimum of 10 cells and groups that include at least 3 patients in both conditions (Flare vs Non Flare). Prior to calculating TFs activities across samples, we standardized the gene expression data on SLE patients based on healthy individuals. Specifically, for each gene, the mean and standard deviation were calculated from the healthy group, and these statistics were then used to scale the gene expression values across SLE patients. Only gene targets identified in the previous step were used to calculate enrichment using the UML method. Finally, the activity of STAT1 and SP1 was calculated at *Level2* using CollecTRI[125].

Immune gene importance evaluation

In this section, we will introduce our pipeline used to obtain a gene importance metric by interpreting cell-type-specific classifiers for disease prediction. All the steps described below were carried out separately for each cell type (excluding RBC, Platelets, Progenitors and Cycling cells). Specifically, the classification task was performed with Gradient Boosted Decision Trees (GBDTs) implemented in the XGBoost library[56] (py-xgboost-gpu: v2.0.3). Furthermore, interpretability was performed using SHapley Additive exPlanation (SHAP) values[58] (v.0.45.1), a powerful approach assigning an importance to each gene by also taking into account their interactions.

Feature Selection. To focus our analysis on cell type specific inflammatory related signatures we considered only genes relevant in annotated Spectra factors and we further reduce the list by removing cell identity genes (ex. *CD3E*, *MS4A1*) as well as non protein coding genes. This filtering gave a final number 935 genes.

Data Processing. We split our data in three parts: the training set, and the validation and testing set, used for hyperparameter tuning and performance evaluation, respectively. We balanced the splits by disease, ensuring that each sample's cells were included in the same set. Initially, we partitioned the data into five splits using the function *sklearn.model_selection.StratifiedGroupKFold*. Three of these splits were assigned to the training set, while one was designated for validation and one for testing. Accounting for both stratification by disease and patient partitioning might lead to an uneven distribution of cells among diseases. To address this, we assigned splits with a well-balanced distribution of cells to the training and testing sets first.

XGBoost Fitting. XGBoost (*xgboost.XGBClassifier)* hyperparameters were tuned using the Optuna library[57] (v.3.6.0), which employs a Tree-structured Parzen Estimator (TPE) sampling algorithm[127] to navigate the hyperparameters space. The performance of each model configuration was estimated using the weighted F1 score on the validation set. To reduce the computational cost, we both pruned unpromising hyperparameters and early-stopped the training when no improvement is achieved over 20 steps before the upper bound of 1500. We considered 50 configurations of XGBoost, taking into account the hyperparameters detailed in the **Supplementary Table 7.** Using the best configuration and its corresponding number of training steps (equivalent to the number of estimators), we retrained the best model on the union of the training and validation sets. This time, we did not apply early stopping and increased the number of training steps by 20%, to account for the larger number of training samples.

SHAP interpretability. To interpret the decision of the selected XGBoost classifier, we employed the widely-adopted Shapley values[128] through the SHAP library. SHAP values were computed with *shap.TreeExplainer* using the *observational* "*three_path_dependant*" approach, which is expected to be more representative of the data itself and its causal dependencies than the *interventional* one[129]. Given the potential resource-intensive nature of handling all SHAP values for every cell and disease, especially in terms of storage, we computed their mean and variance across all samples in batches using the Weldford online algorithm[130]. Given a specific cell type *ct*, we have a SHAP value for every gene in every cell, and for each disease: a matrix of real values $shap^{ct}(c,g,d)$, where *c, g,* and *d* identify the cell, gene, and disease, respectively. The average contribution of a gene *g* for a disease *d* can be computed as $Dshap^{ct}(g,d) = mean_{c \in C} \vee shap^{ct}(c,g,d) \vee$, where *C* is the set of cells.

Gene Selection Validation

To validate our ensemble of important genes through SHAP values, we tested if our selection generalizes to unseen studies. First, we defined a gene set ***GS*** that included genes expressed in at least 5% of the cells. Then, for each of the 8 conditions included in **EXTERNAL**, we selected from ***GS*** the top *k* ranked genes by SHAP importance. We then trained XGBoost in a nested cross-validation setting on the **EXTERNAL** dataset, where we performed both hyperparameter tuning and performance evaluation, using only the selected genes as input features. Next, we computed Weighted F1 (WF1) and Balanced Accuracy Score (BAS) to test the performance considering *k*=5 and *k=10*. Given our selection ***S*** of genes, with size |***S***|, we also tested 20 sets of |***S***| randomly selected genes from ***GS***, excluding the ones in ***S*** (i.e., not top-ranked according to SHAP). Lastly, we compared the performances of the models trained on those subsets of genes against XGBboost trained on the whole set of gene ***GS***. The analysis was repeated for each cell-type independently.

Patient classifier pipeline

In this section, we describe the pipeline used to validate the inflammation atlas as a diagnostic tool. In the following analysis the terms 'patient' and 'sample' are equivalent, since after Data splitting we kept only one sample foreach patient. The pipeline consists in i) integrating an annotated reference dataset with data integration tools that provide batch-corrected embeddings, ii) mapping a query dataset into the reference to obtain its corrected embeddings, iii) transfer the cell annotation labels from the reference, iv) defining a patient embedding space, and v) training a classifier to predict the patient conditions from the embeddings.

Starting from a large annotated reference dataset, we applied four state-of-the-art integration methods, described below, to obtain a batch corrected embedding. We considered different chemistry protocols as the main source of batch effect, thus we corrected for the *chemistry* covariate. Then, an independent query dataset was mapped into the corrected embeddings. This step provides both batch correction and allows us to transfer cell annotation labels from the reference to the query dataset. To define patient-wise embeddings, we averaged each patient's cell embeddings by cell type, resulting in an embedding for each cell-type and each patient. To predict the inflammatory conditions of the patients in the independent query dataset, we fit one classifier for each cell type on the reference patient embeddings. Then, we predicted the inflammatory condition of the query patients by returning the most frequent condition among the predictions of every cell-type-specific classifier.

We validated our pipeline considering three different settings. In the first one, we performed a cross-validation on the **Main** dataset, where each left-out split is considered as a query dataset and the remaining as the reference. Moreover, we tested our diagnostic tool on the **Validation** and **External** datasets, this time using the whole **Main** as a reference.

Integration Methods

In this section we will explain each data integration method, while the tested configurations of hyperparameters can be found in **Extended Data Fig. 7 and Supplementary Table 7**.

scGEN. scGen is defined by two main components: a Variational AutoEncoder (VAE) and a latent space arithmetic method. The VAE estimates a posterior distribution of latent variables through the encoder, from which we can reconstruct the expression matrices via the decoder (*scGen_model.batch_removal()*). Similarly to commonly employed VAEs, scGen approximates the posterior through a variational distribution, modeled by the encoder and defined as a multivariate Gaussian. When the scGen's VAE has been fitted on the reference dataset, latent space arithmetic is employed to correct for the batch effect induced by the chemistry protocol used. Within each cell type, scGen first selects the mean μ_{max} of the most populated batch, and then corrects each batch with mean μ_0 by adding $\delta = \mu_{max} - \mu_0$ to each cell's embedding. Importantly, the cell type has to be inferred when not known. The final corrected count matrix will correspond to the generated count matrix from the arithmetic-corrected embeddings. Following scGen's tutorials, we will refer as corrected embeddings to the ones obtained given the corrected expression matrix as input. In order to apply scGen batch correction on the query dataset, we need to also infer the cell types of those cells. This step was performed through label transfer by nearest neighbors, following a similar approach employed in Human Lung Cell Atlas[113] and introduced in[70]. The idea is to employ (approximate) nearest neighbors through the PyNNDescent[131] (v0.5.11) (*pynndescent.NNDescent().prepare()*), and infer the most probable cell type in the 10 nearest neighbors (*pynndescent.NNDescent().query())* from the already annotated cells in the reference dataset. To account for the shape of the distribution of the neighbors, a Gaussian kernel was applied instead of using the Euclidean distance. The most probable nearest neighbor cell type is then assigned to annotate new cells.

scANVI. We first trained scVI and scANVI on the reference dataset, like the Dataset Integration described before, then we fine-tuned it to the query dataset. Regarding the label transfer, we employed the scANVI *predict()* function with default parameters.

Harmony and Symphony. Harmony[132] and Symphony[74] are two related methods that integrate a reference and map a query dataset to it, respectively. Harmony takes a PCA embedding of cells as input, along with their batch covariates (*chemistry*). Next, the model represents cell states as soft clusters, where each cell identity is defined as a probabilistic assignment across clusters, with the aim of maximizing diversity among batches within those clusters. Cells are iteratively assigned soft-cluster memberships; those assignments are used as weights in a linear mixture model to remove confounding factors. The result is a new batch-corrected embedded space. The Symphony algorithm starts from the linear model parameters inferred by Harmony to map query cells onto the corresponding embedding space. First, it projects the query gene expression profiles into the same uncorrected low-dimensional space as the reference cells. Next, Symphony computes soft-cluster assignments for the query cells based on their proximity to the reference cluster centroids. Finally, Symphony employs the Harmony mixture model components to estimate and regress out batch effects from the query data. Importantly, the reference cell embedding remains stable during this mapping process. We transferred annotation labels from the reference to the query dataset by exploiting cell proximity in the embedding space using nearest neighbors through s*klearn.classifier.KNeighborsClassifier,* Symphony default choice [https://github.com/potulabe/symphonypy].

scPoli. In contrast to other integration methods such as scANVI, scPoli[72] encodes the condition (*chemistry*, *sampleID*) as a learnable conditional embedding and characterizes each cell-type as a prototype in the latent embedding to facilitate the label transfer. In the reference building phase, we first pre-trained the model given the reference dataset and its conditions, and then fine-tuned to optimize the prototypes. In the reference mapping phase, we freezed the model and learned the new conditional

embeddings belonging to the query dataset. The label transfer is performed by simply assigning the cell type belonging to the closest prototype in the latent embedding space. All the methods belong to the scArches[70] class *scarches.models.scPoli*.

Note that scGen and Harmony/Symphony approaches generate one integrated dataset that is independent from the query data, while scANVI and scPoli require a fine-tuning of the reference model for a given query dataset.

Disease Classifiers

Patient Embeddings Definition. After obtaining the corrected embedding from one of the data integration approaches described previously, we need to aggregate the cell-wise embeddings into patient-wise embeddings. We decided to group at the level of the cell-types, by computing the mean embedding across cells belonging to the same cell type and *sampleID.* Only for scPoli, we generated three different types of patient embeddings: the learned patient embeddings (*sample*), the averaged cell-wise latent embeddings (*cell*), and the concatenation of the two (*cell&sample*).

Classifiers Definition and Hyperparameter Tuning. In this phase the aim is to train a classifier for each cell-type on the patient-wise embeddings belonging to the reference dataset. We tested the following classifier types: *sklearn.svm.LinearSVC*, *sklearn.svm.SVC*, and *sklearn.neighbors.KNeighborsClassifier (sklearn v.1.4.1.post1)*. For each classifier type, we trained different configurations defined in **Supplementary Table 7**, and evaluated their performance using a 5-fold cross-validation on the reference patient embeddings. Similarly to what we did to optimize the XGBoost classifier when estimating the immune gene importance, we employed the Optuna library to perform the hyperparameter tuning for each classifier. The best hyperparameters combination was selected according to the weighted F1 score independently of the cell-type.

Majority Voting and Evaluation. The best classifier type according to the average performance over all cell-type is then used to train from scratch the corresponding classifier on the whole reference patient embedding. The predicted condition for a patient is simply the majority voting among the classifiers. In case of a tie of different conditions, we conservatively rejected the prediction of the classifiers. Then, the overall metrics weighted F1-score, BAS, and Matthew Correlation Coefficient (MCC), and the disease-wise metrics Precision, Recall, BAS and F1-score were computed by comparing the predicted inflammatory conditions by each classifier type in the query dataset with the available ground truth. All those metrics were computed with the Sklearn python library, when we refer to the weighted version of a given metric we are using average='weighted' parameter to take into account the unbalance of the inflammatory condition observations.

Note, if a given query patient does not have any cells annotated for a given cell type, the corresponding prediction was set as 'Not Available'. This label was not taken into account during the Majority Voting procedure, and considered as a wrong prediction when evaluating the performances of that cell-type.

Extended Data

Extended Data Figure Legends (1-11)

Extended Data Figure 1. Composition of the inflammation atlas datasets from the Main Core. (**a**) Barplot showing patient count distribution across different diseases, including healthy condition, stratified by technical variables (*studyID* and *chemistry*) and clinical metadata (*age* and *sex*). The donor without age information is shown in white. (**b**) *SCIB* metric results computed on five different embedding spaces, ranked by their overall performances. (**c**) Heatmaps showing the coefficient of determination R^2 from a linear regression between each principal component and one of four confounding factors. The Principal Component Analysis was performed on (**left**) original data (normalized and log-scaled) and (**right**) from scANVI normalized expression (log-scaled). (**d**) Cellular proportions (*Level 1*) across diseases and healthy donors (**Top**). Compositional analysis of *Level 1* populations (excluding Platelets and RBC) between each disease and healthy donors (**Bottom**). The dot size reflects the significance of the result ('Final parameter' != 0), and the color represents the log2FC (Disease vs Healthy).

Extended Data Figure 2. Marker gene expression for the Inflammation Atlas cell populations. Dotplot showing the average expression of canonical gene markers (*x-axis*) characterizing each cell population (*y-axis*), both considering (**a**) Level1 and (**b**) Level2 annotation (UTC, Unconventional T cells[133]; ILC, Innate Lymphoid Cells[134]). The dot size reflects the percentage of cells in a cluster expressing each gene, and the color represents the average expression level.

Extended Data Figure 3. **Inflammation-related signatures across diseases and cell types**. (**a**) Agglomerative hierarchical clustering with complete linkage, performed considering the euclidean distance among columns, based on the corrected immune-related signature activity score computed by disease and cell type (*Level 1*). (**b**) Barplot showing factor scores for each disease obtained from MOFAcell analysis. (**c**) Boxplot displaying the activity of STA1 and SP1 across cell types (annotation *Level2*) in SLE patients from *Perez et al.*[27]. Asterisk *(**)* identifies significant differences with respect to other lineage populations (Wilcoxon rank-sum test corrected for multiple testing; p-value Adj < 0.05); location refers to the direction of change (on top, means upregulated whereas below, refers to downregulated). (**d**) Dotplot showing the uncorrected average expression of the 50 genes defining the IFN Type 1 and 2 signature (*x-axis*) in Non-Naive CD8 T cell population (*Level 1*) across IMID diseases and health (SCGT00 study, *y-axis*). (**e**) Dotplot showing the uncorrected average expression of the 10 top ranked genes defining the IFN Type 1 and 2 signature (*x-axis*) across subpopulation of Non-Naive CD8 T cells (*Level 2*) in IMID diseases and health (SCGT00 study). For panels (**d**) and (**e**), the dot size reflects the percentage of cells of each disease expressing each gene, and the color represents the average expression level.

Extended Data Figure 4. Confusion matrices of predicted inflammatory condition by cell type. Normalized confusion matrices, one for each cell-type, displaying proportion of predictions belonging to each True Condition. Diagonal values correspond to the Recall metric. XGBoost was trained on the (a) scANVI batch corrected and log-scaled cell expression profiles, and **(b)** original normalized and log-scaled cell expression profiles.

Extended Data Figure 5. Functional biomarker discovery using interpretable machine learning analysis. (**a**) Gene list ranked top-to-bottom by importance (absolute SHAP value), coupled with max-normalized expression levels computed per celltype (*Level1*) and considering selected diseases. From **left** to **right**, reporting top ranked genes for Naive CD4 T cells in RA disease as well as for Monocytes and pDC in SLE patients. (**b**) Rank by importance (absolute SHAP value) of the CYBA gene in every combination of celltype (*Level1*) and disease. (**c**) Scatter plot of max-normalized gene expression against SHAP values computed for *CYBA* gene on Monocyte population and specific diseases (COPD and Asthma, from top to bottom). (**d**) *CYBA gene* expression levels in a validation cohort of intestinal biopsies[62] from CD (n=6), UC (n=6) and healthy (n=6) patients across the different monocyte-derived and macrophage populations. Asterisks (*) indicates statistical significant changes using the Wilcoxon signed-rank test and corrected for multiple-testing. (**e**) Rank by importance (absolute SHAP value) of *IFITM1* gene in every combination of celltype (*Level1*) and disease. (**f**) Scatter plot of max-normalized gene expression against SHAP values computed for *IFITM1* gene on Non-Naive CD4 T cells (**top**) and

ILC (**bottom**) populations and disease, COPD (**left**) and Asthma (**right**). In Panels **(a)**, **(b)**, and **(e)** we first dropped the genes expressed in less than 5% of the selected cell population. In Panels (**c**) and (**f**), we limited the visualization to up to 60,000 cells, sampling an equal percentage from each patient corresponding to 5%, 7.5% and 100% of the total cells for Monocytes, Non-Naive CD4 T cells, and pDC, respectively. Cells belonging to samples with or without the given condition are marked in orange or blue, respectively.

Extended Data Figure 6. Additional performance evaluation metrics. Scenario 1. (**a**) Boxplots showing the distribution of Balanced Accuracy Score (balanced by true disease support), F1, Precision and Recall computed during 5-fold cross validation, considering Majority Vote prediction on the left-out split for each inflammatory condition. The average number of samples among 5 splits, with the corresponding ground truth labels, are also reported. (**b**) Boxplots showing the distribution of Balance Accuracy Score (**top**), and Matthew Correlation Coefficient (**bottom**) computed during 5-fold cross validation, considering Majority Vote and cell-type prediction, on the left-out split. (**c**) Heatmap reporting Precision and Recall computed by aggregating the prediction performed by each cell-type on each left-out split during 5-fold cross validation.

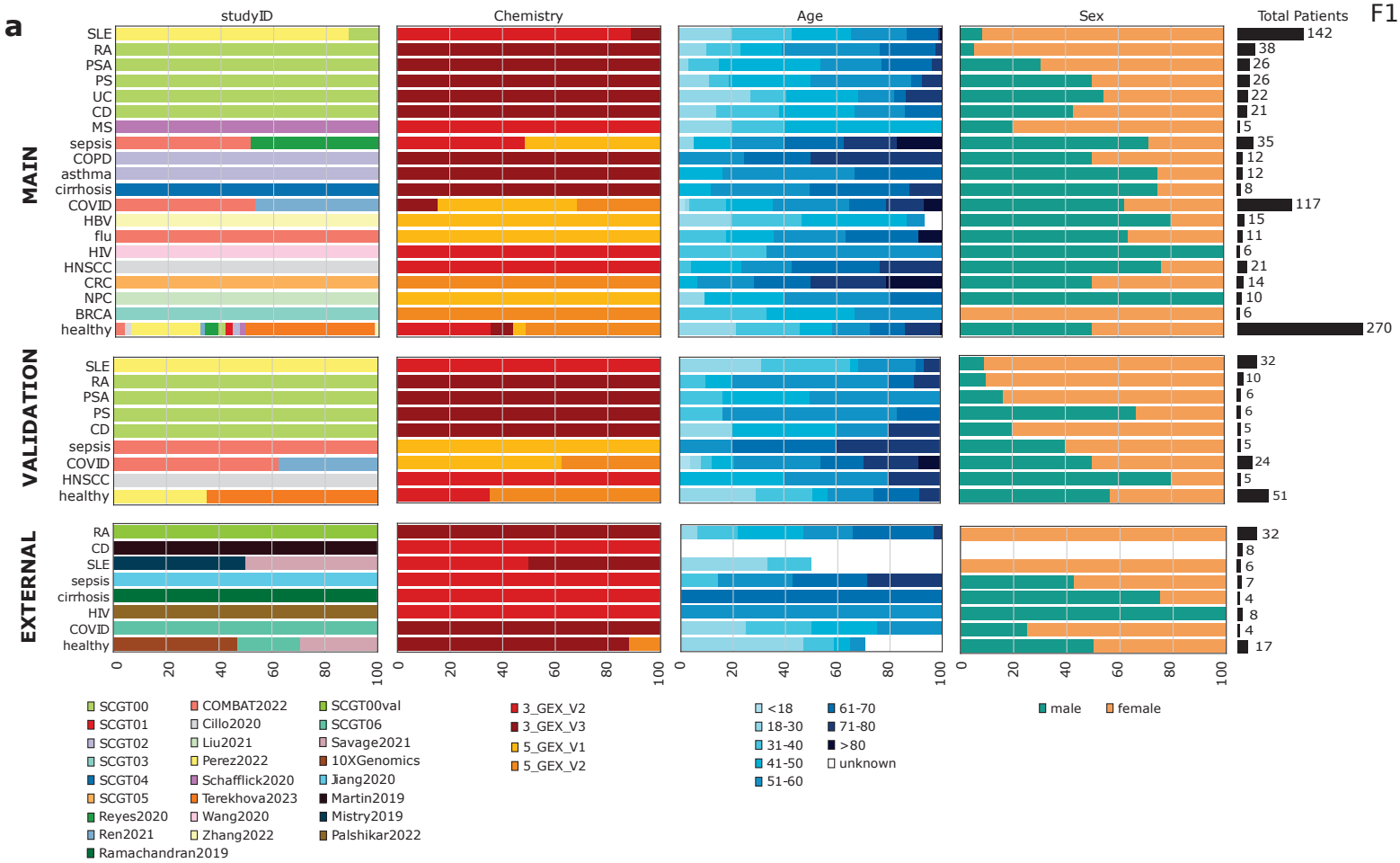
Extended Data Figure 7. Additional performance evaluation metric. Scenario 2 and 3. (**a**) Table with all the data integration method configurations and corresponding performance metrics computed on the Validation dataset (VAL, scenario 2) and External dataset (EXT, scenario 3). The best performing classifiers (SVM with different kernels and kNN) in terms of Weighted F1 are reported. (**b-c**) Performance evaluation from Scenario 2 and 3, respectively, showing (left) the distribution of Weighted Recall and Weighted Precision for all the configurations of each data integration approach, and (right) the mean and standard-deviation of each data integration method. Arrows highlight the scANVI configuration applied in scenario 1.

Extended Data Figure 8. Extended patient classifier workflow schema. (**a**) Definition of reference and query datasets, for Scenarios 1, 2, and 3. (**b**) Integration of the reference dataset and mapping of the query dataset to define the patient-wise embeddings, stratified by cell-type. (**c**) Patient classifier pipeline composed by the hyperparameter tuning of each classifier family, the selection of the best classifier family and the final evaluation of the left-out query dataset.

Extended Data Figure 9. Sample-wise distribution of SHAP-values for *CYBA* and *IFITM1*. Averaged sample-wise SHAP values for each inflammatory condition, i.e., the mean SHAP value of the corresponding cells. The orange and blue groups correspond to the samples with or without the corresponding condition, respectively. In detail, we considered the following combinations of cell-type and gene: (Mono, *CYBA*), (CD8 Non-Naive, *IFITM1*), (CD4 Non-Naive, *IFITM1*), and (ILC, *IFITM1*).

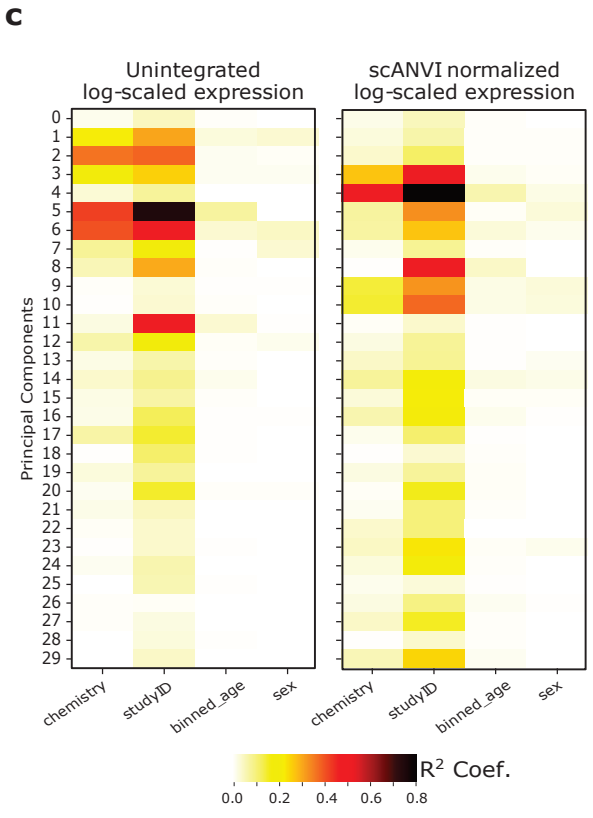
Extended Data Figure 10. Validation of SHAP-based gene selection. Performance evaluation of XGBoost trained with a nested cross-validation on the EXTERNAL dataset. Each point corresponds to the average left-out fold performance, for each best configuration of each fold combination. The scatter plot reports the Weighted F1-score (WF1) and the Balanced Accuracy Score (BAS) computed considering top 5 and 10 genes (among the ones expressed in at least 5% of the total cells), for each inflammatory condition present in EXTERNAL (i.e., healthy, sepsis, CD, SLE, HIV, cirrhosis, RA, COVID), according to the SHAP values. Moreover, for the same number of genes, we report the performance of 20 random selections. Last, we computed the performance of the classifier when trained on the whole gene set, consisting of the genes expressed in at least 5% of the total cells.

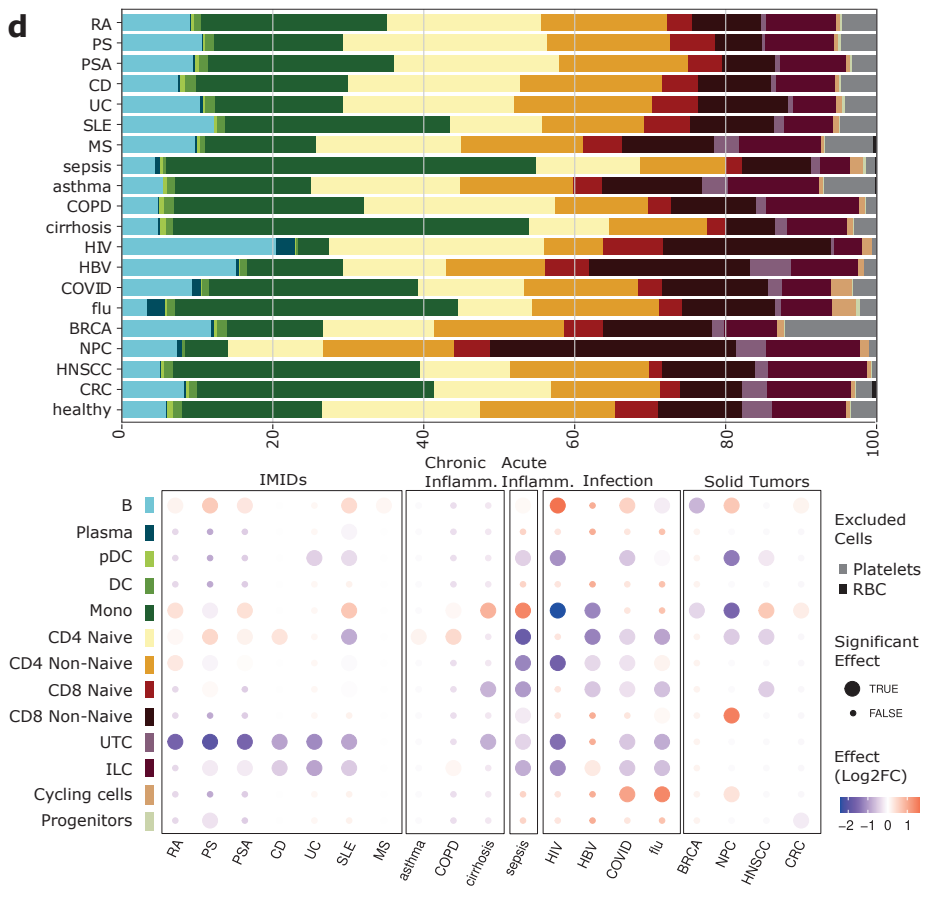
Extended Data Figure 11. Patient classifier performance in Scenarios 2 and 3. (a,c,e,g) Result obtained with the best parameter configuration for each integration and mapping method, considering W-F1 score computed on prediction of VALIDATION dataset. **(b,d,f,h)** Result obtained with the best parameter configuration for each integration and mapping method, considering W-F1 score computed on prediction of EXTERNAL dataset. In Panels (**a**) to (**h**): (**left**) Pointplot of weighted F1-scores for Majority vote and each cell-type. (**center**) F1-score for each combination of cell-type and disease, columns ordered for similarities. (**right**) Normalized confusion matrices displaying proportion of predictions belonging to each true condition. Diagonal values correspond to the Recall metric. Corresponding Majority Vote Weighted F1 score (WF1) and Balanced Accuracy Score (BAS) were reported. Note, scPoli configurations where embedding space = *sample* were not considered.

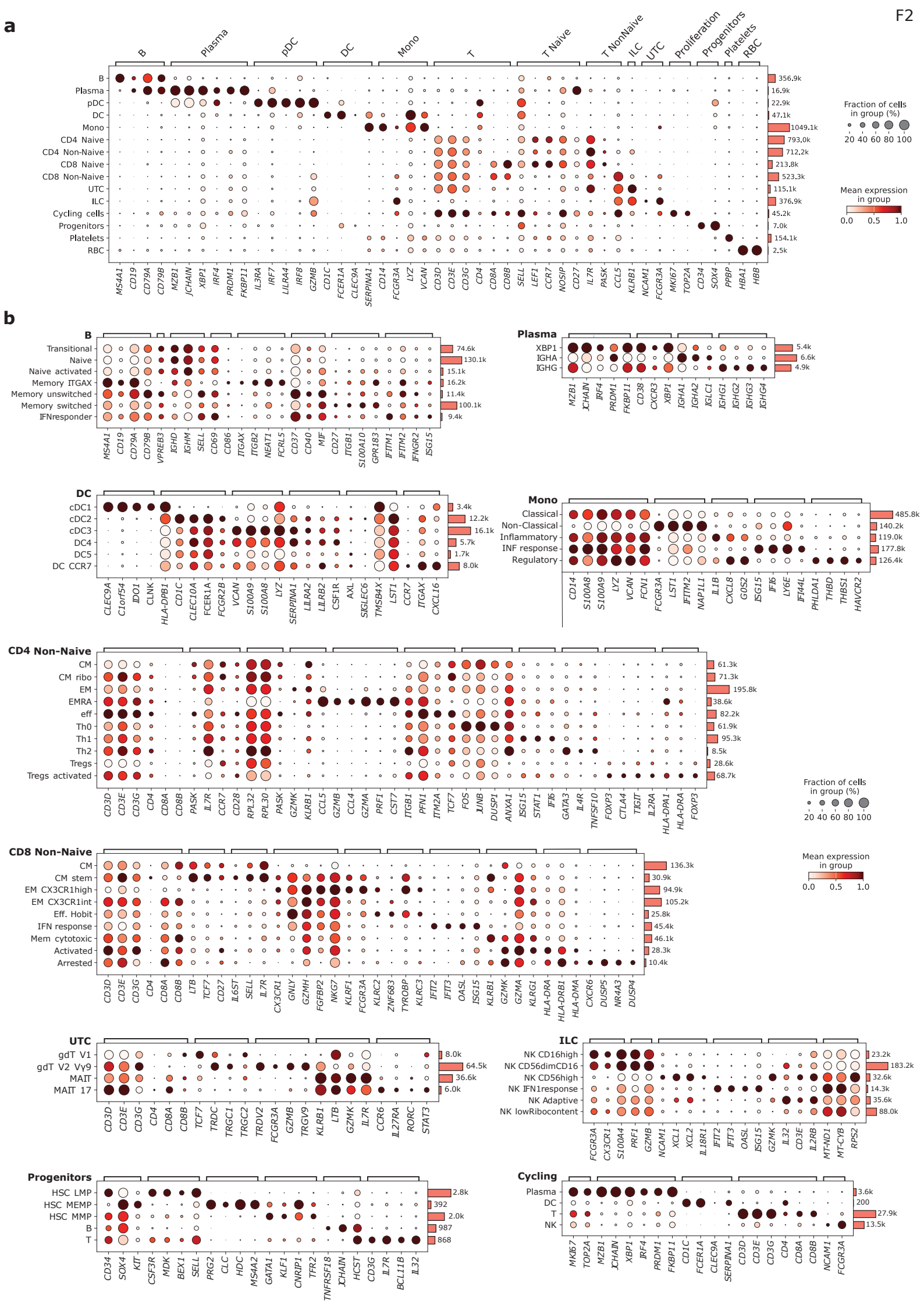


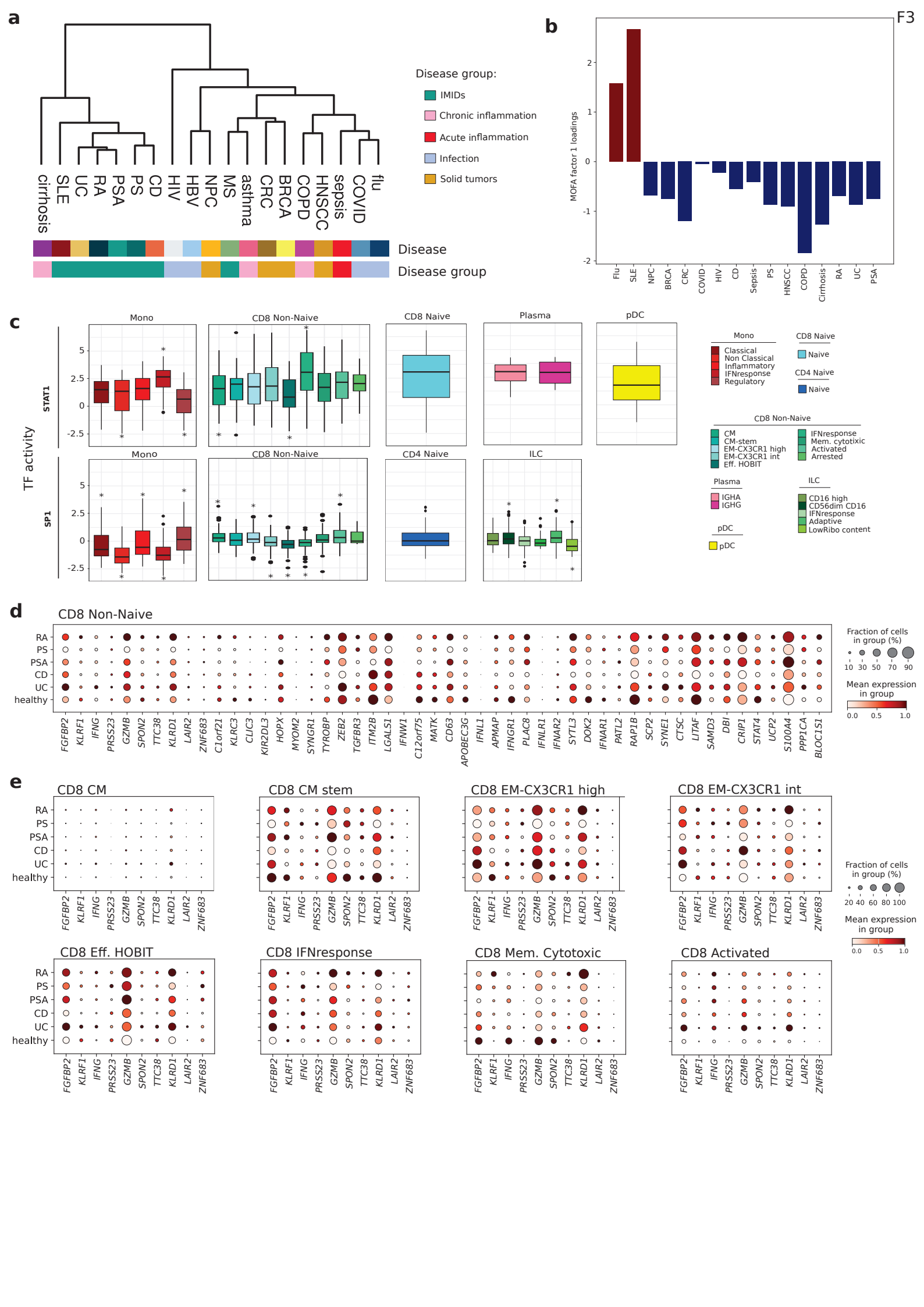
b

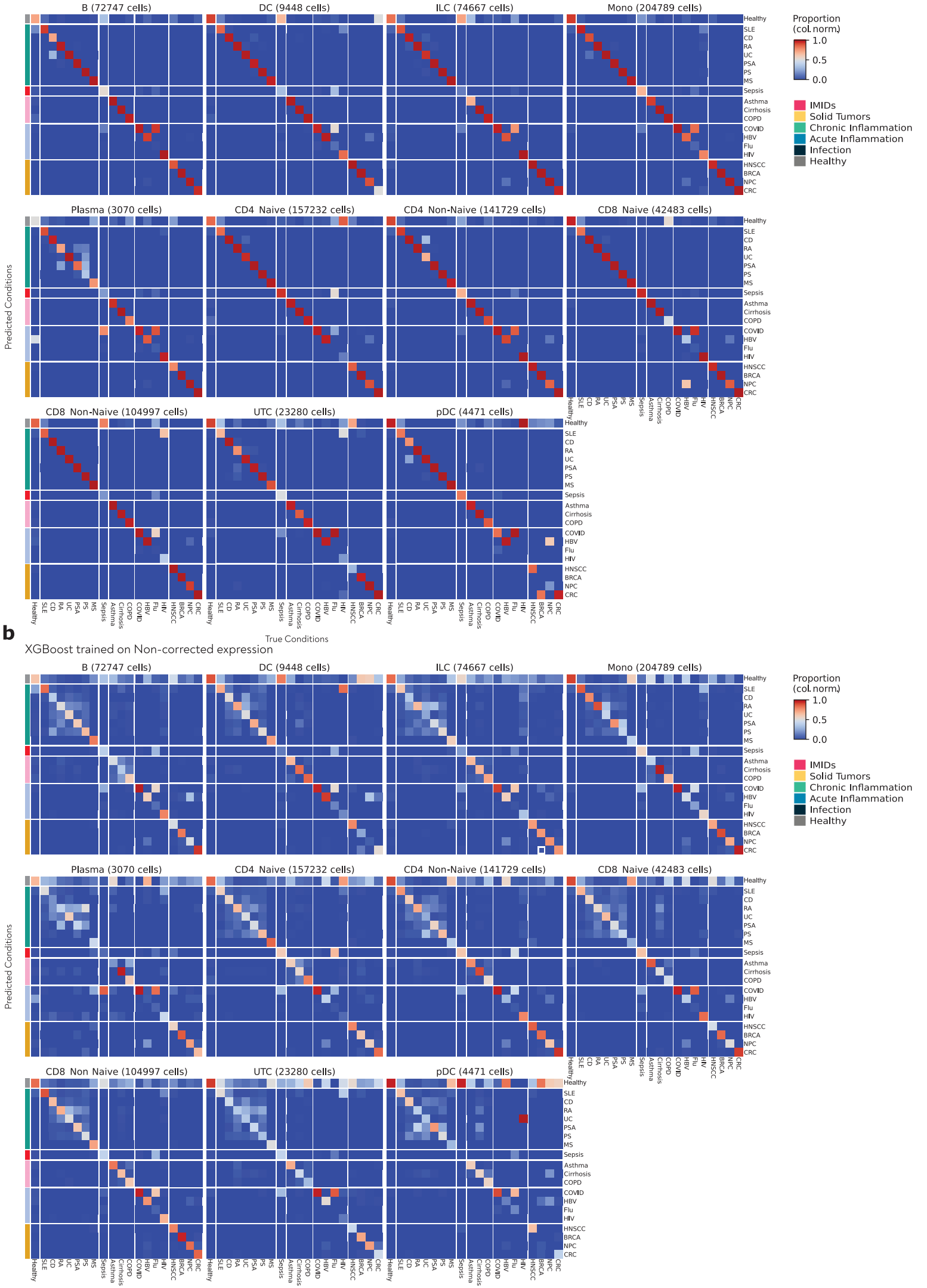
Method	Bio conservation: KMeans NMI	KMeans ARI	Silhouette label	cLISI	Batch correction: Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Aggregate score: Batch correction	Bio conservation	Total
X_scANVI_latent	0.63	0.27	0.52	0.99	0.93	0.40	0.28	0.80	0.92	0.67	0.61	0.63
X_pca_harmony	0.52	0.18	0.49	0.98	0.92	0.43	0.40	0.58	0.96	0.66	0.54	0.59
scgen_corrected_latent	0.50	0.16	0.49	0.99	0.94	0.16	0.09	0.73	0.96	0.58	0.53	0.55
X_pca_scANVI_normalized_exp	0.53	0.18	0.50	0.99	0.80	0.00	0.02	0.54	0.43	0.36	0.55	0.47
X_pca_unintegrated	0.47	0.13	0.49	0.99	0.82	0.00	0.00	0.65	0.00	0.30	0.52	0.43

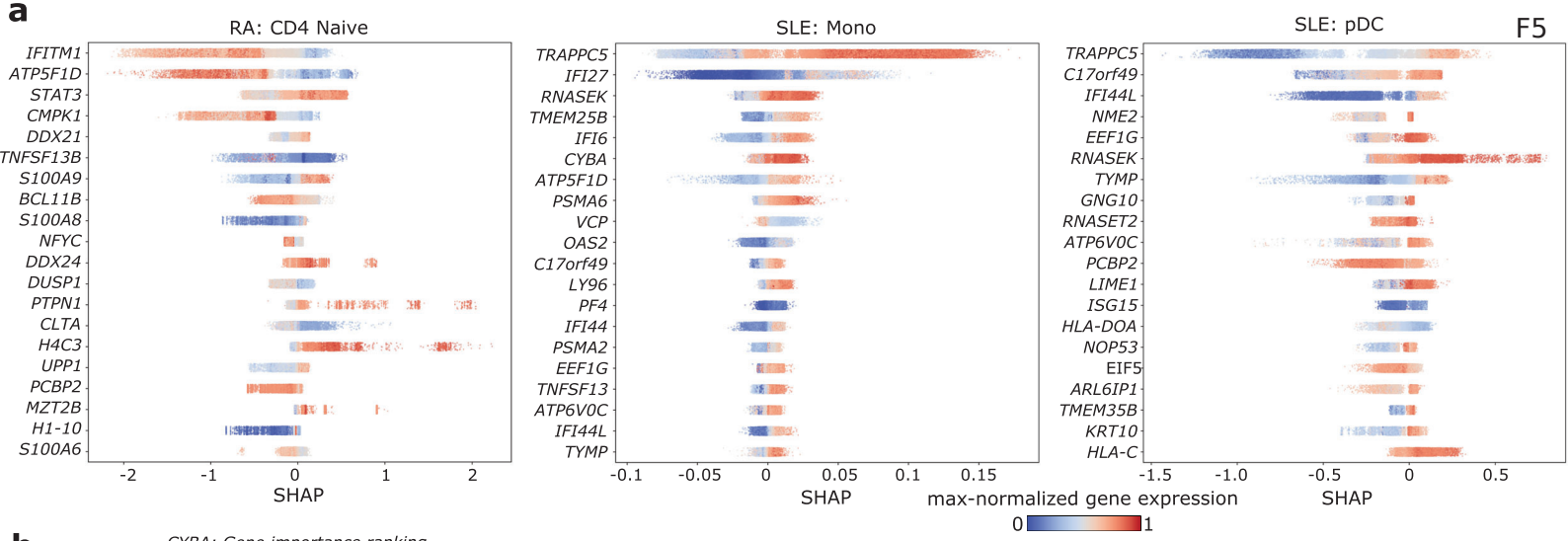


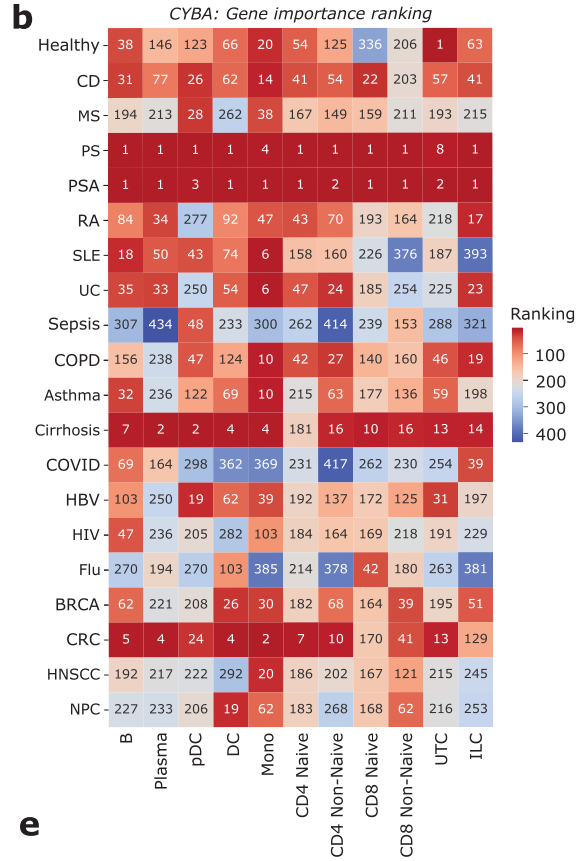


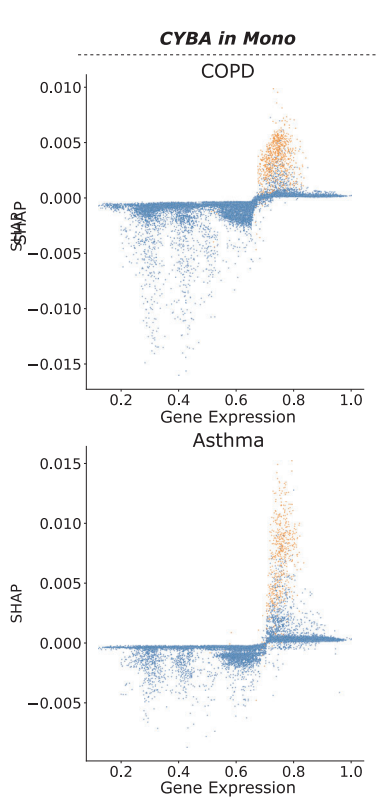

a
B
Plasma
pDC
DC
Mono
T
T Naive
T NonNaive
ILC
UTC
Proliferation
Progenitors
Platelets
RBC
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0.0 0.5 1.0
b
B
Transitional
Naive
Naive activated
Memory ITGAX
Memory unswitched
Memory switched
IFNresponder
Plasma
XBP1
IGHA
IGHG
DC
cDC1
cDC2
cDC3
DC4
DC5
DC CCR7
Mono
Classical
Non-Classical
Inflammatory
INF response
Regulatory
CD4 Non-Naive
CM
CM ribo
EM
EMRA
eff
Th0
Th1
Th2
Tregs
Tregs activated
CD8 Non-Naive
CM
CM stem
EM CX3CR1high
EM CX3CR1int
Eff. Hobit
IFN response
Mem cytotoxic
Activated
Arrested
UTC
gdT V1
gdT V2 Vγ9
MAIT
MAIT 17
ILC
NK CD16high
NK CD56dimCD16
NK CD56high
NK IFN1response
NK Adaptive
NK lowRibocontent
Progenitors
HSC LMP
HSC MEMP
HSC MMP
B
T
Cycling
Plasma
DC
T
NK

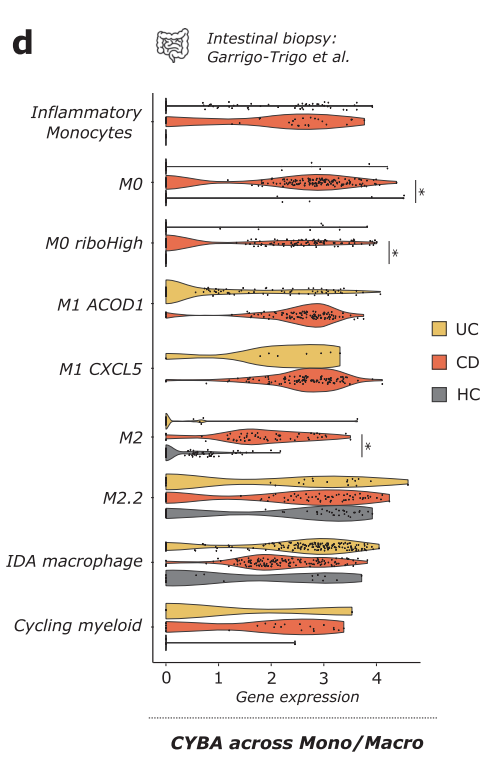

a
cirrhosis
SLE
UC
RA
PSA
PS
CD
HIV
HBV
NPC
MS
asthma
CRC
BRCA
COPD
HNSCC
sepsis
COVID
flu
Disease group:
IMIDs
Chronic inflammation
Acute inflammation
Infection
Solid tumors
Disease
Disease group
b
MOFA factor 1 loadings
Flu
SLE
NPC
BRCA
CRC
COVID
HIV
CD
Sepsis
PS
HNSCC
COPD
Cirrhosis
RA
UC
PSA
c
TF activity
STAT1
SP1
Mono
CD8 Non-Naive
CD8 Naive
Plasma
pDC
CD4 Naive
ILC
Classical
Non Classical
Inflammatory
IFNresponse
Regulatory
Naive
CM
CM-stem
EM-CX3CR1 high
EM-CX3CR1 int
Eff. HOBIT
IFNresponse
Mem. cytotixic
Activated
Arrested
IGHA
IGHG
CD16 high
CD56dim CD16
IFNresponse
Adaptive
LowRibo content
d
CD8 Non-Naive
RA
PS
PSA
CD
UC
healthy
FGFBP2
KLRF1
IFNG
PRSS23
GZMB
SPON2
TTC38
KLRD1
LAIR2
ZNF683
C1orf21
KLRC3
CLIC3
KIR2DL3
HOPX
MYOM2
SYNGR1
TYROBP
ZEB2
TGFBR3
ITM2B
LGALS1
IFNW1
C12orf75
MATK
CD63
APOBEC3G
IFNL1
APMAP
IFNGR1
PLAC8
IFNLR1
IFNAR2
SYTL3
DOK2
IFNAR1
PATL2
RAP1B
SCP2
SYNE1
CTSC
LITAF
SAMD3
DBI
CRIP1
STAT4
UCP2
S100A4
PPP1CA
BLOC1S1
Fraction of cells in group (%)
10 30 50 70 90
Mean expression in group
0.0 0.5 1.0
e
CD8 CM
CD8 CM stem
CD8 EM-CX3CR1 high
CD8 EM-CX3CR1 int
CD8 Eff. HOBIT
CD8 IFNresponse
CD8 Mem. Cytotoxic
CD8 Activated
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0.0 0.5 1.0

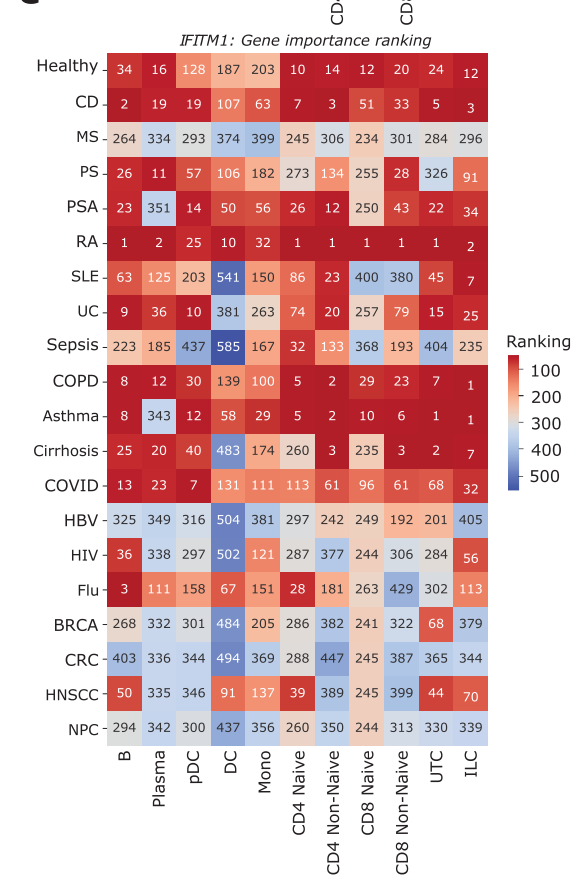

a
XGBoost trained on scANVI-corrected expression
B (72747 cells)
DC (9448 cells)
ILC (74667 cells)
Mono (204789 cells)
Plasma (3070 cells)
CD4 Naive (157232 cells)
CD4 Non-Naive (141729 cells)
CD8 Naive (42483 cells)
CD8 Non-Naive (104997 cells)
UTC (23280 cells)
pDC (4471 cells)
Predicted Conditions
True Conditions
Proportion (col. norm.)
1.0
0.5
0.0
IMIDs
Solid Tumors
Chronic Inflammation
Acute Inflammation
Infection
Healthy
Healthy
SLE
CD
RA
UC
PSA
PS
MS
Sepsis
Asthma
Cirrhosis
COPD
COVID
HBV
Flu
HIV
HNSCC
BRCA
NPC
CRC
b
XGBoost trained on Non-corrected expression
B (72747 cells)
DC (9448 cells)
ILC (74667 cells)
Mono (204789 cells)
Plasma (3070 cells)
CD4 Naive (157232 cells)
CD4 Non-Naive (141729 cells)
CD8 Naive (42483 cells)
CD8 Non Naive (104997 cells)
UTC (23280 cells)
pDC (4471 cells)
Predicted Conditions
True Conditions
Proportion (col. norm.)
1.0
0.5
0.0
IMIDs
Solid Tumors
Chronic Inflammation
Acute Inflammation
Infection
Healthy

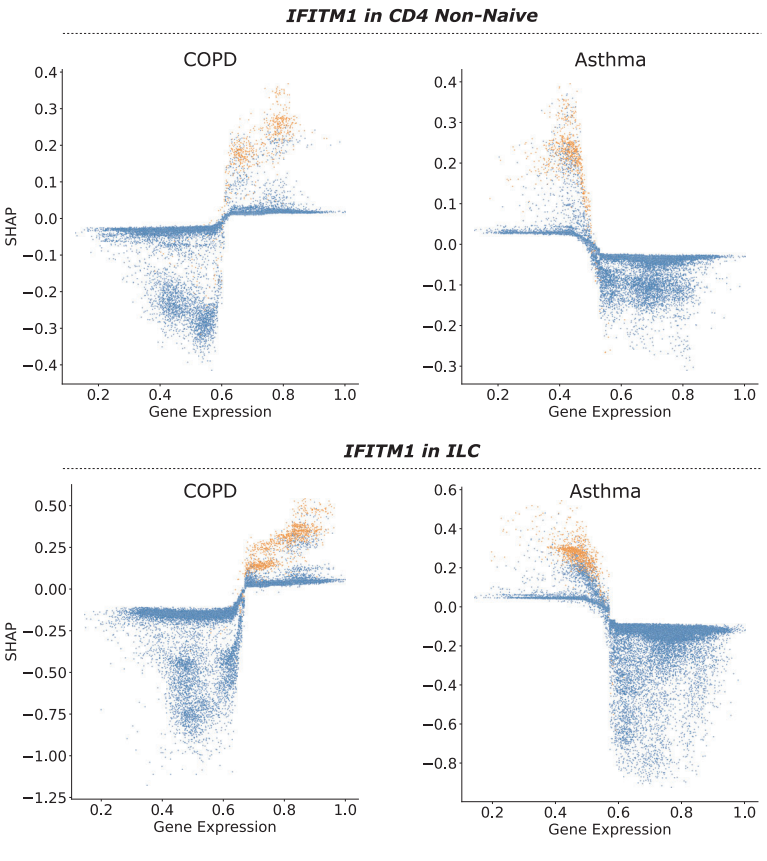

a
RA: CD4 Naive
IFITM1
ATP5F1D
STAT3
CMPK1
DDX21
TNFSF13B
S100A9
BCL11B
S100A8
NFYC
DDX24
DUSP1
PTPN1
CLTA
H4C3
UPP1
PCBP2
MZT2B
H1-10
S100A6
-2 -1 0 1 2
SHAP
SLE: Mono
TRAPPC5
IFI27
RNASEK
TMEM25B
IFI6
CYBA
ATP5F1D
PSMA6
VCP
OAS2
C17orf49
LY96
PF4
IFI44
PSMA2
EEF1G
TNFSF13
ATP6V0C
IFI44L
TYMP
-0.1 -0.05 0 0.05 0.1 0.15
SHAP
SLE: pDC
TRAPPC5
C17orf49
IFI44L
NME2
EEF1G
RNASEK
TYMP
GNG10
RNASET2
ATP6V0C
PCBP2
LIME1
ISG15
HLA-DOA
NOP53
EIF5
ARL6IP1
TMEM35B
KRT10
HLA-C
-1.5 -1.0 -0.5 0 0.5
SHAP
max-normalized gene expression
0
1

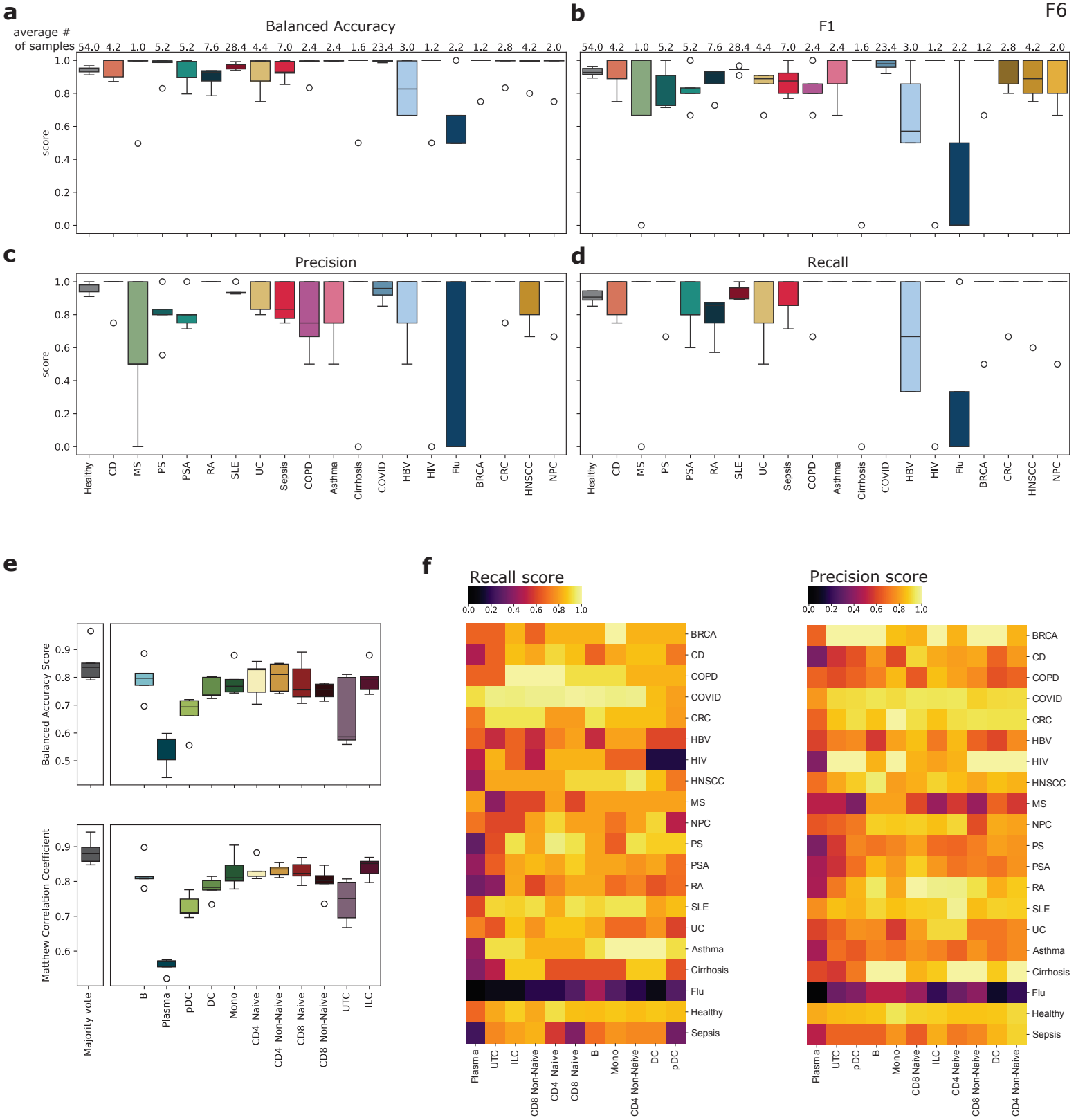

b
CYBA: Gene importance ranking
Healthy 38 146 123 66 20 54 125 336 206 1 63
CD 31 77 26 62 14 41 54 22 203 57 41
MS 194 213 28 262 38 167 149 159 211 193 215
PS 1 1 1 1 4 1 1 1 1 8 1
PSA 1 1 3 1 1 1 2 1 1 2 1
RA 84 34 277 92 47 43 70 193 164 218 17
SLE 18 50 43 74 6 158 160 226 376 187 393
UC 35 33 250 54 6 47 24 185 254 225 23
Sepsis 307 434 48 233 300 262 414 239 153 288 321
COPD 156 238 47 124 10 42 27 140 160 46 19
Asthma 32 236 122 69 10 215 63 177 136 59 198
Cirrhosis 7 2 2 4 4 181 16 10 16 13 14
COVID 69 164 298 362 369 231 417 262 230 254 39
HBV 103 250 19 62 39 192 137 172 125 31 197
HIV 47 236 205 282 103 184 164 169 218 191 229
Flu 270 194 270 103 385 214 378 42 180 263 381
BRCA 62 221 208 26 30 182 68 164 39 195 51
CRC 5 4 24 4 2 7 10 170 41 13 129
HNSCC 192 217 222 292 20 186 202 167 121 215 245
NPC 227 233 206 19 62 183 268 168 62 216 253
B
Plasma
pDC
DC
Mono
CD4 Naive
CD4 Non-Naive
CD8 Naive
CD8 Non-Naive
UTC
ILC
Ranking
100
200
300
400


c
CYBA in Mono
COPD
SHAP
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.2 0.4 0.6 0.8 1.0
Gene Expression
Asthma
SHAP
0.015
0.010
0.005
0.000
−0.005
0.2 0.4 0.6 0.8 1.0
Gene Expression


d
Intestinal biopsy:
Garrigo-Trigo et al.
Inflammatory Monocytes
M0
M0 riboHigh
M1 ACOD1
M1 CXCL5
M2
M2.2
IDA macrophage
Cycling myeloid
UC
CD
HC
0 1 2 3 4
Gene expression
CYBA across Mono/Macro


e
IFITM1: Gene importance ranking
Healthy 34 16 128 187 203 10 14 12 20 24 12
CD 2 19 19 107 63 7 3 51 33 5 3
MS 264 334 293 374 399 245 306 234 301 284 296
PS 26 11 57 106 182 273 134 255 28 326 91
PSA 23 351 14 50 56 26 12 250 43 22 34
RA 1 2 25 10 32 1 1 1 1 1 2
SLE 63 125 203 541 150 86 23 400 380 45 7
UC 9 36 10 381 263 74 20 257 79 15 25
Sepsis 223 185 437 585 167 32 133 368 193 404 235
COPD 8 12 30 139 100 5 2 29 23 7 1
Asthma 8 343 12 58 29 5 2 10 6 1 1
Cirrhosis 25 20 40 483 174 260 3 235 3 2 7
COVID 13 23 7 131 111 113 61 96 61 68 32
HBV 325 349 316 504 381 297 242 249 192 201 405
HIV 36 338 297 502 121 287 377 244 306 284 56
Flu 3 111 158 67 151 28 181 263 429 302 113
BRCA 268 332 301 484 205 286 382 241 322 68 379
CRC 403 336 344 494 369 288 447 245 387 365 344
HNSCC 50 335 346 91 137 39 389 245 399 44 70
NPC 294 342 300 437 356 260 350 244 313 330 339
B
Plasma
pDC
DC
Mono
CD4 Naive
CD4 Non-Naive
CD8 Naive
CD8 Non-Naive
UTC
ILC
Ranking
100
200
300
400
500


f
IFITM1 in CD4 Non-Naive
COPD
Asthma
SHAP
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0.2 0.4 0.6 0.8 1.0
Gene Expression
IFITM1 in ILC
COPD
Asthma
SHAP
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.2 0.4 0.6 0.8 1.0
Gene Expression


a
Balanced Accuracy
average # of samples
54.0 4.2 1.0 5.2 5.2 7.6 28.4 4.4 7.0 2.4 2.4 1.6 23.4 3.0 1.2 2.2 1.2 2.8 4.2 2.0
score
b
F1
c
Precision
d
Recall
Healthy CD MS PS PSA RA SLE UC Sepsis COPD Asthma Cirrhosis COVID HBV HIV Flu BRCA CRC HNSCC NPC

e
Balanced Accuracy Score
Matthew Correlation Coefficient
Majority vote B Plasma pDC DC Mono CD4 Naive CD4 Non-Naive CD8 Naive CD8 Non-Naive UTC ILC
f
Recall score
Precision score
BRCA CD COPD COVID CRC HBV HIV HNSCC MS NPC PS PSA RA SLE UC Asthma Cirrhosis Flu Healthy Sepsis
Plasma UTC ILC CD8 Non-Naive CD4 Naive CD8 Naive B Mono CD4 Non-Naive DC pDC
Plasma UTC pDC B Mono CD8 Naive ILC CD4 Naive CD8 Non-Naive DC CD4 Non-Naive

a

Method	Emb Space	Emb Dim	PEmb Dim	annot	wF1 VAL	wF1 EXT	BAS VAL	BAS EXT	MCC VAL	MCC EXT	wPRE VAL	wPRE EXT	wREC VAL	wREC EXT	wSPE VAL	wSPE EXT	bestClf VAL	bestClf EXT
HS	cell	20	0	NU	0.84	0.47	0.69	0.24	0.8	0.38	0.87	0.46	0.84	0.5	0.97	0.93	SVC	SVC
scPoli	cellSample	200	100	1&2	0.66	0.46	0.42	0.24	0.62	0.45	0.64	0.42	0.71	0.55	0.89	0.89	SVC	SVC
scPoli	sample	200	100	1&2	0.6	0.46	0.36	0.24	0.55	0.46	0.6	0.42	0.65	0.56	0.86	0.9	SVC	SVC
HS	cell	30	0	NU	0.81	0.45	0.69	0.24	0.76	0.37	0.88	0.47	0.81	0.48	0.93	0.93	SVC	SVC
HS	cell	50	0	NU	0.86	0.42	0.76	0.24	0.83	0.34	0.9	0.47	0.86	0.44	0.95	0.92	SVC	SVC
scPoli	cell	100	20	1&2	0.96	0.41	0.95	0.18	0.95	0.27	0.97	0.49	0.96	0.36	0.99	0.96	LSVC	LSVC
scPoli	cell	200	100	1&2	0.94	0.4	0.88	0.16	0.92	0.24	0.96	0.55	0.94	0.34	0.99	0.97	LSVC	LSVC
HS	cell	200	0	NU	0.94	0.37	0.92	0.18	0.9	0.25	0.96	0.46	0.92	0.35	0.98	0.95	LSVC	LSVC
scPoli	sample	20	100	1&2	0.59	0.36	0.44	0.18	0.53	0.26	0.66	0.43	0.63	0.38	0.85	0.9	SVC	SVC
scPoli	cell	50	20	1&2	0.94	0.36	0.9	0.15	0.9	0.21	0.97	0.52	0.92	0.28	0.99	0.97	LSVC	LSVC
scPoli	sample	100	100	1&2	0.63	0.35	0.5	0.18	0.56	0.28	0.71	0.44	0.65	0.38	0.87	0.9	SVC	SVC
scGen	cell	20	0	2	0.96	0.34	0.91	0.15	0.94	0.23	0.97	0.45	0.95	0.33	0.99	0.94	SVC	SVC
HS	cell	100	0	NU	0.85	0.33	0.78	0.19	0.8	0.25	0.9	0.46	0.84	0.35	0.94	0.92	SVC	SVC
scPoli	cell	50	100	1&2	0.97	0.32	0.96	0.11	0.95	0.17	0.99	0.49	0.96	0.24	0.99	0.98	LSVC	LSVC
scPoli	cell	30	20	1&2	0.92	0.32	0.84	0.13	0.89	0.18	0.95	0.47	0.91	0.24	0.99	0.98	LSVC	LSVC
scPoli	cellSample	20	100	1&2	0.64	0.31	0.49	0.14	0.57	0.19	0.68	0.43	0.66	0.3	0.88	0.91	SVC	SVC
scPoli	sample	30	20	1&2	0.84	0.3	0.77	0.14	0.8	0.19	0.88	0.45	0.84	0.29	0.96	0.93	SVC	SVC
scGen	cell	30	0	2	0.96	0.3	0.92	0.12	0.93	0.17	0.98	0.45	0.94	0.26	0.99	0.95	SVC	SVC
scPoli	cell	100	100	1&2	0.96	0.29	0.9	0.15	0.93	0.17	0.98	0.48	0.94	0.24	0.99	0.95	LSVC	LSVC
scANVI	cell	20	0	2	0.95	0.29	0.9	0.11	0.91	0.15	0.97	0.45	0.93	0.23	0.99	0.96	SVC	SVC
scPoli	cellSample	30	100	1&2	0.74	0.28	0.58	0.14	0.69	0.17	0.78	0.44	0.76	0.28	0.93	0.91	SVC	SVC
scPoli	cell	20	100	1&2	0.92	0.27	0.79	0.12	0.87	0.14	0.96	0.48	0.9	0.2	0.99	0.97	SVC	SVC
scGen	cell	100	0	2	0.98	0.27	0.95	0.12	0.96	0.17	0.99	0.46	0.97	0.24	1.0	0.95	SVC	SVC
scPoli	sample	30	100	1&2	0.7	0.27	0.56	0.15	0.65	0.19	0.75	0.44	0.72	0.29	0.89	0.91	SVC	SVC
scPoli	sample	20	20	1&2	0.91	0.26	0.87	0.15	0.89	0.19	0.93	0.45	0.91	0.28	0.97	0.92	SVC	SVC
scANVI	cell	200	0	2	0.97	0.25	0.96	0.12	0.95	0.16	0.98	0.47	0.96	0.22	0.99	0.96	LSVC	SVC
scGen	cell	200	0	2	0.97	0.25	0.92	0.12	0.95	0.15	0.99	0.45	0.96	0.23	1.0	0.94	SVC	SVC
scGen	cell	50	0	2	0.98	0.24	0.95	0.11	0.96	0.14	0.99	0.45	0.97	0.22	1.0	0.94	SVC	SVC
scPoli	cellSample	30	20	1&2	0.91	0.24	0.91	0.11	0.88	0.14	0.93	0.46	0.9	0.21	0.97	0.95	SVC	SVC
scPoli	cell	20	50	1&2	0.89	0.23	0.74	0.09	0.82	0.11	0.94	0.44	0.85	0.17	0.99	0.96	SVC	SVC
scPoli	cell	30	100	1&2	0.91	0.23	0.83	0.08	0.86	0.11	0.95	0.49	0.89	0.15	0.99	0.98	LSVC	LSVC
scANVI	cell	30	0	2	0.98	0.23	0.95	0.12	0.96	0.17	0.98	0.5	0.97	0.21	0.99	0.97	SVC	SVC
scPoli	cellSample	20	20	1&2	0.89	0.21	0.84	0.1	0.84	0.11	0.92	0.43	0.88	0.2	0.98	0.93	SVC	SVC
scPoli	cellSample	100	100	1&2	0.68	0.2	0.62	0.14	0.61	0.16	0.75	0.44	0.69	0.24	0.91	0.9	SVC	SVC
scANVI	cell	100	0	2	0.96	0.2	0.93	0.11	0.93	0.14	0.97	0.47	0.94	0.2	0.99	0.96	SVC	SVC
scPoli	sample	50	20	1&2	0.8	0.18	0.7	0.12	0.74	0.12	0.87	0.47	0.79	0.19	0.94	0.95	SVC	SVC
scPoli	cellSample	50	100	1&2	0.71	0.16	0.52	0.11	0.64	0.11	0.83	0.45	0.72	0.19	0.93	0.93	SVC	SVC
scANVI	cell	50	0	2	0.96	0.16	0.91	0.11	0.93	0.13	0.97	0.46	0.94	0.19	0.99	0.95	SVC	SVC
scPoli	cellSample	50	20	1&2	0.83	0.16	0.74	0.09	0.78	0.1	0.89	0.45	0.82	0.16	0.96	0.95	SVC	SVC
scPoli	sample	50	100	1&2	0.69	0.16	0.55	0.12	0.63	0.11	0.76	0.44	0.71	0.2	0.9	0.92	SVC	SVC
scPoli	cellSample	20	50	1&2	0.83	0.14	0.77	0.11	0.78	0.11	0.88	0.44	0.82	0.19	0.94	0.92	SVC	SVC
scPoli	sample	20	50	1&2	0.64	0.13	0.4	0.09	0.57	0.08	0.74	0.44	0.64	0.15	0.95	0.94	LSVC	LSVC
scPoli	sample	100	20	1&2	0.83	0.12	0.79	0.11	0.78	0.1	0.87	0.44	0.83	0.19	0.94	0.91	SVC	SVC
scPoli	cell	20	20	1&2	0.93	0.12	0.86	0.06	0.9	0.04	0.95	0.41	0.92	0.08	0.99	0.96	SVC	SVC
scPoli	cellSample	100	20	1&2	0.86	0.1	0.87	0.09	0.83	0.05	0.89	0.24	0.86	0.15	0.95	0.9	SVC	SVC
random	na	nan	nan	NaN	0.2	0.1	0.09	0.1	-0.0	-0.0	0.21	0.21	0.19	0.11	0.81	0.89	rand	rand

Annotation (annot)

1&2 = Level 1 and 2
2 = Level 2

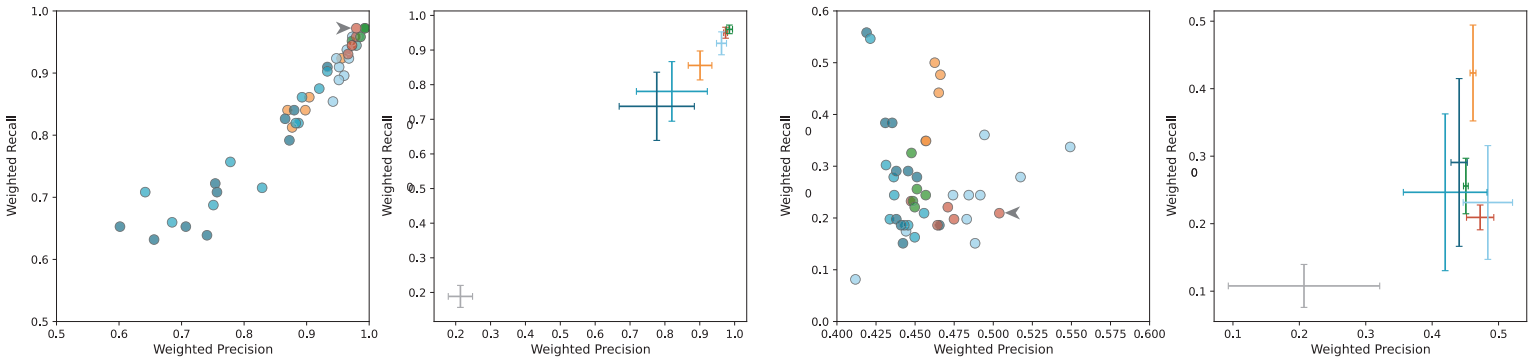
NU = Not Utilized

Performance metrics

wF1 = weighted F1
BAS = Balanced accuracy score
MCC = Matthew Correlation Coef.
wPRE = weighted Precision
wREC = weighted Recall
wSPE = weighted Specificity

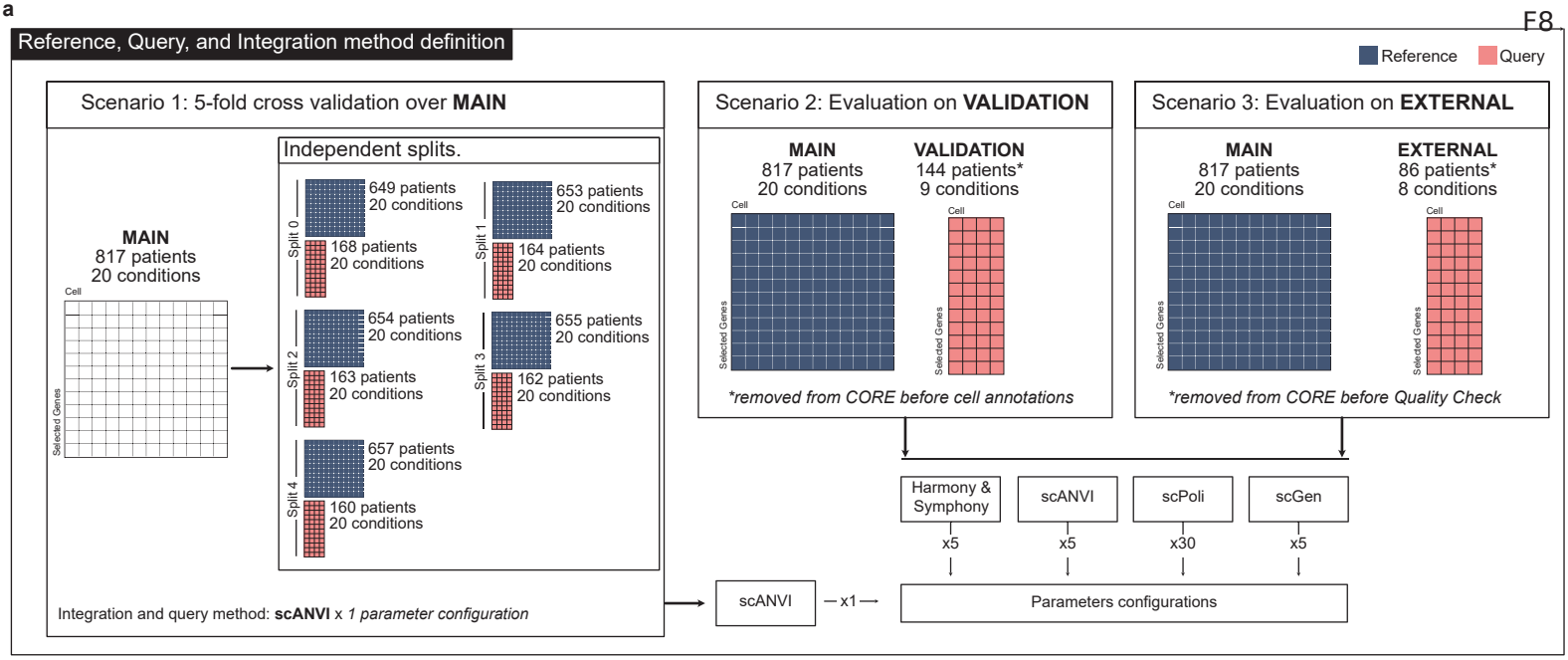
b **Scenario 2:** Validation

c **Scenario 3:** External

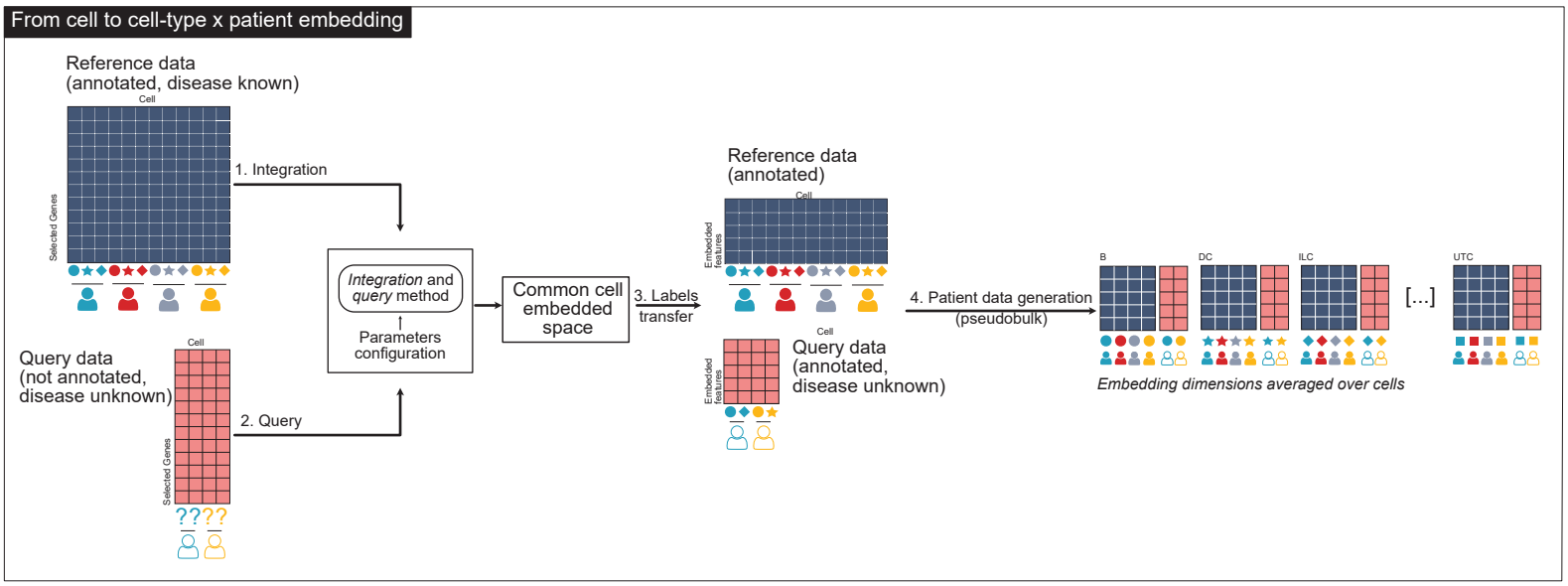


Integration method : Harmony + Symphony | scANVI | scGen | scPoli sample | scPoli cell & sample | scPoli cell | random

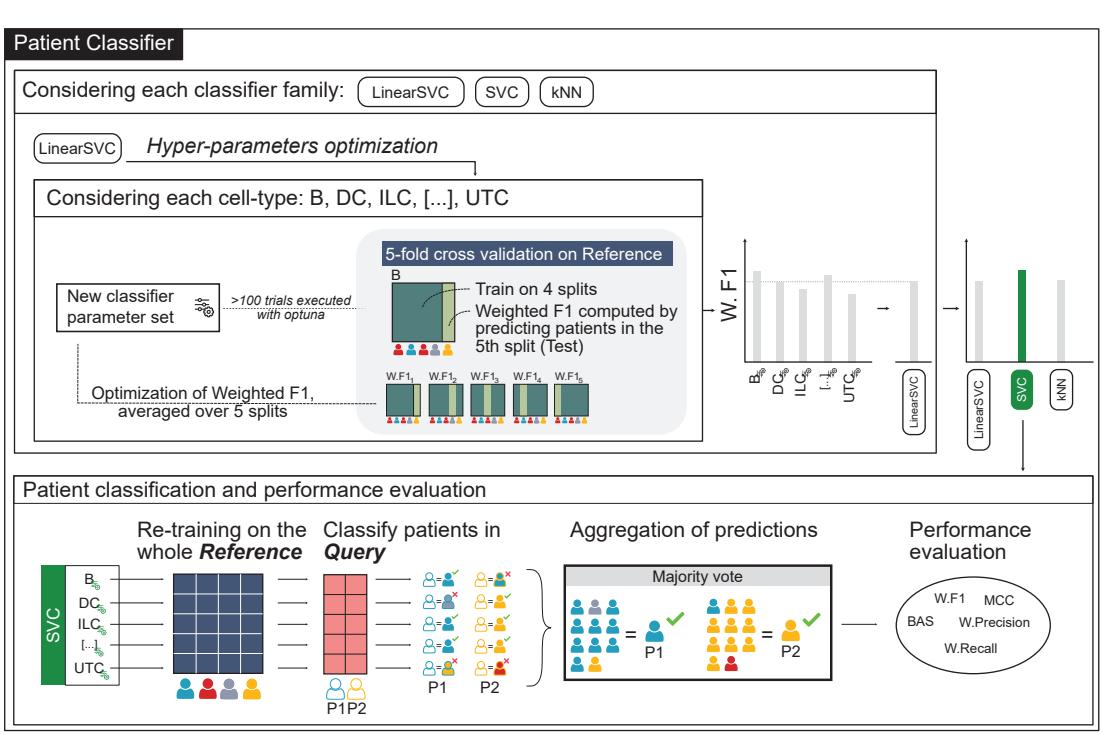
a

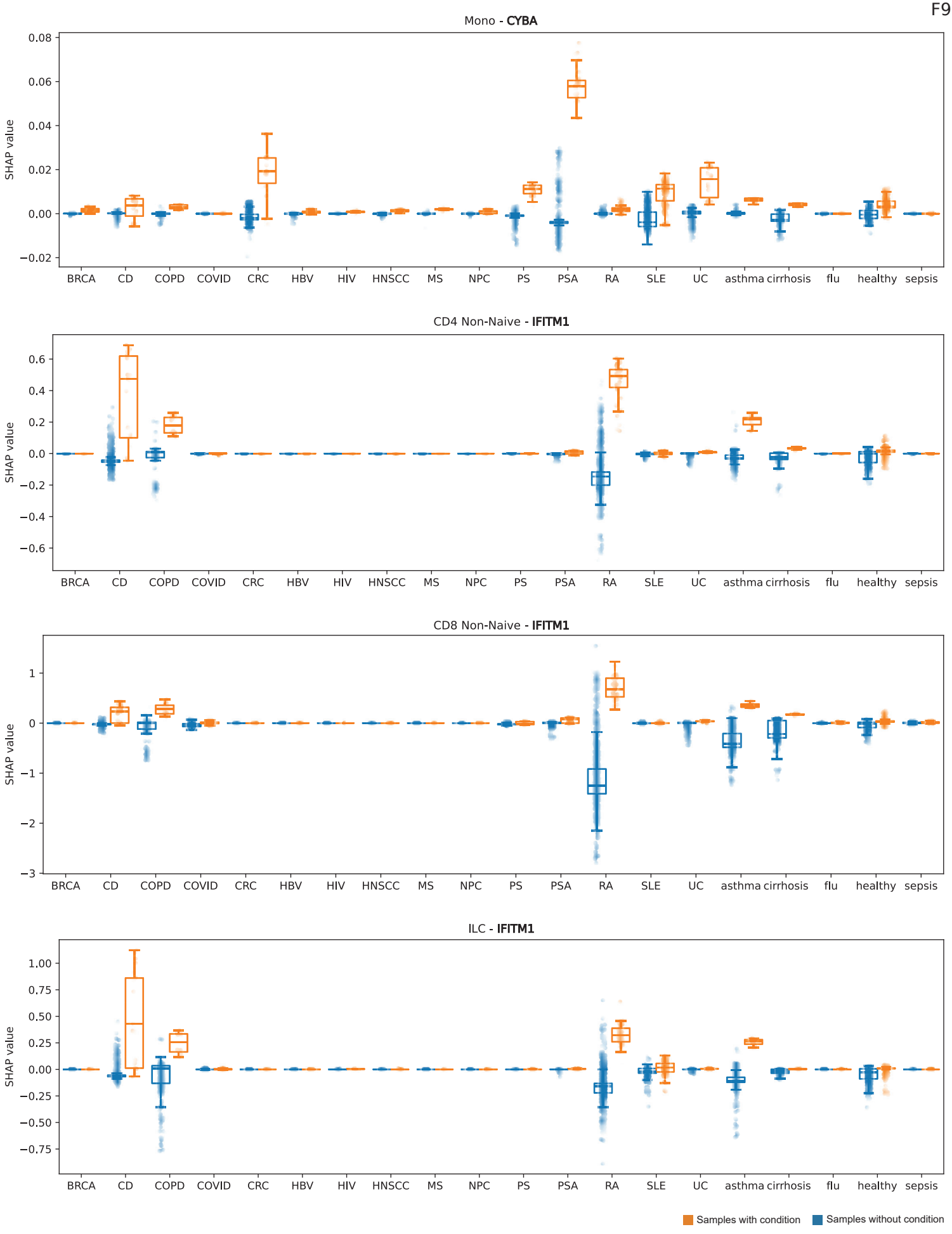

Reference, Query, and Integration method definition
Reference
Query
Scenario 1: 5-fold cross validation over MAIN
MAIN
817 patients
20 conditions
Cell
Selected Genes
Independent splits.
Split 0
649 patients
20 conditions
168 patients
20 conditions
Split 1
653 patients
20 conditions
164 patients
20 conditions
Split 2
654 patients
20 conditions
163 patients
20 conditions
Split 3
655 patients
20 conditions
162 patients
20 conditions
Split 4
657 patients
20 conditions
160 patients
20 conditions
Integration and query method: scANVI x 1 parameter configuration
Scenario 2: Evaluation on VALIDATION
MAIN
817 patients
20 conditions
VALIDATION
144 patients*
9 conditions
*removed from CORE before cell annotations
Scenario 3: Evaluation on EXTERNAL
MAIN
817 patients
20 conditions
EXTERNAL
86 patients*
8 conditions
*removed from CORE before Quality Check
Harmony & Symphony
scANVI
scPoli
scGen
x5
x5
x30
x5
scANVI
x1
Parameters configurations

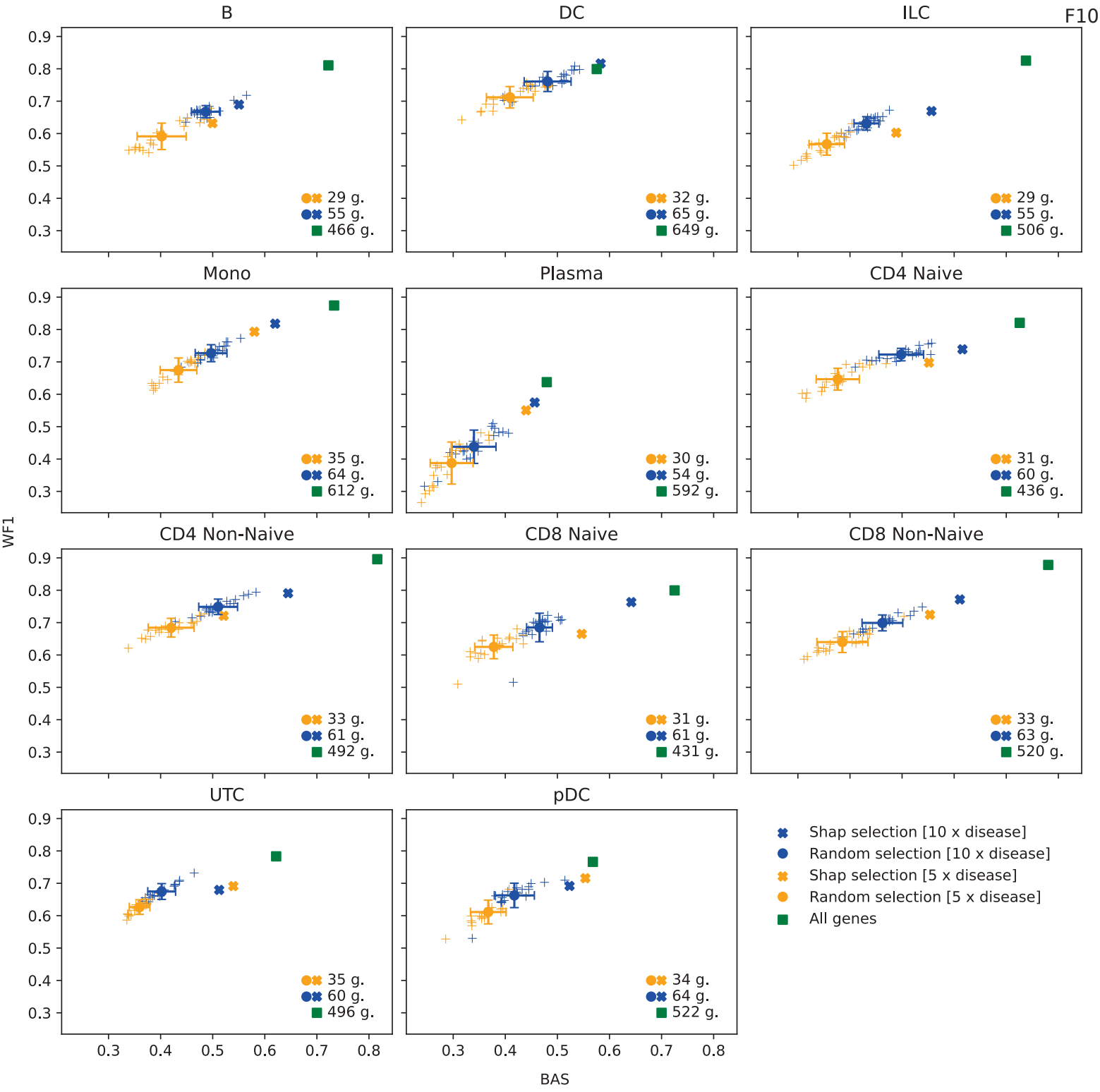
b

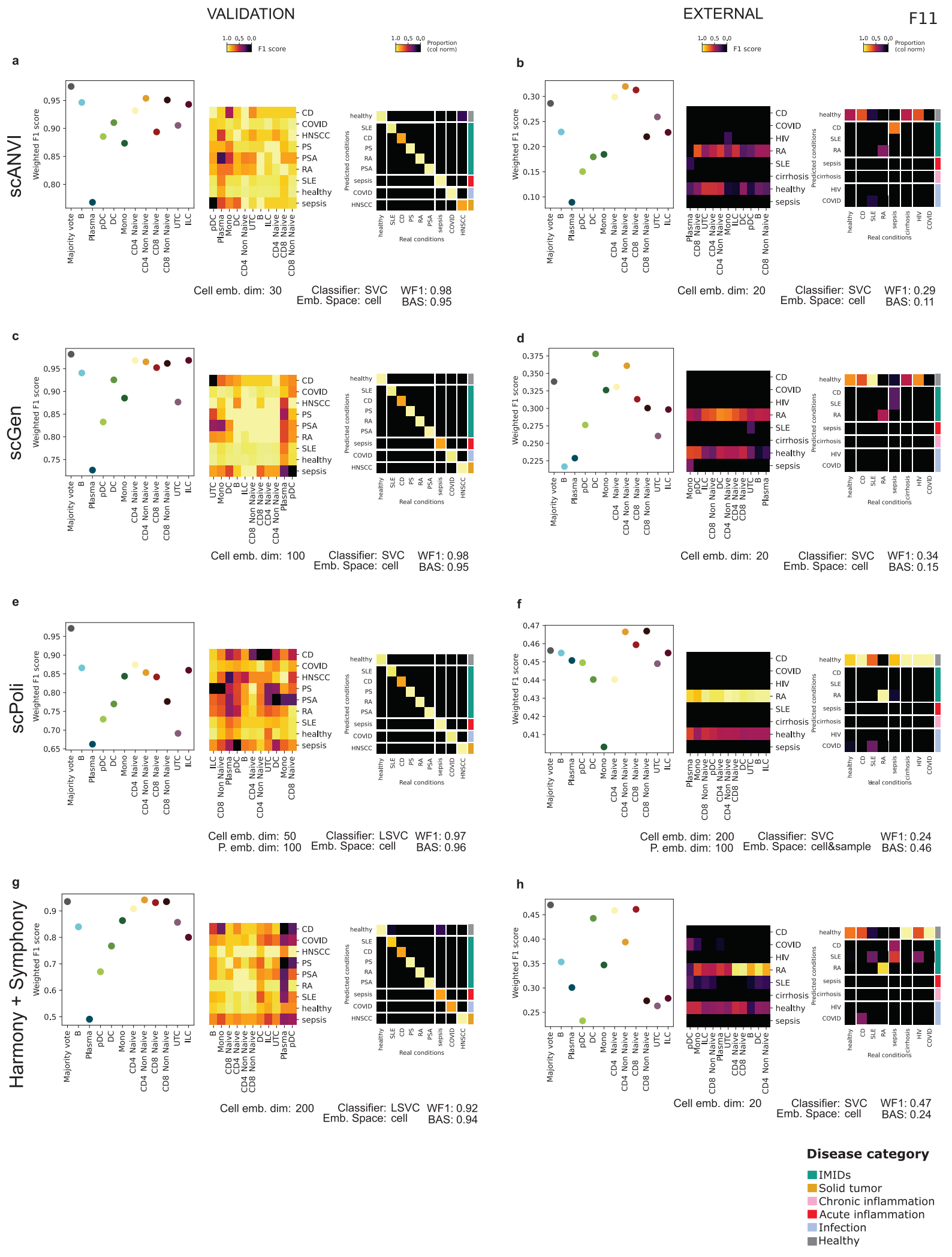

From cell to cell-type x patient embedding
Reference data
(annotated, disease known)
Cell
Selected Genes
1. Integration
Integration and query method
Parameters configuration
Common cell embedded space
3. Labels transfer
Reference data
(annotated)
Embedded features
Query data
(annotated, disease unknown)
4. Patient data generation
(pseudobulk)
B
DC
ILC
[...]
UTC
Embedding dimensions averaged over cells
Query data
(not annotated, disease unknown)
2. Query
????

c


Patient Classifier
Considering each classifier family:
LinearSVC
SVC
kNN
LinearSVC
Hyper-parameters optimization
Considering each cell-type: B, DC, ILC, [...], UTC
New classifier parameter set
>100 trials executed with optuna
Optimization of Weighted F1, averaged over 5 splits
5-fold cross validation on Reference
Train on 4 splits
Weighted F1 computed by predicting patients in the 5th split (Test)
W.F1
Patient classification and performance evaluation
Re-training on the whole Reference
Classify patients in Query
Aggregation of predictions
Majority vote
Performance evaluation
W.F1
MCC
BAS
W.Precision
W.Recall
P1
P2


Mono - CYBA
CD4 Non-Naive - IFITM1
CD8 Non-Naive - IFITM1
ILC - IFITM1
SHAP value
BRCA CD COPD COVID CRC HBV HIV HNSCC MS NPC PS PSA RA SLE UC asthma cirrhosis flu healthy sepsis
Samples with condition
Samples without condition


B
29 g.
55 g.
466 g.
DC
32 g.
65 g.
649 g.
ILC
29 g.
55 g.
506 g.
Mono
35 g.
64 g.
612 g.
Plasma
30 g.
54 g.
592 g.
CD4 Naive
31 g.
60 g.
436 g.
CD4 Non-Naive
33 g.
61 g.
492 g.
CD8 Naive
31 g.
61 g.
431 g.
CD8 Non-Naive
33 g.
63 g.
520 g.
UTC
35 g.
60 g.
496 g.
pDC
34 g.
64 g.
522 g.
WF1
BAS
Shap selection [10 x disease]
Random selection [10 x disease]
Shap selection [5 x disease]
Random selection [5 x disease]
All genes


VALIDATION
EXTERNAL
F1 score
Proportion (col norm)
scANVI
scGen
scPoli
Harmony + Symphony
a
b
c
d
e
f
g
h
Weighted F1 score
Predicted conditions
Real conditions
Majority vote
Cell emb. dim: 30
Classifier: SVC
Emb. Space: cell
WF1: 0.98
BAS: 0.95
Cell emb. dim: 20
Classifier: SVC
Emb. Space: cell
WF1: 0.29
BAS: 0.11
Cell emb. dim: 100
Classifier: SVC
Emb. Space: cell
WF1: 0.98
BAS: 0.95
Cell emb. dim: 20
Classifier: SVC
Emb. Space: cell
WF1: 0.34
BAS: 0.15
Cell emb. dim: 50
P. emb. dim: 100
Classifier: LSVC
Emb. Space: cell
WF1: 0.97
BAS: 0.96
Cell emb. dim: 200
P. emb. dim: 100
Classifier: SVC
Emb. Space: cell&sample
WF1: 0.24
BAS: 0.46
Cell emb. dim: 200
Classifier: LSVC
Emb. Space: cell
WF1: 0.92
BAS: 0.94
Cell emb. dim: 20
Classifier: SVC
Emb. Space: cell
WF1: 0.47
BAS: 0.24
Disease category
IMIDs
Solid tumor
Chronic inflammation
Acute inflammation
Infection
Healthy

Publication 3. Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis[3]

Joan Clària* , Ferran Aguilar* , Juan-José Lozano*, **Laura Jiménez-Gracia***, Juan C. Nieto*, Berta Romero-Grimaldo, Xavi Marcos-Fa , Emma Giarracco, Emmanuel Weiss, Jonel Trebicka, Inmaculada Hernàndez, Javier Fernandez, Mireia Casulleras, Cristina López-Vicario, Sinan Muldur, Alex Hopke, Alexandru Vlagea, Ana M. Aransay, Domenica Marchese, Mauro Bernardi, Rajiv Jalan, Paolo Angeli, Giuliana Magri, Andrea Cerutti, Daniel Irimia, Holger Heyn†, Vicente Arroyo†, Richard Moreau†. **Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis.** *JHEP Reports*. 6, 1-15 (2024)

* These authors contributed equally

† Joint senior authors

DOI: https://doi.org/10.1016/j.jhepr.2024.101184

Impact Factor 2023: 9.5 (Q1)

Raw & Processed data: Available upon corresponding author's request.

Code: https://github.com/LJimenezGracia/AD-ACLF_cirrhosis_HSAtreatment

[3] In this paper, I conducted all the computational analysis of the scRNA-seq data. However, I did not handle patient collection, perform *in vitro* or *in vivo* experiments, process the samples for bulk or single-cell RNA-sequencing, analyze bulk RNA-seq data or run other functional assays. Therefore, I will only discuss the results from the scRNA-seq analysis, which I can take credit from.

Research article

JHEP|Reports

Albumin reprograms the B cell transcriptional landscape and improves neutrophil antimicrobial function in patients with decompensated cirrhosis

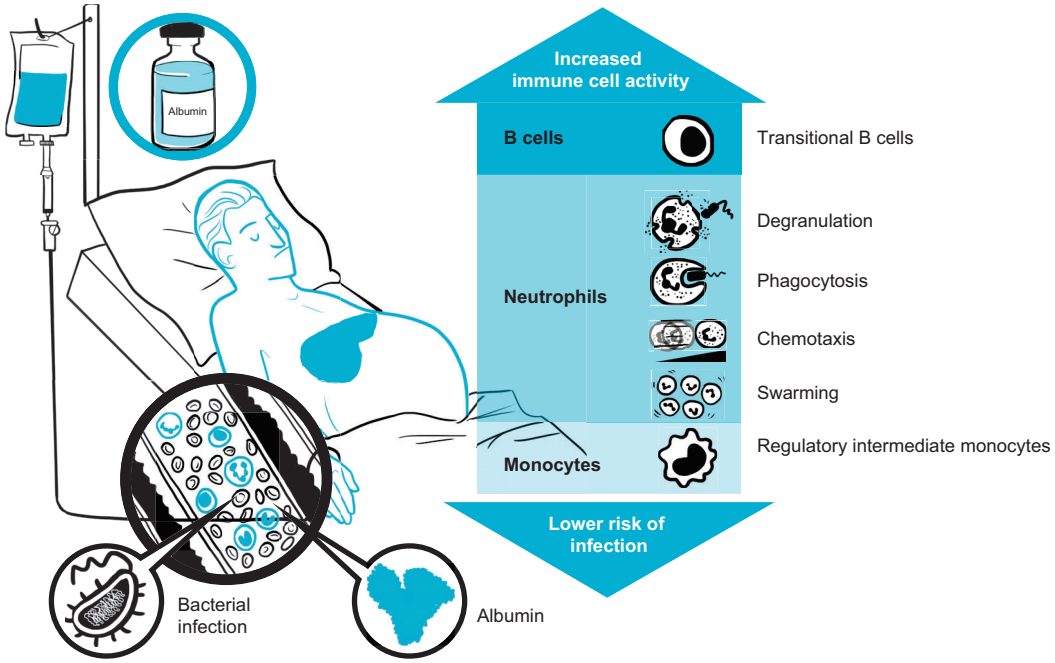
Authors

Joan Clària, Ferran Aguilar, Juan-José Lozano, …, Holger Heyn, Vicente Arroyo, Richard Moreau

Correspondence

richard.moreau@inserm.fr (R. Moreau), jclaria@clinic.cat (J. Clària).

Graphical abstract



Highlights:

- Patients with acutely decompensated cirrhosis are immunocompromised and have higher predisposition to infections.
- scRNA-seq and FACS analyses demonstrate the expansion of the B-cell compartment in mononuclear leukocytes exposed to albumin.
- Polymorphonuclear leukocytes exposed to albumin also exhibit enhanced neutrophil antimicrobial functions.
- Analysis of RNA-seq data in patients receiving albumin shows upregulated signatures related to B cells and neutrophils.
- These findings provide mechanisms by which albumin reduces the incidence of infections in patients with acutely decompensated cirrhosis.

Impact and implications:

Patients with acutely decompensated cirrhosis receiving albumin as treatment have a lower incidence of infections. The reason for this protection is currently unknown, but the present study provides data that support the ability of albumin to boost the antimicrobial functions of immune cells in these patients. Moreover, these findings encourage the design of controlled clinical studies specifically aimed at investigating the effects of albumin administration on the immune system.

https://doi.org/10.1016/j.jhepr.2024.101184

© 2024 The Author(s). Published by Elsevier B.V. on behalf of European Association for the Study of the Liver (EASL). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). JHEP Reports, 2024, 6, 1–15

Research article

JHEP|Reports

Albumin reprograms the B cell transcriptional landscape and improves neutrophil antimicrobial function in patients with decompensated cirrhosis

Joan Clària[1,2,3,4,*,†], **Ferran Aguilar**[1,†], **Juan-José Lozano**[3,†], **Laura Jiménez-Gracia**[4,5,†], **Juan C. Nieto**[4,5,†], Berta Romero-Grimaldo[1,2,3], Xavi Marcos-Fa[6], Emma Giarracco[6], Emmanuel Weiss[1,7,8], Jonel Trebicka[1,9], Inmaculada Hernàndez[4,5,10], Javier Fernandez[1,2], Mireia Casulleras[1,2], Cristina López-Vicario[1,2,3], Sinan Muldur[11], Alex Hopke[11], Alexandru Vlagea[2], Ana M. Aransay[3,12], Domenica Marchese[4,5], Mauro Bernardi[13], Rajiv Jalan[1,14], Paolo Angeli[1,15], Giuliana Magri[16], Andrea Cerutti[6,17], Daniel Irimia[11], Holger Heyn[4,5,‡], Vicente Arroyo[1,‡], Richard Moreau[1,7,18,*,‡]

JHEP Reports **2024**. vol. 6 | 1–15

Check for updates

Background & Aims: Patients with acutely decompensated (AD) cirrhosis are immunocompromised and particularly susceptible to infections. This study investigated the immunomodulatory actions of albumin by which this protein may lower the incidence of infections.

Methods: Blood immunophenotyping was performed in 11 patients with AD cirrhosis and 10 healthy volunteers (HV). Bulk and single-cell RNA sequencing (scRNA-seq) and flow cytometry were performed in peripheral blood mononuclear cells (PBMCs) from 20 patients with AD cirrhosis and 34 HV exposed to albumin. Albumin's effects on degranulation, phagocytosis, chemotaxis, and swarming of neutrophils from six patients with AD cirrhosis and nine HV were assessed by measuring myeloperoxidase enzymatic activity, the engulfment of fluorescent-labeled *Escherichia coli* and zymosan, and interactions of neutrophils with *Candida albicans* at single-cell resolution in microfluidic chambers, respectively. Whole blood RNA sequencing (RNA-seq) analyses were performed in 49 patients admitted for severe AD cirrhosis, of whom 30 received albumin during hospitalization.

Results: Compared with HV, patients with AD cirrhosis showed severe lymphopenia and defective neutrophil antimicrobial function. Bulk and scRNA-seq analyses revealed significantly (false discovery rate [FDR] <0.05) increased signatures related to B cells, myeloid cells, and $CD4^+$ T cells in PBMCs incubated with albumin. Changes in the B cell population were confirmed by flow cytometry. Neutrophils exposed to albumin also exhibited augmented chemotactic and degranulation responses, enhanced phagocytosis, and increased pathogen-restrictive swarming. RNA-seq data analysis in patients who had received albumin revealed specific upregulation of signatures related to B cells and neutrophils together with transcriptional changes in $CD4^+$ T cells (FDR <0.05).

Conclusions: The finding that albumin promotes the transcriptional reprogramming and expansion of the B cell compartment and improves neutrophil antimicrobial functions indicates mechanisms that may lower the incidence of infections in patients with severe AD cirrhosis receiving albumin therapy.

© 2024 The Author(s). Published by Elsevier B.V. on behalf of European Association for the Study of the Liver (EASL). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Acute decompensation is the most frequent cause of nonelective hospital admission of patients with cirrhosis, and its most severe form, acute-on-chronic liver failure (ACLF), is characterized by hepatic and/or extrahepatic organ failures and high risk of short-term death.[1,2] These patients exhibit intense systemic inflammation, indicated by leukocytosis,[2,3] elevated blood levels of cytokines, inflammatory lipid mediators, and C-reactive protein (CRP).[1,3–5] These patients also exhibit immunosuppression, indicated by decreased responses of specific monocyte subsets to bacterial products,[6,7] decreased capacity to kill microbes,[6,8,9] and reduced lymphocyte function.[3] Immunosuppression predisposes patients with acutely decompensated (AD) cirrhosis to secondary infections and re-escalation of end-organ dysfunction and mortality.[10] To date, there are no therapies to prevent or treat patients with AD cirrhosis who are immunocompromised or have an impaired immune response. This unmet need is more striking given the ramping prevalence of patients with AD cirrhosis infected with Gram-positive or Gram-negative bacteria resistant to antibiotics.[11]

Albumin therapy is common in patients with AD cirrhosis, including patients treated with large-volume paracentesis and

* Corresponding authors. Address: EF CLIF, Travessera de Gràcia 11, 7th Floor, 08021, Barcelona, Spain.
E-mail addresses: richard.moreau@inserm.fr (R. Moreau), jclaria@clinic.cat (J. Clària).
† Co-first authors.
‡ Co-senior authors.
https://doi.org/10.1016/j.jhepr.2024.101184

patients with spontaneous bacterial peritonitis (SBP) or hepatorenal syndrome–acute kidney injury (HRS-AKI).[12] Recently, the ANSWER study has demonstrated a lower incidence of SBP and other infections in patients with AD cirrhosis receiving long-term weekly administration of albumin.[13] In addition, albumin administration has emerged as an efficacious anti-inflammatory therapy, reducing cytokine levels in patients with AD cirrhosis and ACLF.[14] Although albumin pleiotropic effects have classically been attributed to its oncotic and antioxidant properties,[15] recent data on leukocytes from patients with AD cirrhosis have shown that albumin modulates cytokine production in response to bacterial DNA by interacting with the endosomal TLR9 signaling pathway.[16] The albumin immune restorative effects have also been linked to inactivation of the immunosuppressive effects of PGE_2.[8] However, at present, a comprehensive and deep understanding of the immunomodulatory actions of albumin in the setting of AD cirrhosis is still lacking.

To assess this, we designed *in vitro* experiments and performed bulk and single-cell RNA-sequencing (scRNA-seq) and functional assays in peripheral blood mononuclear cells (PBMCs) and neutrophils from patients with AD cirrhosis and healthy volunteers (HV). Confirmatory *in vivo* whole blood RNA sequencing (RNA-seq) studies were performed in 30 patients with severe AD cirrhosis receiving albumin therapy and in 19 non-albumin-treated patients with AD cirrhosis from the PREDICT study.[17] The results of this investigation revealed that albumin induces specific immune cell gene signatures primarily related to the revitalization of B cell and neutrophil defensive functions in patients with AD cirrhosis. These findings expand our understanding of the mechanisms whereby albumin lowers the rate of infections in patients with cirrhosis and improves survival.

Patients and methods

Patients

For immunophenotyping and *in vitro* experiments in PBMCs, peripheral blood was obtained from 20 patients with AD cirrhosis recruited at the Liver Intensive Care Unit of the Hospital Clínic of Barcelona and 34 HV from the Barcelona Hospital Clínic Blood Bank (Table S1). For neutrophil experiments, peripheral blood was obtained from six patients with AD cirrhosis and nine age-matched HV (Table S2). Whole blood bulk RNA-seq was performed in 10 HV and 49 patients with AD cirrhosis from the PREDICT study,[17] of whom 30 had received albumin therapy (see Supplementary methods for selection criteria). All patients provided written informed consent to participate, and approval was granted by the Ethics Committee of the Hospital Clínic of Barcelona (#HCB/2016/0710, #HCB/2018/0899 and #HCB/2015/0427).

Immunophenotyping and isolation of PBMCs and neutrophils

See Supplementary methods.

PBMC incubations

PBMCs (1.5–3.0 × 10^6 cells/ml) seeded in RPMI 1640 medium were incubated with either human serum albumin (HSA, Albutein®, Grifols, Barcelona, Spain), recombinant albumin expressed in *Oryza sativa* (Sigma-Aldrich, St Louis, MO, USA) (both at 15 mg/ml), or a vehicle control for 24 h at 37 °C in a 5% CO_2 incubator. For comparison, experiments with the isooncotic compound mannitol (Sigma-Aldrich) (15 mg/ml), albumin-depleted FBS (15% total volume), and IgG from human serum (Sigma-Aldrich) (15 mg/ml) were also performed. Albumin-depleted FBS was prepared using the Pierce™ Albumin Depletion Kit, and depletion was verified using the bicinchoninic acid assay. In some experiments, PBMCs were incubated with the neonatal Fc receptor (FcRn) blocking antibody ADM31 (Aldevron, Fargo, ND, USA) (10 μg/ml) before the addition of albumin.

Bulk RNA-seq in whole blood

Isolation of total RNA from whole blood collected in Tempus tubes, assessment of RNA concentration and integrity, library preparation, sequencing, and processing of RNA-seq data are described in the Supplementary methods.

Bulk RNA-seq and real-time PCR in PBMCs

See Supplementary methods.

scRNA-seq in PBMCs

PBMCs (1.5 × 10^6 cells/ml) seeded in RPMI 1640 medium were incubated with either HSA (15 mg/ml) or the vehicle (culture medium) for 2 h at 37 °C in a 5% CO_2 incubator. At the end of the incubation period, cells were rapidly (within 30 min) transferred on ice to the CNAG Single Cell Genomics platform (see Supplementary methods).

Neutrophil degranulation and phagocytosis assays

See Supplementary methods.

Neutrophil chemotaxis and swarming assays

A chemotaxis–phagocytosis assay in microfluidic arenas[18] was used to test the role of HSA on the ability of neutrophils to migrate directionally toward and to phagocytose *Candida albicans* yeast. A swarming assay was used to test the contribution of HSA to the ability of neutrophils to contain the growth of *Candida* clusters. To this purpose, arrays of *Candida*-adherent spots (200-μm diameter) were printed using a microarray printing platform (Picospotter PolyPico, Galway, Ireland) and a solution of poly-l-lysine (Sigma-Aldrich) (see Supplementary methods).

Results

Characterization of the blood immune cell landscape in patients with AD cirrhosis

The peripheral immune cell landscape was characterized by flow cytometry in a representative population of patients with AD cirrhosis in whom HSA administration was indicated as standard of care. Compared with HV, patients with AD cirrhosis showed neutrophilia, severe lymphopenia, and monocytosis (Fig. 1A). Lymphopenia affected $CD8^+$ lymphocytes and natural killer (NK) cells (Fig. 1B). The analysis of $CD45^+CD19^+$ B lymphocytes, which included the study of distinct B cell subsets defined using specific gating strategies (Fig. 1C), also revealed a decline of B and transitional B cells in patients with AD cirrhosis, suggesting impaired B-cell development and maturation (Fig. 1D).

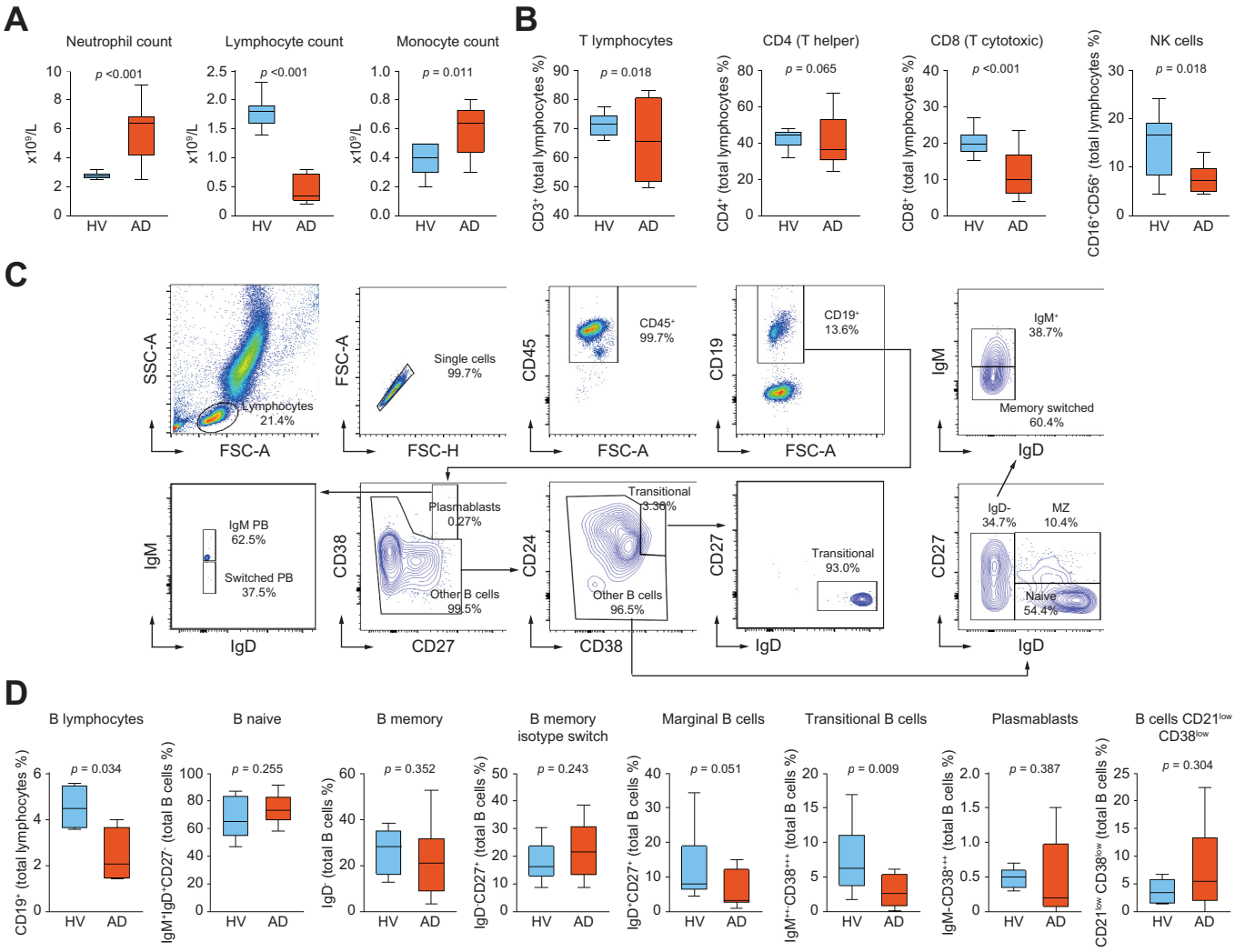


Fig. 1. Peripheral immune cell landscape of patients with AD cirrhosis. (A) Peripheral blood cell counts in 10 HV and 11 patients with AD cirrhosis. (B) Peripheral blood immunophenotyping of T cells, $CD4^+$ and $CD8^+$ T cells, and NK cells in HV and patients with AD cirrhosis. (C) Scatterplots of the flow cytometry analysis showing the gating strategy to identify the different B cell subtypes, including plasmablasts ($CD45^+CD19^+CD38^{++}CD27^{++}$), transitional ($CD45^+CD19^+CD38^{++}CD24^+IgD^+CD27^-$), naïve ($CD45^+CD19^+IgD^+CD27^-CD24^{low}IgM^+$), marginal ($CD45^+CD19^+IgD^+CD27^+CD24^{high}IgM^{++}$), and switched ($CD45^+CD19^+IgD^-CD27^{+/-}IgM^-$). (D) Proportions of B cells and subpopulations in the peripheral blood of HV and patients with AD cirrhosis. Box and whisker graphs in panels A, B, and D represent the median (IQR). Lower and upper box borders indicate the 25th and 75th percentiles, respectively. Lines within each box indicate median percentage. Whiskers above and below each box indicate maximum and minimum values, respectively. Significance between groups was obtained using Wilcoxon–Mann–Whitney tests. AD, acutely decompensated; HV, healthy volunteers; NK, natural killer.

In vitro response of mononuclear leukocytes from patients with AD cirrhosis to HSA

To investigate the direct effect of HSA on immune cells avoiding any confounding factor present in the patient's bloodstream, we first performed experiments *in vitro* in PBMCs isolated from patients with AD cirrhosis. To accurately assess the global transcriptional changes elicited by albumin, we performed bulk RNA-seq and identified differentially expressed (DE) gene sets by gene set enrichment analysis (GSEA). This analysis identified two clusters of gene sets related to B-cell-mediated immunity, B-cell activation, immunoglobulin complexes, immunoglobulin receptor binding, immunoglobulin production, immune response mediated by immunoglobulins, and the B cell receptor signaling pathway that were significantly enriched in PBMCs incubated with HSA (Fig. 2A). Among the gene clusters enriched by HSA, we also identified three gene sets related to the Fc receptor signaling pathway (Fig. 2A). HSA also induced the enrichment of two additional gene set clusters, one related to T cells and DNA and protein complexes and the other to antimicrobial responses, phagocytosis recognition, and defense response against bacteria (Fig. 2A). To exclude the possibility that the effect of HSA on PBMCs could be caused by the presence of stimulatory serum factors bound to its molecule and retained during its manufacturing, we repeated the experiments using recombinant human albumin from *Oryza sativa*. These experiments confirmed that the activation of the B

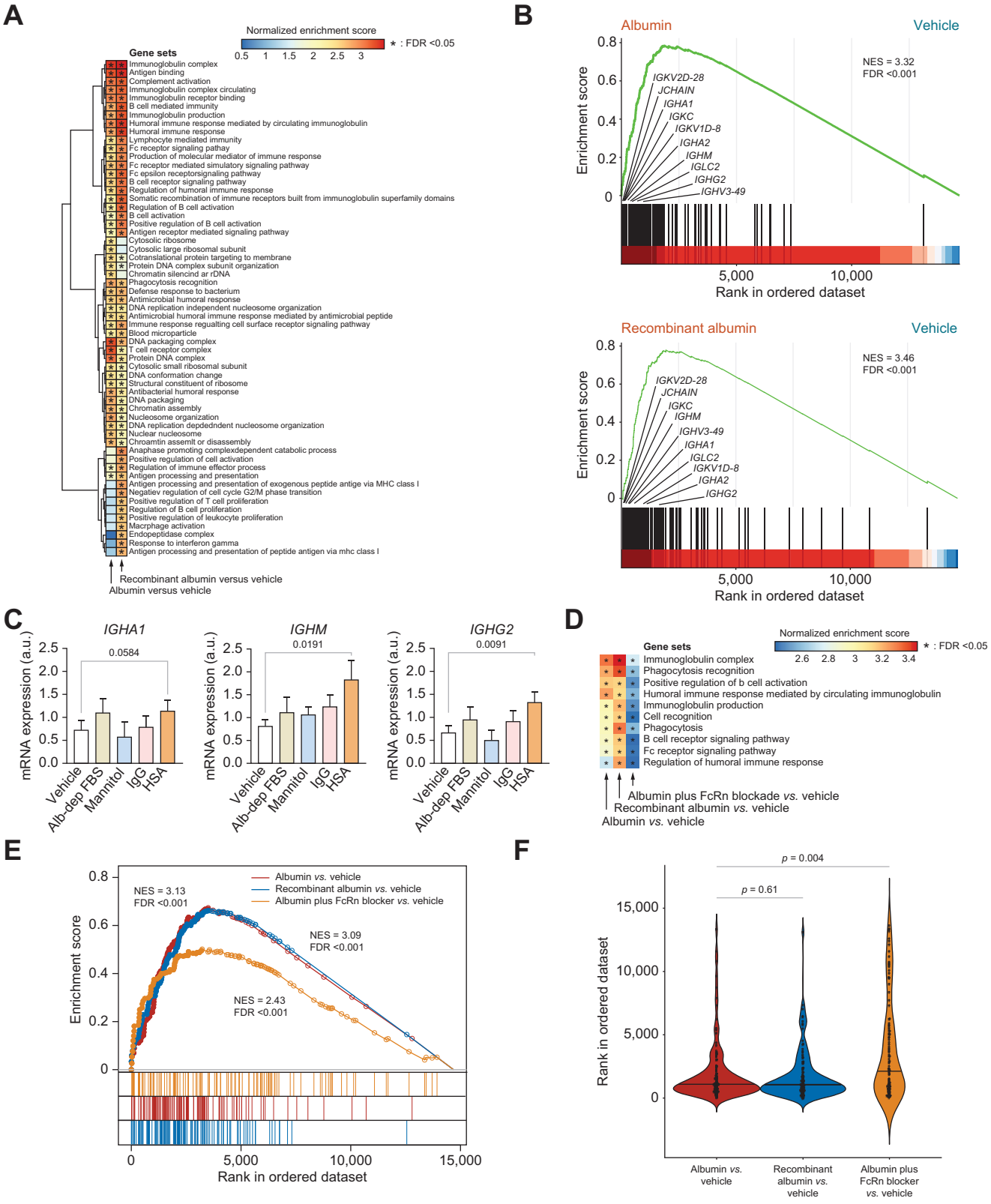


Fig. 2. Changes in the transcriptional landscape of PBMCs isolated from patients with AD cirrhosis and incubated *in vitro* with HSA. (A) GSEA was run on RNA-seq data to generate ranked lists of genes for the two following pairwise comparisons *vs*. vehicle: HSA and recombinant albumin (both at 15 mg/ml). Color gradient corresponds to increasing values of the NES of gene sets from the less (in blue) to the most (in red) upregulated. Asterisks in the heat map indicate FDR <0.05. (B) GSEA enrichment plots of the immunoglobulin complex gene set in two comparisons *vs*. vehicle: HSA (top) and recombinant albumin (bottom). The hash plot under GSEA curves shows where the members of the gene set appear in the ranked list of genes. The genes shown on the plots are representative of leading-edge genes (*i.e.* top-scoring genes). (C) Changes in the expression of three representative leading-edge genes coding for immunoglobulins in response to the vehicle control, 15% albumin-

cell transcriptome by albumin is an intrinsic property of this molecule (Fig. 2A). Similar results were observed in PBMCs isolated from HV, indicating that the effect of albumin on B cells was not exclusive to patients with cirrhosis (Fig. S1). Moreover, enrichment analysis identified several leading-edge genes coding for immunoglobulins increased by both HSA and recombinant albumin (Fig. 2B). To assess whether the effect of HSA on immunoglobulin expression was related to its oncotic properties, we compared the effects of HSA with those of the iso-oncotic control mannitol, IgG, and albumin-depleted FBS and demonstrated that leading-edge immunoglobulin genes were only upregulated by HSA (Fig. 2C).

FcRn is implicated in the B cell response to HSA

Having shown that albumin enhances the expression of gene sets related to the Fc receptor signaling pathway, we explored whether the B cell stimulatory effect of HSA was mediated by its binding to FcRn, an Fc receptor that binds albumin with high affinity.[19] The normalized enrichment score (NES) for the immunoglobulin complex in PBMCs incubated with HSA in the presence of an FcRn-blocking antibody was lower than that in PBMCs incubated with HSA alone (Fig. 2D and E). Consistent with this, immunoglobulin genes showed a less preferred position in the ordered ranked list of genes in PBMCs incubated with HSA in the presence of the FcRn blocker than in PBMCs incubated with HSA alone (Fig. 2F). Although these findings suggest some implication of the FcRn receptor in the transcriptional reconfiguration of B cells elicited by HSA, other receptors or mechanisms are also likely involved in this transcriptional rearrangement.

Profile of the *in vitro* effects of HSA at the scRNA-seq level

We next used scRNA-seq to gain a deeper insight into the multitiered complexity of the cellular composition of PBMCs exposed *in vitro* to HSA for 2 h. We jointly analyzed 66,064 human PBMCs from nine patients with AD cirrhosis clustered into B lymphocytes, T lymphocytes, and myeloid cells. Each lineage was subclustered to define fine-grained cell populations using canonical gene markers (Figs. S2 and S3). We then analyzed 1,946 B lymphocytes and identified 10 principal populations, including two subpopulations that acquired a transitional-like B cell profile characterized by a higher expression of *CD79B*, *IGHD*, and *TCL1A* (transitional 1 B cell) along with a higher expression of *CD55* (transitional 2 B cell) (Fig. 3A). HSA increased the frequency of B cells expressing a transitional-like profile while reducing naive B cells (Fig. 3B and C). To further confirm that HSA induced transcriptional changes resembling those of transitional-like B cells, we assessed the expression of specific genes described by Steward *et al.*[20] This analysis proved the HSA-induced expression of *TCL1A*, *VPREB3*, *PCDH9*, *IGHM*, and *IGHD* genes in transitional B cells (Fig. S4). Importantly, HSA-induced increased frequency of transitional-like B cells observed at the transcriptional level was confirmed by flow cytometry at 24 h (Fig. 3D and E). Interestingly and consistent with the essential role of albumin in *in vitro* B-cell growth,[21] a significant increase in the B cell population was detected by flow cytometry (Fig. 3E). However, albumin did not expand any other B cell subset (Fig. S5A). In addition, no effects on B cells were observed in response to albumin-depleted FBS, mannitol, or IgG (Fig. S5B).

In addition to B cells, we also assessed 21,677 myeloid cells, which clustered into 13 different cell populations of monocytes (n = 7), dendritic cells (n = 3), and myeloid-derived suppressor cells and granulocyte-monocyte progenitors (n = 3) (Fig. 3F). Fig. 3G and H shows that HSA increased the frequency of monocytes with an HAVCR2$^+$ intermediate transcriptional signature and plasmacytoid dendritic cells. The HSA-induced shift in the HAVCR2$^+$ intermediate monocyte population was unlikely attributed to the upregulation of surface proteins resulting from cell attachment, as no changes were observed in the expression of genes associated with monocyte activation (*i.e. FCGR3A*, *CD80*, *CD86*, *ICAM1*, *ITGAM*, *CCR2*, and *HLA*)[22] (Fig. S6A). HSA did not change the frequency of other myeloid cell populations (Fig. S6B).

We also analyzed 17,467 T lymphocytes, including 12,692 CD4$^+$ T cells, 3,483 CD8$^+$ T cells, and 1,292 unconventional T cells. Interestingly, HSA modified the frequency distribution of the 10 populations captured within the CD4$^+$ T cell compartment (Fig. 3I–K and Fig. S7). Specifically, HSA decreased the frequency of activated memory CD4$^+$ T cells while increasing central memory ITGB1$^+$ CD4$^+$ T cells (Fig. 3K). Finally, no significant changes were found upon HSA exposure within the 13 populations of CD8$^+$ T cells and unconventional T cells (Fig. S8). Together, these observations confirmed at the single-cell level the ability of HSA to reprogram the transcriptional landscape of B cells and other mononuclear leukocytes isolated from patients with AD cirrhosis.

In vitro response of polymorphonuclear leukocytes (neutrophils) from patients with AD cirrhosis to HSA

To have a more comprehensive view of the effects of HSA on blood immune cells, we designed *in vitro* experiments in neutrophils from patients with AD cirrhosis and HV. Given that neutrophils express fewer genes than any other leukocyte and that their lower RNA content complicates high-throughput transcriptomic analysis,[23] we tested the effects of HSA on neutrophil function, including degranulation, phagocytosis, chemotaxis, and swarming. As expected, neutrophils from

depleted FBS (Alb-dep FBS), mannitol (15 mg/ml), IgG (15 mg/ml), and HSA (15 mg/ml). Significant differences between groups were assessed using *t* tests. (D) Heat map of the NES for the 10 representative B cell–related gene sets obtained for each of the following three comparisons *vs.* vehicle: HSA, recombinant albumin, and FcRn blocker (10 μg/ml) plus HSA *vs*. vehicle. (E) Enrichment plots of the immunoglobulin complex in the three GSEA comparisons described in D. The leading-edge genes are indicated by solid dots. The hash plots under GSEA curves show where the members of the gene set appear in each of the three ranked lists of genes. (F) Rank order of each gene of the immunoglobulin complex gene set. The higher in the rank, the lower the importance. Values of *p* are ranked from Kruskal–Wallis tests, followed by Mann–Whitney *U* tests. All FDR values are computed with adjustments for multiple testing and gene set size in the GSEA analysis. AD, acutely decompensated; FcRn, neonatal Fc receptor; FDR, false discovery rate; GSEA, gene set enrichment analysis; HSA, human serum albumin; HV, healthy volunteers; NES, normalized enrichment score; PBMC, peripheral blood mononuclear cell; RNA-seq, RNA sequencing.

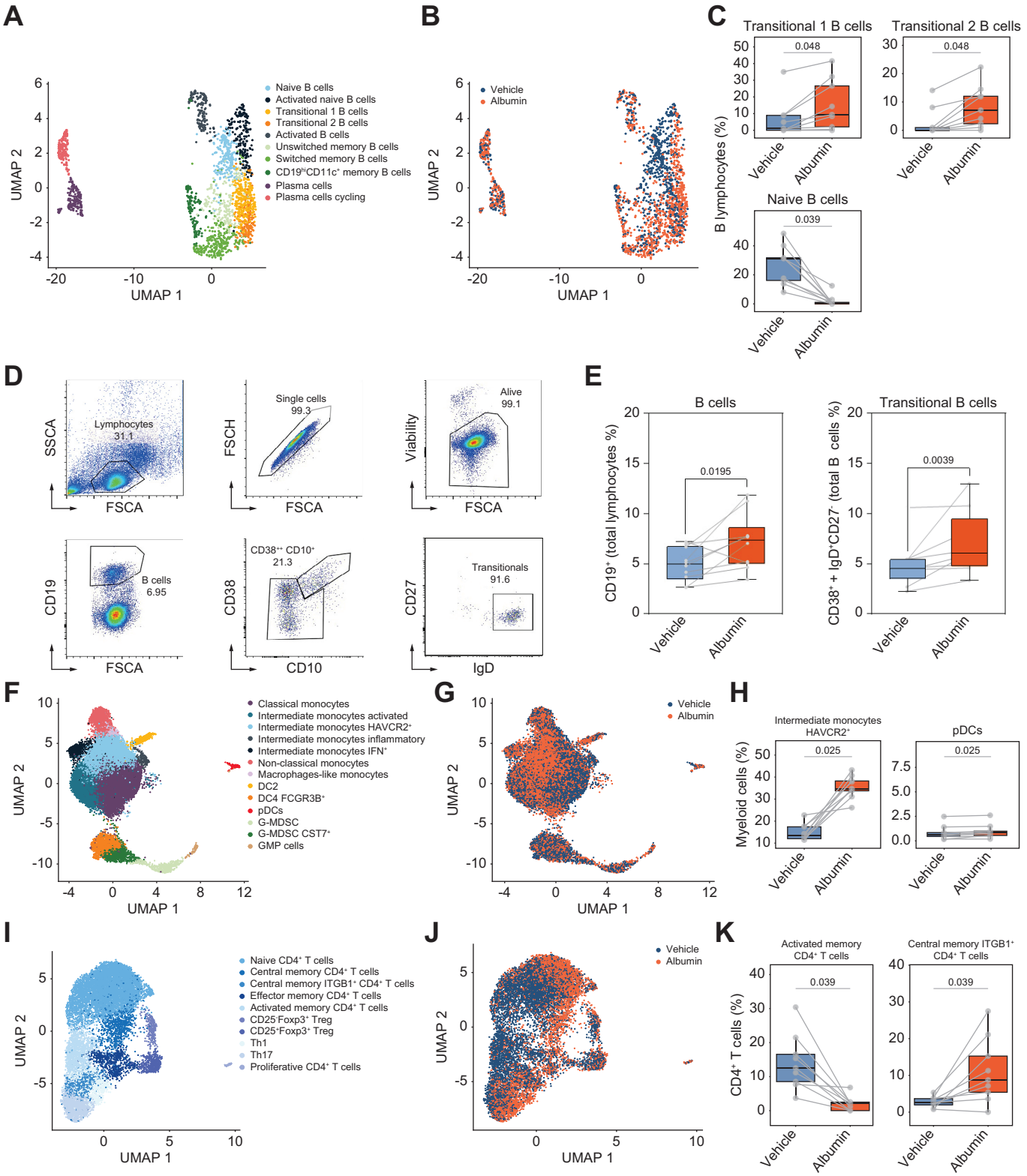


Fig. 3. scRNA-seq and flow cytometry analyses identify specific immune cell changes in PBMCs exposed *in vitro* to HSA. (A) UMAP of 1,946 patients' B cells exposed to HSA and the vehicle, colored by cell types. (B) Overlay of HSA and vehicle exposure on the B lymphocyte UMAP. (C) Box plots for the proportion of B cell populations that significantly changed after HSA exposure. (D, E) Scatterplots of the flow cytometry analysis showing the gating strategy to identify B and transitional B cells and boxplots showing their proportions after HSA exposure. (F) UMAP of 21,819 patients' myeloid cells exposed to HSA and the vehicle, colored by cell populations. (G) Overlay of HSA and vehicle exposure on the myeloid cell UMAP. (H) Box plots for the proportion of myeloid cell types that significantly changed after

patients with AD cirrhosis showed impaired antimicrobial functions, including impaired phagocytosis, which is consistent with previous studies,[9] as well as reduced neutrophil chemotaxis and swarming (Fig. S9). Incubation of neutrophils from both patients with AD cirrhosis and HV with either HSA or recombinant human albumin increased degranulation, as assessed by myeloperoxidase (MPO) activity in the supernatant (Fig. 4A), an essential antimicrobial protein localized mainly to the azurophil or primary neutrophil granules.[24] Incubation of neutrophils from both patients with AD cirrhosis and HV with either HSA or recombinant human albumin also increased phagocytosis, as monitored by the ingestion of fluorescent-labeled zymosan particles by neutrophils (Fig. 4B). The phagocytosis findings were confirmed using fluorescent-labeled *Escherichia coli* (Fig. S10). We finally tested the effect of HSA on neutrophil swarming, a defensive process in which neutrophils undergo phases of highly directed and coordinated migration, followed by accumulation at sites of infection that culminate in the containment and killing of the intruding pathogen. Here, we used microfluidic devices that enable monitoring at single-cell resolution the neutrophil interactions with *Candida* through morphology changes from yeast to hyphae and hyphae growth.[18] We also tested the effect of HSA on neutrophil phagocytosis of live *Candida*. Neutrophils from patients displayed severe defects in their ability to control fungal growth, as measured by the time it took for *Candida* hyphae to escape the neutrophil swarm and the higher total area covered by fungal growth (Fig. S9). The addition of HSA improved the ability of neutrophils to restrict fungal growth in a concentration-dependent manner (higher albumin leading to greater restriction) (Fig. 4C). Imaging showed that in the presence of HSA, neutrophils swarmed around any *Candida* hyphae clusters, phagocytosing them and delaying their growth (Fig. 4D and Supplementary video). In vehicle controls, neutrophils delayed the growth of *Candida* but were not efficient enough against clusters of *Candida* hyphae. These findings reveal the ability of HSA to improve the defensive functions of neutrophils isolated from immunocompromised patients with AD cirrhosis.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.jhepr.2024.101184

Translational study in patients with AD cirrhosis receiving HSA as therapy

The study included 49 patients with AD cirrhosis at imminent risk of ACLF at entry (T1) who developed ACLF during the index hospitalization (T2) (Fig. S11A). Of the 49 patients, 30 received HSA between T1 and T2 (albumin group), and 19 remained free of HSA (non-albumin group). We assumed that investigating patients exhibiting the most severe forms of AD cirrhosis with the use of longitudinal data would limit the effect of interindividual variability on the assessment of HSA effects. The clinical characteristics and laboratory data of all patients at T1 and T2 are given in Table S3. Characteristics were similar in terms of standard laboratory values, including clinical differential blood counts for neutrophils, monocytes, and lymphocytes (Fig. S12A), which strongly correlated with RNA-seq-inferred blood counts (Fig. S12B). As expected, patients presented organ failures, higher white blood cell counts, and elevated levels of inflammatory markers at T2 compared with HV (Tables S3 and S4).

We next considered characteristics at T1 and T2 in patients in the albumin and non-albumin groups. In both groups, progression to ACLF from T1 to T2 was associated with significant increases in the model for end-stage liver disease (MELD) score. However, except for serum creatinine, which increased at T2, there were no significant within-group changes regarding the rest of longitudinally collected standard laboratory data and circulating levels of inflammatory mediators (Table 1). However, there were some in between-group differences. The delay between T1 and T2 was shorter in the albumin group than in the non-albumin group. Compared with patients in the non-albumin group, those in the albumin group had significantly higher CRP, IL-6, and IL-10 levels at T1 and higher Chronic Liver Failure Consortium organ failure (CLIF-C OF) score and creatinine, CRP, IL-6, and MCP-1 levels at T2. Moreover, a higher percentage of patients in the albumin group died by 28 and 90 days. The difference in survival was expected considering the higher prevalence of SBP and HRS-AKI in the albumin group.

Transcriptional characteristics of all patients at T1 and T2

Analyzing differentially expressed genes (DEGs) in two comparisons, T1 *vs*. HV and T2 *vs*. HV, we observed, as expected, strong similarities between the two disease stages. Thus, the number of DEGs was high in the two comparisons; that is, there were 4,615 DEGs in T1 relative to HV and 4,529 in T2 relative to HV. Of the 4,615 DEGs associated with T1, 3,929 (85%) overlapped with the DEGs assigned to T2 (Fig. S11B, top). By comparing effect-size changes in T1 *vs*. HV with T2 *vs*. HV, we observed a highly concordant magnitude of changes between the two signatures (Fig. S11B, bottom). Next, we used Quantitative Set Analysis for Gene Expression (QuSAGE) to analyze the differential expression of blood transcription modules (BTMs). Among the 258 annotated BTMs, the total number of DE BTMs (false discovery rate [FDR] <0.05) was 190 in T1 *vs*. HV and 186 in T2 *vs*. HV (Fig. S11C, top; Table S5A). These findings indicated extensive changes in the blood transcription module space in both T1 and T2. Strikingly, only 20 BTMs had specific differential expression relative to HV, in either T1 (12 modules) or T2 (eight modules), whereas 179 DE BTMs were shared with a concordant sign, of which 108 were upregulated

HSA exposure. (I) UMAP of 12,692 patients' $CD4^+$ T cells exposed to HSA and the vehicle, colored by cell types. (J) Overlay of HSA and vehicle exposure on $CD4^+$ T cell UMAP. (K) Box plots for the proportion of $CD4^+$ T cells that significantly changed after HSA exposure. Fig. 3A–C and F–K have been designed using scRNA-seq in PBMCs from nine patients with AD cirrhosis. (D) and (E) were designed based on results obtained by flow cytometry in PBMCs from 10 age-matched HV. Significant differences between groups were assessed using paired *t* tests. AD, acutely decompensated; HSA, human serum albumin; HV, healthy volunteers; PBMC, peripheral blood mononuclear cell; pDC, plasmacytoid dendritic cell; scRNA-seq, single-cell RNA sequencing; UMAP, Uniform Manifold Approximation and Projection.

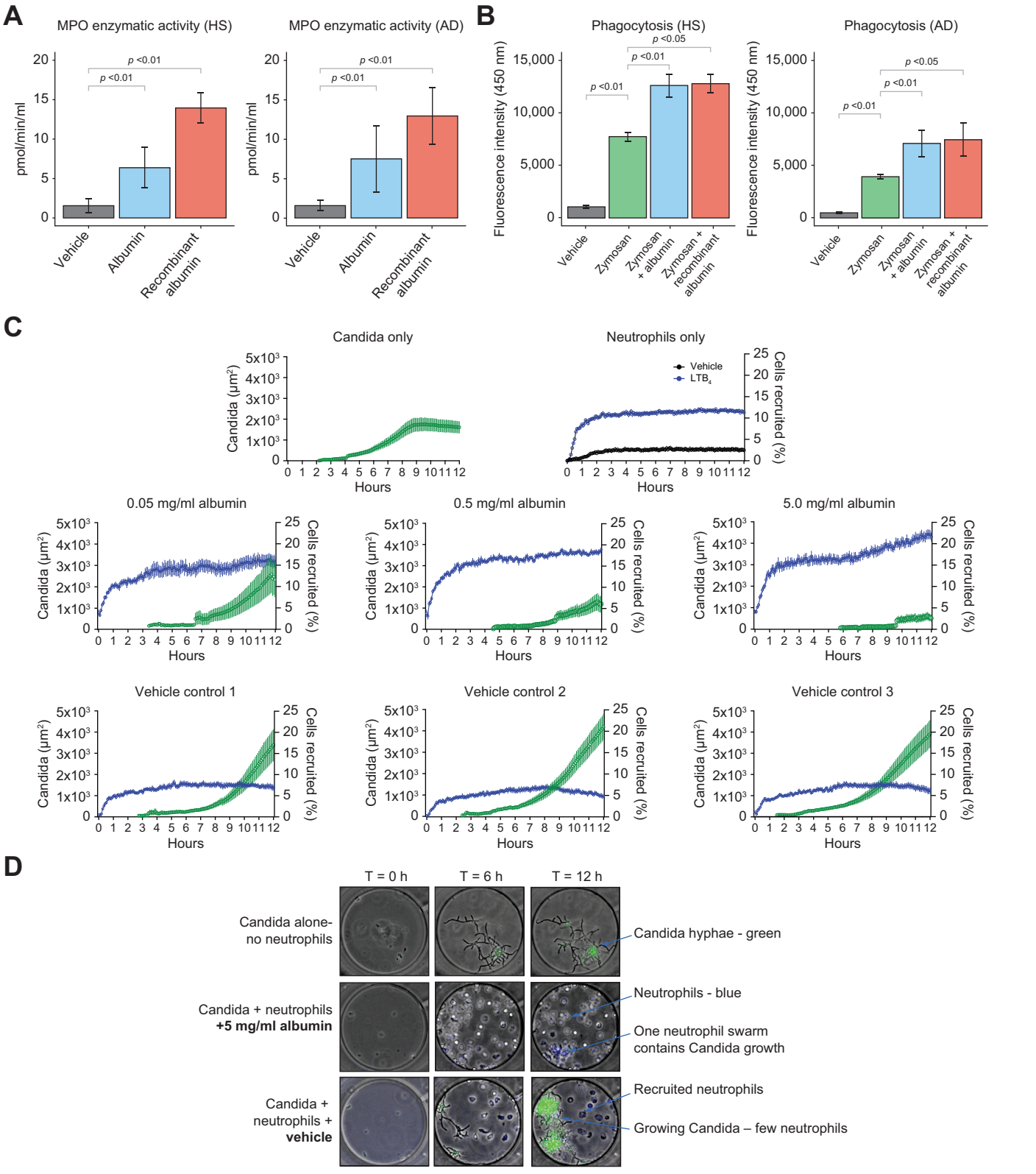


Fig. 4. Effects of HSA on the host defense function of neutrophils from patients with AD cirrhosis. (A) Neutrophils were incubated with cell medium (vehicle), HSA, or recombinant human albumin (both at 15 mg/ml) for 2 h at 37 °C in a 5% CO_2 incubator. Neutrophil degranulation was assessed by measuring the MPO enzymatic activity in the cell supernatants. (B) Phagocytic capacity assessed by incubating neutrophils with FITC-conjugated zymosan bioparticles alone or in the presence of HSA and recombinant albumin for 60 min and compared with vehicle control. (C) Quantification of *Candida albicans* growth and neutrophil recruitment in the microfluidic device. The size of *Candida* hyphae clusters was quantified based on the area of green fluorescence and is shown in green. The number of neutrophils recruited was quantified and shown in blue. For the neutrophils-only condition, we compared the neutrophils entering the chambers in the presence of the vehicle

and 71 downregulated (Table S5A). Fig. S11C (bottom) shows the 20 top gene modules for shared upregulated BTMs and 20 top gene modules for shared downregulated BTMs. Shared upregulated BTMs were related to innate immunity, including those related to Toll-like receptors and inflammatory signaling, interferon-alpha response, and innate immune cells (*i.e.* neutrophils, monocytes, and dendritic cells). Shared downregulated BTMs were related to T and NK cells and antigen presentation. These results were consistent with those obtained when analyzing gene signatures for the 10 main immune cell types (Fig. S13A) and 29 fine immune cell types (Fig. S13B). Collectively, our analyses of bulk blood RNA-seq data provided consistent results that highlighted the similarity of the transcriptional landscape in T1 and T2, both disease's stages being characterized by simultaneity between increases in gene signatures related to innate immunity and decreases in gene signatures related to adaptive immunity.

Whole blood gene signatures associated with HSA treatment

We compared differential gene expression between T2 and T1 for the albumin and non-albumin groups and found that gene signatures differed between the two groups. Indeed, 269 DEGs (92 upregulated, 177 downregulated) were specific for the albumin group, whereas 103 DEGs (64 upregulated, 39 downregulated) were specific for the non-albumin group; only 36 DEGs with a concordant sign were shared by the two groups (Fig. 5A). Identification of DEGs using volcano plots (Fig. 5B) illustrated the group specificity of both upregulated and downregulated genes. A broad variety of immunoglobulin genes (see below) and several major neutrophil genes (*CD177, OLFM4, PRG2, MPO, BPI, RETN, LCN2, CEACAM8, MCEMP1*) were upregulated in the albumin group (Fig. 5B, top), but not in the non-albumin group (Fig. 5B, bottom). These findings agree with the observed *in vitro* responses of B cells and neutrophils to albumin. Because analysis of DEGs drew our attention to immunoglobulin genes, we compared the effect-size changes between T2 and T1 within each group for each of the 125 genes (including 120 immunoglobulin genes) that are included in the Gene Ontology gene set labeled 'GOCC Immunoglobulin Complex'. We found that effect-size changes were greater in the albumin group (33 upregulated immunoglobulin genes) than in the non-albumin group (only 6) (Fig. 5C). Gene coding for constant regions of immunoglobulin heavy chains (*IGHM*, *IGHG2*, *IGHG3*, *IGHG4*, and *IGHA2*) were specifically upregulated in the albumin group. These findings are consistent with the results of SingleR analysis, which showed increases in gene signatures for plasmablasts that were specific for the albumin group (Fig. 5D, top). Immunoglobulin genes that were specifically upregulated in the albumin group also included genes for constant regions of immunoglobulin light chains kappa (*IGKC*) and lambda (*IGLC1*), genes for the V regions of the variable domain of immunoglobulin heavy chains, and genes for the V regions of the variable domain of light chains kappa and lambda. The three genes that were specifically upregulated in the non-albumin group were genes coding for the V regions of the variable domain of immunoglobulin heavy chains (Fig. 5C). Together these findings indicate an extensive upregulation of gene coding for immunoglobulins that were specific for patients who had received HSA.

Next, we applied QuSAGE to identify DE BTMs between T2 and T1 within each group. We first observed that the number of DE BTMs was greater in the albumin group than in the non-albumin group (36 *vs*. 6, respectively; Table S5B). Only one DE BTM, related to endoplasmic reticulum, was specifically upregulated in the non-albumin group (Fig. 5E). In sharp contrast, 31 DE BTMs (13 downregulated, 18 upregulated) were specific for the albumin group (Fig. 5E). Downregulated modules were related to erythropoiesis, cytoskeleton, or cell junction (Fig. 5E), whereas upregulated BTMs were related to activated dendritic cells, complement and other receptors in dendritic cells, cytokines, and chemokines (Fig. 5E). Upregulated BTMs specific for the albumin group also comprised modules related to B cells (including those related to enriched in B cells, plasma cells, and immunoglobulins), mismatch repair, cell cycle, and mitosis. Of note, upregulation of BTMs related to B cells in the albumin group was confirmed after adjusting the transcriptomics data by disease severity, as estimated by the MELD score (Fig. S14). Moreover, a direct and significant positive correlation was observed between the mean daily HSA dose administered to the patients and the expansion of the B cell compartment (Fig. S15). However, no evidence for HSA induction of NF-κB, a transcription factor involved in B-cell development and survival,[25,26] was observed in our study (Fig. S16). Of note, RNA-seq-inferred signatures for T and NK cells (Fig. 5D, middle and bottom panels; and Fig. S17A and B) as well as gene modules related to these cells (Table S5B), whose downregulation is a hallmark of patients with AD cirrhosis without and with ACLF (Fig. S11C), remained all downregulated among patients treated with HSA. Nevertheless, we observed that a BTM labeled 'mitotic cell cycle in stimulated $CD4^+$ T cells (M4.11)' was specifically upregulated among patients treated with HSA (Fig. 5E), which suggests some activation of transcription in $CD4^+$ T cells. Consistent with this, we found no overlap between member genes of the M4.11 BTM and genes used for SingleR analysis of $CD4^+$ T cells (Fig. S18). These findings were validated by scRNA-seq, in which the gene signature score for this BTM across all $CD4^+$ T cells computed using the Ucell package[27] showed a significantly higher cell density in HSA-treated cells relative to the vehicle (Fig. S19).

(black) with the number of neutrophils in the presence of LTB_4 chemoattractant (100 nM; blue). Twelve chambers per condition were quantified, and three of them are shown. (D) Microscopy images of neutrophil–*Candida albicans* interactions in microfluidic chambers. A similar number of *Candida* yeast was loaded in each chamber (time T-0). Neutrophils were loaded outside the chambers. Neutrophils migrated to the chambers, attracted by *Candida*-released molecules. In the presence of HSA, neutrophils actively phagocytosed *Candida*. Neutrophils swarmed around any *Candida* hyphae clusters, delaying their growth. In vehicle controls, neutrophils delayed the growth of *Candida* but were not efficient enough to contain clusters of *Candida* hyphae. (A)–(D) were designed using functional assays in freshly isolated peripheral neutrophils from six patients with AD cirrhosis and five age-matched HV. Significant differences between groups were assessed using paired *t* tests. AD, acutely decompensated; HSA, human serum albumin; HV, healthy volunteers; LTB_4, leukotriene B_4; MPO, myeloperoxidase.

Table 1. Characteristics of patients at time 1 (T1) and time 2 (T2) and deaths in the albumin and non-albumin groups.

	Albumin group (n = 30)		Non-albumin group (n = 19)		*p* value			
Characteristics	**T1**	**T2**	**T1**	**T2**	**T2 *vs.* T1 (albumin group)**	**T2 *vs.* T1 (non-albumin group)**	**T1 (between groups)**	**T2 (between groups)**
Markers of organ function								
MELD score, mean ± SD	19.2 ± 5.3	27.0 ± 6.6	19.0 ± 5.7	24.0 ± 7.5	<0.01	0.01	0.92	0.17
CLIF-C OF score, mean ± SD	7.0 ± 1.0	10.5 ± 2.6	7.4 ± 1.3	8.9 ± 1.5	<0.01	<0.01	0.26	0.01
Organ system failure, n (%)								
Liver failure	6 (20)	7 (23)	4 (21)	6 (32)	1	0.71	1	0.76
Kidney failure	0 (0)	23 (77)	0 (0)	13 (68)	<0.01	<0.01	1	0.76
Circulatory failure	0 (0)	10 (33)	1 (5)	1 (5)	<0.01	1	1	0.69
Cerebral failure	0 (0)	6 (20)	0 (0)	0 (0)	<0.01	1	0.82	0.31
Coagulation failure	0 (0)	4 (13)	0 (0)	4 (22)	0.12	0.10	1	0.01
Respiratory failure	0 (0)	9 (30)	0 (0)	0 (0)	<0.01	1	1	0.02
ACLF grade, no. (%)								
1	0 (0)	13 (43)	0 (0)	13 (72)	<0.01	<0.01	1	0.1
2–3	0 (0)	17 (57)	0 (0)	5 (28)	<0.01	0.05	1	0.1
Precipitating events, n (%)								
Infection as precipitant at T1 or T2	14 (47)	20 (67)	5 (26)	7 (37)	0.19	0.73	0.5	0.03
Alcohol-related hepatitis as precipitant at T1 or T2	14 (50)	14 (47)	7 (39)	7 (37)	1	1	0.4	0.21
Laboratory data								
International normalized ratio, median (IQR)	1.6 (1.4–1.7)	1.6 (1.5–2.0)	1.5 (1.2–1.8)	1.4 (1.3–2.2)	0.08	0.48	0.7	0.27
Total bilirubin (mg/L), median (IQR)	3.7 (2.0–9.0)	3.6 (2.4–11.3)	3.2 (1.3–9.5)	5.0 (1.0–13.5)	0.83	0.99	0.72	0.52
Serum creatinine (mg/dl), median (IQR)	1.2 (1.0–1.6)	2.7 (1.9–3.4)	1.4 (1.1–1.6)	2.1 (1.4–2.4)	<0.01	0.01	0.59	0.02
Serum sodium (mmol/L), mean ± SD	131 ± 5.8	133 ± -8.4	134.0 ± 6.9	132 ± 6.9	0.31	0.53	0.18	0.74
Serum albumin (g/dl), median (IQR)	2.6 (2.2–3.2)	3.0 (2.3–3.5)	2.7 (2.5–3.4)	3.0 (2.8–3.1)	0.21	0.36	0.23	0.89
White cell count ($\times 10^3/mm^3$), median (IQR)	8.3 (7.0–10.3)	11.9 (6.2–13.7)	6.6 (4.6–8.6)	8.1 (5.8–12.4)	0.06	0.12	0.14	0.31
Absolute lymphocyte count ($\times 10^3/mm^3$), median (IQR)	0.9 (0.6–1.3)	1.1 (0.7–1.6)	1.1 (0.8–1.5)	1.2 (0.8–1.8)	0.25	0.54	0.23	0.48
Absolute monocyte count ($\times 10^3/mm^3$), median (IQR)	0.8 (0.6–1.1)	1.0 (0.6–1.3)	0.6 (0.3–1.0)	0.7 (0.6–0.9)	0.1	0.48	0.28	0.08
Absolute neutrophil count ($\times 10^3/mm^3$), median (IQR)	6.1 (4.2–7.7)	8.9 (4.2–11.8)	4.0 (3.4–6.4)	4.5 (3.0–7.8)	0.12	0.73	0.15	0.05
C-reactive protein (mg/L), median (IQR)	31.3 (19.7–54.0)	33.5 (19.4–90.4)	13.3 (7.4–25.0)	16.1 (11.4–21.8)	0.5	0.69	<0.01	0.01
Blood levels of protein mediators of inflammation (pg/ml), median (IQR)								
Eotaxin	66.0 (43.9–97.1)	86.3 (58.6–104.1)	71.0 (43.7–121.0)	64.3 (50.1–121.9)	0.1	1	0.47	0.53
Granulocyte-colony stimulating factor	18.8 (4.4–43.2)	22.5 (9.0–73.8)	23.9 (7.6–99.6)	16.9 (3.6–70.9)	0.59	0.75	0.63	0.61
IFN-α2	10.0 (2.2–24.8)	16.8 (2.4–33.4)	12.8 (7.6–22.0)	17.6 (8.2–22.7)	0.43	0.5	0.68	0.97
IFN-γ	33.3 (7.0–86.0)	23.8 (12.1–97.2)	19.2 (7.2–98.7)	25.9 (7.9–45.8)	0.93	0.84	0.76	0.57
IL-1α	4.3 (1.7–7.2)	2.0 (0.8–6.1)	2.5 (1.1–3.3)	2.0 (0.3–3.9)	0.3	0.71	0.12	0.45
IL-1β	5.4 (3.1–10.8)	4.7 (2.3–13.0)	5.2 (1.8–9.7)	5.2 (1.9–8.7)	0.93	0.89	0.71	0.64
IL-6	24.7 (13.6–46.5)	52.5 (24.3–168.9)	9.3 (6.8–18.1)	16.9 (8.8–23.1)	0.16	0.1	<0.01	<0.01
IL-8	5.5 (3.5–10.7)	12.6 (4.0–18.1)	4.3 (2.6–13.0)	4.3 (1.0–9.0)	0.19	0.6	0.44	0.03

IL-10	10.0 (3.9–30.4)	13.1 (3.8–26.4)	2.4 (1.2–3.6)	4.7 (3.2–5.8)	0.97	0.08	0.03	0.15
Interferon-inducible protein 10	231 (123–374)	231 (169–615)	214 (166–482)	237 (141–487)	0.34	1	0.69	0.71
Monocyte chemoattractant protein 1	179 (147–257)	349 (230–533)	176 (158–279)	198.5 (100–281)	<0.01	0.8	0.74	<0.01
Macrophage inflammatory protein 1α	11.6 (8.4–23.5)	15.2 (9–28.9)	8.5 (2–18)	13.3 (4.5–19.4)	0.38	0.41	0.21	0.26
Macrophage inflammatory protein 1β	15.7 (11.7–19.7)	19.9 (13.8–30.6)	15.9 (12–22.7)	15.2 (11.7–19.6)	0.08	0.7	0.56	0.15
Tumor necrosis factor	30.8 (18.1–42)	35.3 (24–66.6)	37.6 (18.8–47.2)	35.6 (20.8–49.3)	0.31	0.89	0.59	0.62
Vascular endothelial growth factor	4.3 (0.8–8.2)	6.3 (2.6–16.4)	5.2 (1.9–12.7)	7.6 (4.7–11.7)	0.17	0.25	0.57	0.53

Normal values for blood inflammatory mediators are shown in Table S4. There was no significant difference between the albumin and non-albumin groups in terms of age (mean ± SD, 61.0 ± 7.9 and 59.1 ± 10.8 years, respectively) and number (%) of males (21 [70] and 12 [63]; respectively). A higher percentage of patients in the albumin group than in the no-albumin group died by 28 days (30% [9/30] and 0%, respectively) and 90 days (67% [20/30] and 16% [3/19], respectively).
ACLF, acute-on-chronic liver failure; CLIF-C OF, Chronic Liver Failure Consortium organ failure; MELD, model for end-stage liver disease.

Discussion

The results of the current investigation contribute to understanding the mechanisms by which albumin administration is associated with a lower rate of infections and better survival in patients with AD cirrhosis. The investigation was performed *in vitro* in peripheral leukocytes and *in vivo* in patients with AD cirrhosis receiving albumin therapy. The investigation used cutting-edge technologies (*i.e.* bulk and scRNA-seq and swarming assays in microfluidic arenas) and gold-standard functional assays to assess leukocyte antimicrobial functions. The results of this investigation revealed that albumin modulates the peripheral immune system by attenuating B-cell depletion and revitalizing neutrophil functions in patients with AD cirrhosis.

To our knowledge, this is the first study to use longitudinal transcriptomics to characterize the effects of albumin on immune cells in patients with AD cirrhosis. The design of the study, particularly in those aspects related to the selection of patients, was performed after considering that patients hospitalized with ACLF, who are known to present an extremely dynamic clinical course with marked changes in the magnitude of systemic inflammation,[2,17] were clearly poor candidates for inclusion. Other phenotypes, including 'unstable' and 'stable' AD cirrhosis, were also not considered because they are associated with moderate systemic inflammation, which may further decrease during hospitalization, and we were interested in including only patients with intense systemic inflammation to better assess immune changes induced by albumin. For this reason, we chose patients with AD cirrhosis with imminent risk of developing ACLF, in whom the magnitude of systemic inflammation was similar to that seen in ACLF. We included only patients who received albumin for well-established indications (paracentesis, prevention of HRS-AKI associated with SBP, or HRS-AKI treatment)[12] even though this would introduce the bias of higher severity in patients receiving albumin. We did not select patients receiving albumin within 1 month before T1, those with a too-long interval between the two assessments of bulk blood RNA-seq, and those with a delay of >10 days between the last dose of albumin and T2. Finally, although albumin dosage and duration of treatment varied according to indications, we decided to analyze data in all patients irrespective of albumin dosage. The similarity of transcriptional characteristics as well as most clinical characteristics, standard laboratory values, and levels of cytokines in patients with AD cirrhosis before and after developing ACLF confirmed our assumption that patients included in the study were in a relatively steady state of systemic inflammation during the study.

The major finding of the scRNA-seq and bulk RNA-seq in blood cells from patients with AD cirrhosis was that albumin administration triggered signals that caused an expansion of the B cell compartments, and likely the $CD4^+$ T cell compartments, while activating some mononuclear myeloid cells. These findings are extremely relevant, considering that the prevalence of bacterial infections at admission in patients with AD cirrhosis is very high (37.3 and 25.1% in patients with and without ACLF, respectively).[10,28] Among the uninfected patients at admission, 46% with and 18% without ACLF develop bacterial infection during hospitalization. The existence of a profound impairment of the innate immune system, which is highlighted by impaired neutrophil and monocyte antimicrobial functions and lymphocyte depletion, likely explains such a high

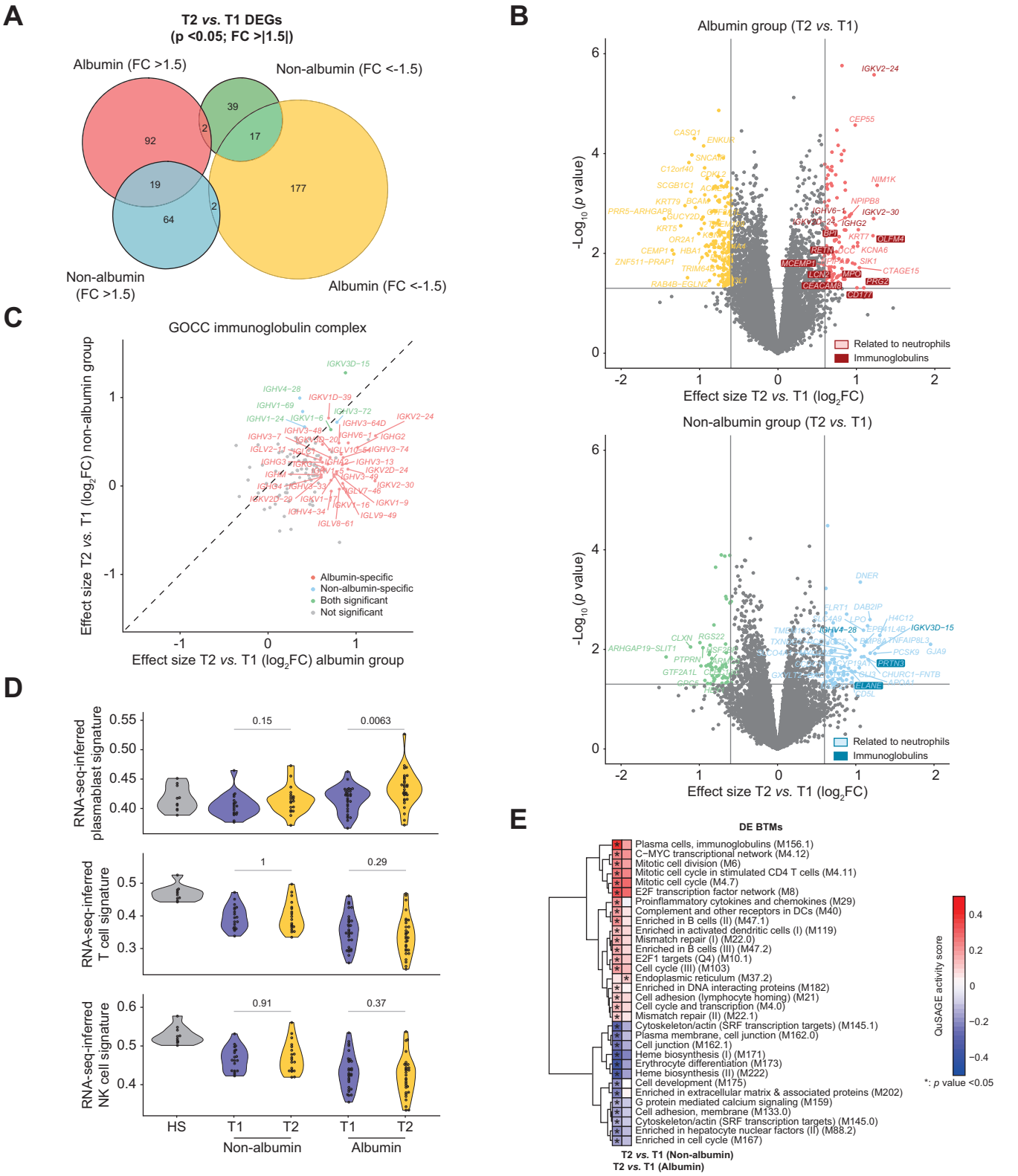


Fig. 5. Whole blood gene signatures associated with HSA treatment. (A) Euler plot showing DEGs between T2 and T1 for the albumin and non-albumin groups. *DEGs were defined by an* absolute FC greater than 1.5 and p <0.05. (B) Volcano plots showing differential expression effect size ($\log_2$ FC) between T2 and T1 plotted against significance ($-\log_{10} p$) for the albumin (top) and non-albumin (bottom) groups. In both volcano plots, gray points indicate genes with no significant difference in expression between T2 and T1 (with absolute FC <1.5 and p >0.05, i.e. $-\log_{10} p$ <1.3). Colored points indicate DEGs, either upregulated or downregulated. The neutrophil-related genes are highlighted. (C) Differential expression effect size ($\log_2$ FC) between T2 and T1 in the albumin group compared with corresponding differential effect size in the non-albumin group. Only genes included in the Gene Ontology gene set 'GOCC_Immunoglobulin Complex' are shown here. DEGs with a concordant sign were considered as shared. (D) Violin plots of RNA-seq inferred signatures at T1 and T2 for plasmablasts, NK cells, and T cells, determined using the

risk of infections in these patients.[29] Of note, because changes in immune-cell transcriptome seen after albumin administration occurred in the absence of changes in albuminemia, the effects of albumin on the transcriptome of immune cells cannot be merely explained by correction of hypoalbuminemia but rather by qualitative differences between the 'commercial' albumin and the circulating endogenous albumin in patients with AD cirrhosis, which is known to be post-translationally modified and highly oxidized.[30]

Some mechanisms could explain the effect of albumin on immune cells. Immune cells can be activated by osmotic stress.[31] Therefore, the movement of water from cells to the extracellular milieu (osmotic stress) could be a mechanism by which albumin exerted immunomodulatory effects in our *in vitro* cell experiments. However, we did not see the same response in PBMCs incubated with mannitol, a compound sharing the iso-oncotic properties of albumin. Another mechanism could be the induction of NF-κB, a master transcription factor in B cells[25] that has been reported to be activated by albumin in a human renal proximal tubule-derived cell line.[26] However, in our study, we did not find any sign of NF-κB induction in response to albumin. Because extracellular acidification is known to reprogram immune cell responsiveness,[32] another possibility is that a change in pH caused by albumin, which is a highly soluble acidic molecule, could account for the activation of immune cells. This can also occur at the intracellular level because albumin is rapidly internalized by mononuclear myeloid cells[16] and intracellular acidification results in changes in immune cell performance.[33] In contrast, the interaction of albumin with FcRn or with voltage-gated proton channels, which leads to stimulation of phagocytosis and degranulation,[19,34,35] respectively, could mechanistically explain the activation of neutrophil antimicrobial function by albumin. In this regard, we observed that the immunostimulatory actions of albumin in mononuclear leukocytes were partially inhibited by an antibody blocking FcRn.

This study has some limitations. First, there is a difference in severity between the two study groups. Second, our results are mostly descriptive, which makes it difficult to directly prove immune restoration by albumin. The third limitation is related to the study design. Indeed, we leveraged the prespecified collection of blood for RNA-seq in patients of the PREDICT study at admission (when they did not have ACLF) and at the time of ACLF development to identify patients who did or did not receive albumin during the progression to ACLF. Thus, we do not report the results of a controlled study specifically designed to investigate the effects of prespecified systematic albumin administration, for example, on the development of ACLF. The ideal protocol would consist in daily administration of albumin with the objective of, for example, maintaining blood levels of albumin above 3.0 g/dl, as in the ATTIRE trial.[36] Nevertheless, it is noteworthy that conducting a prospective randomized controlled trial of albumin for preventing ACLF in patients who present with pre-ACLF would face major methodological issues. Indeed, there are currently no criteria to identify patients with pre-ACLF among those who present for AD cirrhosis.[17] Therefore, performing randomized controlled trials that enroll well-matched patients with AD cirrhosis is a big challenge, and criteria for stratifying patients according to the risk of ACLF is an unmet medical need.

In conclusion, the results of this study suggest that albumin promotes the expansion of the B cell compartment and improves neutrophil antimicrobial functions. These findings contribute to understanding the clinical benefits of albumin therapy such as reducing the incidence rate of infections in patients with AD cirrhosis.[13]

Affiliations

[1]*European Foundation for the Study of Chronic Liver Failure (EF CLIF), European Association for the Study of the Liver (EASL)–CLIF Consortium, and Grifols Chair, Barcelona, Spain;* [2]*Hospital Clínic-IDIBAPS, Barcelona, Spain;* [3]*CIBER of Hepatic and Digestive Diseases (CIBERehd); Barcelona, Spain;* [4]*Universitat de Barcelona, Barcelona, Spain;* [5]*Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain;* [6]*Translational Clinical Research Program, Institute "Hospital del Mar" for Medical Investigations (IMIM), Barcelona, Spain;* [7]*Institut National de la Santé et de la Recherche Médicale (INSERM), Université Paris-Cité, Centre de Recherche sur l'Inflammation (CRI), Paris, France;* [8]*Assistance Publique–Hôpitaux de Paris (AP-HP), Hôpital Beaujon, Département d'Anesthésie et de Réanimation, DMU Parabol, Clichy, France;* [9]*Department of Internal Medicine B, University of Münster, Münster, Germany;* [10]*Institute for Research in Biomedicine (IRB), Barcelona, Spain;* [11]*Center for Engineering in Medicine and Surgery, Massachusetts General Hospital, Shriners Burns Hospital, Harvard Medical School, Boston, MA, USA;* [12]*Center for Cooperative Research in Biosciences (CIC bioGUNE), Parque Tecnológico de Bizkaia, Derio, Spain;* [13]*Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy;* [14]*Liver Failure Group, Institute for Liver Disease Health, University College London, Royal Free Hospital, London, UK;* [15]*Unit of Internal Medicine and Hepatology (UIMH), Department of Medicine – DIMED, University of Padova, Padova, Italy;* [16]*Immunology Unit, Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain;* [17]*Catalan Institute for Research and Advanced Studies (ICREA); Barcelona, Spain;* [18]*Assistance Publique – Hôpitaux de Paris (AP-HP), Hôpital Beaujon, Service d'Hépatologie, Clichy, France*

Abbreviations

ACLF, acute-on-chronic liver failure; AD, acutely decompensated; BTM, blood transcription module; DE, differentially expressed; DEG, differentially expressed genes; FC, fold change; FcRn, neonatal Fc receptor; FDR, false discovery rate; GSEA, gene set enrichment analysis; HRS-AKI, hepatorenal syndrome–acute kidney injury; HSA, human serum albumin; HV, healthy volunteers; MELD, model for end-stage liver disease; MPO, myeloperoxidase; NES, normalized enrichment score; NK, natural killer; PBMC, peripheral blood mononuclear cell; QuSAGE, Quantitative Set Analysis for Gene Expression; RNA-seq, RNA sequencing; SBP, spontaneous bacterial peritonitis; scRNA-seq, single-cell RNA sequencing; UMAP, Uniform Manifold Approximation and Projection.

SingleR software, in the albumin and non-albumin groups. Baseline values in HV are also shown. Values of *p* are from Kruskal–Wallis tests. (E) Heat map of 32 DE BTMs between T2 and T1 that were specific for either the albumin or non-albumin group. Thirty-one DE BTMs were specific for the albumin group, whereas only one DE BTM ('endoplasmic reticulum [M37.2]') was specific for the non-albumin group. BTMs were hierarchically clustered based on QuSAGE activity scores obtained in the albumin group. Asterisks denote $p < 0.05$ obtained when comparing probability density functions with QuSAGE. Color represents QuSAGE activity score. BTM, blood transcription module; DE, differentially expressed; DEG, differentially expressed gene; FC, fold change; HSA, human serum albumin; NK, natural killer; QuSAGE, Quantitative Set Analysis for Gene Expression; RNA-seq, RNA sequencing; T1, time 1; T2, time 2.

Financial support

This research was supported by EF CLIF, a private, non-profit organization receiving unrestricted grants from Grifols and Fundació Privada Cellex. EF CLIF is a partner and/or coordinator of several European Union (EU) Horizon 2020 program projects (Nos. 825694 and 847949). This study was also supported by Ministerio de Ciencia e Innovacion/Agencia Estatal Investigación MCIN/AEI, PID2022-138970OB-I00 10.13039/501100011033/FEDER, UE (to JC) and RTI2018-093894-B-I00 (to AC). LJ-G has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Science, Innovation and Universities. EW was an EF CLIF Research Fellow who received grants from Fundació Privada Cellex and the Fondation Bettencourt-Schueller. IH received a Marie Skłodowska-Curie grant agreement (No. 75451) from the EU's Horizon 2020 Research and Innovation Program. The funders had no influence on the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest

MB is part of the speakers' bureau for Grifols SA, Octapharma AG, Shire/Takeda, CLS Behring GmbH, and PPTA and is a consultant for Grifols SA, CLS Behring GmbH, Martin Pharmaceuticals, and Shire/Takeda. HH is a co-founder and shareholder of Omniscope, a scientific advisory board member of MiRXES, and a consultant to Moderna. RJ has research collaborations with Yaqrit and Takeda. He is the inventor of OPA, which has been patented by UCL and licensed to Mallinckrodt Pharma. He is also the founder of Yaqrit Ltd, a spin-out company from University College London and Thoeris Ltd. JCN is a scientific consultant of Omniscope. The remaining authors disclose no conflicts.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Study concept and design: VA, RM. Acquisition of data and sample measurements: MC, CLV, BRG, XM, AV, AMA, EW, JT, JF, MB, RJ, PA. Bioinformatic and statistical analyses: JJL, IH, FA, LJG, JCN, AMA, DM. Drafting of the manuscript: JC, VA, RM. Critical revision of the manuscript for important intellectual content: EW, JC, JJL, IH, JT, FA, JF, LJG, JCN, AV, AMA, MB, RJ, PA, GM, AC, HH. Study supervision: RM, VA, HH.

Data availability statement

All data associated with this work is presented in the main manuscript or Supplementary material. For availability of any other type of data, contact the corresponding authors.

Acknowledgements

We thank Dr Anna Bosch, Dr Lidia García-Campmany, Ms Anna Curto, Ms Cristina Sanchez, and Mr Alex Amoros for their help.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhepr.2024.101184.

References

Author names in bold designate shared co-first authorship

[1] Moreau R, Jalan R, Ginès P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology 2013;144:1426–1437.
[2] Arroyo V, Moreau R, Jalan R. Acute-on-chronic liver failure. N Engl J Med 2020;382:2137–2145.
[3] **Weiss E**, **de la Grange P**, Defaye M, et al. Characterization of blood immune cells in patients with decompensated cirrhosis including ACLF. Front Immunol 2021;11:619039.
[4] **Clària J**, **Stauber RE**, Coenraad MJ, et al. Systemic inflammation in decompensated cirrhosis: characterization and role in acute-on-chronic liver failure. Hepatology 2016;64:1249–1264.
[5] López-Vicario C, Checa A, Urdangarin A, et al. Targeted lipidomics reveals extensive changes in circulating lipid mediators in patients with acutely decompensated cirrhosis. J Hepatol 2020;73:817–828.
[6] Bernsmeier C, Pop OT, Singanayagam A, et al. Patients with acute-on-chronic liver failure have increased numbers of regulatory immune cells expressing the receptor tyrosine kinase MERTK. Gastroenterology 2015;148:603–615.
[7] Korf H, du Plessis J, van Pelt J, et al. Inhibition of glutamine synthetase in monocytes from patients with acute-on-chronic liver failure resuscitates their antibacterial and inflammatory capacity. Gut 2019;68:1872–1883.
[8] O'Brien AJ, Fullerton JN, Massey KA, et al. Immunosuppression in acutely decompensated cirrhosis is mediated by prostaglandin E_2. Nat Med 2014;20:518–523.
[9] Stadlbauer V, Mookerjee RP, Wright GA, et al. Role of Toll-like receptors 2, 4, and 9 in mediating neutrophil dysfunction in alcoholic hepatitis. Am J Physiol Gastrointest Liver Physiol 2009;296:G15–G22.
[10] Fernández J, Acevedo J, Wiest R, et al. Bacterial and fungal infections in acute-on-chronic liver failure: prevalence, characteristics and impact on prognosis. Gut 2018;67:1870–1880.
[11] **Fernández J**, **Piano S**, **Bartoletti M**, et al. Management of bacterial and fungal infections in cirrhosis: the MDRO challenge. J Hepatol 2021;75:101–117.
[12] European Association for the Study of the Liver. EASL Clinical Practice Guidelines for the management of patients with decompensated cirrhosis. J Hepatol 2018;69:406–460.
[13] Caraceni P, Riggio O, Angeli P, et al. Long-term albumin administration in decompensated cirrhosis: an open-label randomised trial. Lancet 2018;391:2417–2429.
[14] **Fernández J**, **Clària J**, Amorós A, et al. Effects of albumin treatment on systemic and portal hemodynamics and systemic inflammation in patients with decompensated cirrhosis. Gastroenterology 2019;157:149–162.
[15] Bihari S, Bannard-Smith J, Bellomo R. Albumin as a drug: its biological effects beyond volume expansion. Crit Care Resusc 2020;22:257–265.
[16] Casulleras M, Flores-Costa R, Duran-Güell M, et al. Albumin internalizes and inhibits endosomal TLR signaling in leukocytes from patients with decompensated cirrhosis. Sci Transl Med 2020;12:eaax5135.
[17] Trebicka J, Fernandez J, Papp M, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. J Hepatol 2020;73:842–854.
[18] Ellett F, Jalali F, Marand AL, et al. Microfluidic arenas for war games between neutrophils and microbes. Lab Chip 2019;19:1205–1216.
[19] Pyzik M, Rath T, Lencer WI, et al. FcRn: the architect behind the immune and nonimmune functions of IgG and albumin. J Immunol 2015;194:4595–4603.
[20] Stewart A, Ng JC, Wallis G, et al. Single-cell transcriptomic analyses define distinct peripheral B cell subsets and discrete development pathways. Front Immunol 2021;12:602539.
[21] Polet H, Spieker-Polet H. Serum albumin is essential for in vitro growth of activated human lymphocytes. J Exp Med 1975;142:949–959.
[22] Kapellos TS, Bonaguro L, Gemünd I, et al. Human monocyte subsets and phenotypes in major chronic inflammatory diseases. Front Immunol 2019;10:2035.
[23] Grieshaber-Bouyer R, Radtke FA, Cunin P, et al. The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. Nat Commun 2021;12:2856.
[24] Amulic B, Cazalet GL, Hayes KD, et al. Neutrophil function: from mechanisms to disease. Annu Rev Immunol 2012;30:459–489.
[25] Sasaki Y, Iwai K. Roles of the NF-κB pathway in B-lymphocyte biology. Curr Top Microbiol Immunol 2016;393:177–209.
[26] Drumm K, Bauer B, Freudinger R, et al. Albumin induces NF-κB expression in human proximal tubule-derived cells (IHKE-1). Cell Physiol Biochem 2002;12:187–196.
[27] Andreatta M, Carmona SJ. UCell: robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J 2021;19:3796–3798.
[28] Trebicka J, Fernandez J, Papp M, et al. PREDICT identifies precipitating events associated with the clinical course of acutely decompensated cirrhosis. J Hepatol 2021;74:1097–1108.
[29] Arroyo V, Moreau R, Kamath PS, et al. Acute-on-chronic liver failure in cirrhosis. Nat Rev Dis Primers 2016;2:16041.
[30] Alcaraz-Quiles J, Casulleras M, Oettl K, et al. Oxidized albumin triggers a cytokine storm in leukocytes through P38 mitogen-activated protein kinase: role in systemic inflammation in decompensated cirrhosis. Hepatology 2018;68:1937–1952.
[31] Ip WK, Medzhitov R. Macrophages monitor tissue osmolarity and induce inflammatory response through NLRP3 and NLRC4 inflammasome activation. Nat Commun 2015;6:6931.
[32] Jiang W, Le J, Wang PY, et al. Extracellular acidity reprograms macrophage metabolism and innate responsiveness. J Immunol 2021;206:3021–3031.
[33] Murase M, Kawasaki T, Hakozaki R, et al. Intravesicular acidification regulates lipopolysaccharide inflammation and tolerance through TLR4 trafficking. J Immunol 2018;200:2798–2808.
[34] Vidarsson G, Stemerding AM, Stapleton NM, et al. FcRn: an IgG receptor on phagocytes with a novel role in phagocytosis. Blood 2006;108:3573–3579.

[35] Zhao R, Dai H, Arias RJ, et al. Direct activation of the proton channel by albumin leads to human sperm capacitation and sustained release of inflammatory mediators by neutrophils. Nat Commun 2021;12:3855.

[36] China L, Freemantle N, Forrest E, et al. A randomized trial of albumin infusions in hospitalized patients with cirrhosis. N Engl J Med 2021;384:808–817.

Keywords: Multiorgan failure; Inflammation; Immunosuppression; Gene expression.
Received 27 December 2023; received in revised form 26 July 2024; accepted 31 July 2024; Available online 8 August 2024

SUPPLEMENTARY MATERIAL

Albumin Reprograms the B cell Transcriptional Landscape and Improves Neutrophil Antimicrobial Function in Patients with Decompensated Cirrhosis.

Joan Clària, Ferran Aguilar, Juan-José Lozano, Laura Jiménez-Gracia, Juan C. Nieto, Berta Romero-Grimaldo, Xavi Marcos-Fa, Emma Giarracco, Emmanuel Weiss, Jonel Trebicka, Inmaculada Hernàndez, Javier Fernandez, Mireia Casulleras, Cristina López-Vicario, Sinan Muldur, Alex Hopke, Alexandru Vlagea, Ana M. Aransay, Domenica Marchese, Mauro Bernardi, Rajiv Jalan, Paolo Angeli, Giuliana Magri, Andrea Cerutti, Daniel Irimia, Holger Heyn, Vicente Arroyo, Richard Moreau

SUPPLEMENTARY METHODS

Selection of Patients from the PREDICT Study

The PREDICT study was a large-scale prospective observational multicenter investigation aimed to identify precipitants and short-term (3-month) clinical course phenotypes in patients hospitalized for acutely decompensated (AD) cirrhosis without acute-on-chronic liver failure (ACLF) (*1*). AD cirrhosis was defined by the acute development of ascites, encephalopathy and/or gastrointestinal hemorrhage (*1*), according to previous studies (*2,3*). AD cirrhosis develops in the context of intense systemic inflammation (*2-5*). ACLF is the extreme expression of AD cirrhosis and is characterized by extrahepatic organ failure(s) and high short-term mortality (*2,3*). The PREDICT study consisted of a prespecified longitudinal collection of clinical and standard laboratory data and biological samples, including whole-blood for RNA-seq. Investigators used an electronic clinical-report form to report clinical data, including treatments and standard laboratory data (*1*). Data were continuously monitored on-line by the Data Management Center of the European Foundation for the Study of Chronic Liver Failure. Among patients of the PREDICT study, three distinct clinical phenotypes were identified according to the clinical course. Approximately 20% of the patients had pre-ACLF at enrollment because these patients developed ACLF within 3 months. It was the group with the highest grade of systemic inflammation at enrollment. The other two clinical phenotypes (unstable and stable decompensated cirrhosis) included patients with lower grade of systemic inflammation, which may rapidly improve under standard medical therapy (*1*). Unfortunately, among patients admitted for AD cirrhosis without ACLF, there are currently no available markers enabling to distinguish those who will develop ACLF from those who will not develop this complication (1). Therefore, it is currently not possible to design studies, interventional, in which the inclusion criterion would be the presence of AD cirrhosis at imminent risk of ACLF. For this reason, prospective observational studies are of most interest in the assessment of treatments in these patients.

The selection of patients from PREDICT followed this criteria: i) all AD patients were at imminent risk of ACLF at enrollment (T1) and developed ACLF during the index hospitalization (T2); ii) in all patients, bulk blood RNA-seq data were obtained at T1 and T2; iii) None of the forty-nine patients had received albumin before T1; iv) Thirty out of the forty-nine patients had received albumin between T1 and T2 and composed the albumin group whereas the nineteen others remained free of albumin between T1 and T2 and composed the non-albumin group. In the albumin group, the median time

between T1 and T2 for blood collections for RNA-seq was 11.5 days (interquartile range [IQR], 6.2-18.7). The median duration of HSA treatment was 2.5 days (IQR, 2-6) and the dose was 40.0 g per day (IQR, 22.5-57.5). The time between the last HSA administration and subsequent blood collection was 1.5 days (IQR, 0.0-2.7), indicating that we explored patients soon after HSA administration. In the non-albumin group, the median time between T1 and T2 was 29 days (IQR, 16-52). The 60 first patients with AD cirrhosis and imminent risk of ACLF enrolled in the PREDICT study were eligible. Of these, 11 (9 who had received albumin prior baseline transcriptomics and 2 showing poor-quality control in the transcriptomic analysis) were excluded. The *in vivo* study was therefore performed in 49 patients. The electronic case-reports form of the PREDICT study prespecified the following indications for intravenous albumin: large-volume therapeutic paracentesis, spontaneous bacterial peritonitis, hepatorenal syndrome-acute kidney injury, and other indication. Among the 30 patients who had received intravenous albumin between enrollment and ACLF development and were analyzed in the present study (**Figure S11A**), the primary indications for albumin were large-volume therapeutic paracentesis, spontaneous bacterial peritonitis, and hepatorenal syndrome-acute kidney injury (HRS-AKI, formerly type 1 hepatorenal syndrome).

Immunophenotyping

Extended phenotyping of blood immune cells was performed by flow cytometry using the following panels of antibodies (clones are indicated within brackets): TBNK panel including CD4 (SK3), CD56 (MY31), CD8 (RPA-T8), CD16 (3G8), CD14, CD45 (2D1), CD3 (SP34-2) and CD19 (SJ25C1); B cell panel including CD21 (B-ly4), IgD (IA6-2), CD24 (ML5), CD38 (HIT2), CD27 (M-T271), CD45 (2D1), IgM (G20-127) and CD19 (SJ25C1); dendritic cell panel including CD3 (SK7), CD16 (3G8), CD19 (SJ25C1), CD20 (L27), CD14 (MφP9), CD56 (NCAM16.2), CD1c (F10), CD123 (7G3), CD11c (B-ly6), CD34 (8G12) + CD117 (YB5.B8), CD45 (2D1), HLA-DR (L243) and CD141 (1A4); and T regulatory cell panel including HLA-DR (L243), CD25 (M-A251), CD127 (HIL-7R-M21), CD45RO (UCHL1), CD3 (SK7), CD4 (L200) and CCR4 (1G1). All antibodies were purchased at BD Biosciences (San Jose, CA). For immunophenotyping, the antibodies were titrated and used to stain 100 µL of EDTA-collected whole blood during 20 minutes in the dark. Then erythrocytes were lysed by incubating 20 minutes with BD FACS™ lysing solution and washed twice. After lysis, the cells were resuspended in phosphate-buffered saline. After compensation, data acquisition was performed with a BD FACSCanto II flow cytometer (Becton Dickinson, Franklin Lakes, NJ). For analysis BD FACSDiva (Becton Dickinson) software was used. A minimum of 100,000 total events were acquired. The graphical layout was created using the FlowJo software (Becton Dickinson). Immunophenotyping and isolation of PBMCs

To assess changes in B cell phenotypes induced by albumin, PBMCs from ten healthy subjects were isolated and seeded at a density of $3x10^6$ cells/ml in RPMI 1640 medium with BAFF (Enzo Life Sciences, Farmingdale, NY) (10 ng/ml) as a survival factor, with or without albumin (15 mg/ml) and incubated at 37ºC, 5% CO_2 for 24 hours. At the end of the incubation period, cells were centrifuged at 400 g for 5 min and resuspended in cold DPBS. Then, cells were incubated for 30 min with the LIVE/DEAD yellow viability marker (L34959, Thermo Fisher Scientific), and washed with PBS. Thereafter, cells were stained in MACS buffer (pH 7.4) for 20 min with the following anti-human antibodies. After washing with PBS, the cells were fixed with paraformaldehyde at 4% for 12 min and then washed again with PBS. Finally, cells were resuspended in 250 µl of cold PBS. The analysis was performed on a Cytek Aurora spectral flow cytometer (Cytek Biosciences, Fremont, CA).

Antigen	Isotype	Conjugated dye	Reference	Company
CD10	Mouse IgG1, κ	BV421	312217	Biolegend (San Diego, CA)
CD19	Mouse IgG1, κ	Pacific Blue	302224	Biolegend
CD24	Mouse IgG2a, κ	PE-Cy7	311120	Biolegend
CD27	Mouse IgG1, κ	BV650	302827	Biolegend
CD38	Mouse IgG1, κ	APC-Fire810	CUSTOM	Biolegend
IgD	Goat F(ab')2 IgG	FITC	2032-02	Southern Biotech (Birmingham, AL)

PBMC and neutrophil isolation

Peripheral venous blood (20 ml) was obtained by venipuncture and collected into EDTA-coated, sterile, pyrogen-free tubes (Becton Dickinson). Blood samples were centrifuged at 200 g for 10 min and sedimented cells were diluted with DPBS without calcium and magnesium ($DPBS^{--}$) up to a volume of 20 ml. Diluted blood was layered over 13.3 ml of Ficoll-Hypaque and centrifuged at 500 g for 25 min with the break-off. PBMCs obtained from the mononuclear cell layer were then incubated with pre-warmed ammonium-chloride-potassium lysis buffer for 10 min at room temperature to remove red blood cells and then centrifuged at 400 g for 5 min. The resultant pellet was washed with $DPBS^{--}$. Isolated PBMCs were re-suspended in RPMI 1640 medium containing penicillin (100 U/ml), streptomycin (100 U/ml), and L-glutamine (4 mM) without fetal bovine serum (FBS) and viable cells were then automatically counted. After 30 minutes of resting, $3x10^6$ viable cells were incubated at 37°C and used for bulk and scRNA-seq. Neutrophils were also isolated using the Ficoll-Hypaque method described above. Briefly, after separation of the PBMC layer in the supernatant, neutrophils were collected from the pellet and incubated with pre-warmed ammonium-chloride-potassium lysis buffer for 10 min at room temperature to remove red blood cells and then centrifuged at 400 g for 5 min. The red blood lysis procedure was repeated twice, and the resultant pellet was washed with $DPBS^{--}$. Isolated neutrophils were resuspended in RPMI 1640 medium containing penicillin, streptomycin, and L-glutamine without FBS and used in the assays. For the chemotaxis and swarming assays, blood samples from healthy donors were purchased from RBC components (Alston, MA) and neutrophils were isolated from the blood using the STEMCELL direct neutrophil isolation kit (Vancouver, Canada), according to the manufacturer's protocols. Isolated neutrophils were counted and re-suspended in Iscove's Modified Dulbecco's Medium (IMDM).

scRNA-seq in PBMCs

PBMCs were seeded in RPMI 1640 medium at a density of $1.5x10^6$ cells/ml and incubated with either human serum albumin (Albutein®) (15 mg/ml) or vehicle (culture medium) for two hours at 37ºC in a 5% CO_2 incubator. At the end of the incubation period, cells were rapidly (within 30 min) transferred on ice to the Single Cell Genomics platform from the CNAG. PBMCs were centrifuged for 5 min at 500 g at 4°C. The supernatant was removed, and the pellet was re-suspended in 1xPBS supplemented with 0.05% BSA. Samples were filtered with a 40-µm strainer (pluriSelect) and cell concentration and viability were verified by counting with a TC20™ Automated Cell Counter (Bio-Rad Laboratories, Hercules, CA). PBMC samples from two patients and two healthy subjects were directly used for cell

partition into Gel Bead in Emulsion with a Target Cell Recovery of 5,000 cells each. For all the other cases, donor matched albumin or vehicle samples were labeled by cell hashing and pooled in one 10x Genomics channel to avoid technical biases. Cell hashing was performed following the manufacturer's instructions (Cell hashing and Single Cell Proteogenomics Protocol Using TotalSeq™ Antibodies; BioLegend, San Diego, CA). Briefly, for each sample, 1-2 million cells were re-suspended in 100 µl of Cell Staining Buffer (BioLegend) and incubated for 10 min at 4ºC with 5 µl of Human TruStain FcX™ Fc Blocking reagent (BioLegend). Next, a specific TotalSeq-A antibody-oligo conjugate (anti-human Hashtag 1-4, Biolegend) was added to each of the two conditions (albumin and vehicle) and incubated on ice for 30 min. Cells were then washed three times with cold PBS-0.05% BSA and centrifuged for 5 min at 500 g at 4°C. Finally, cells were re-suspended in an appropriate volume of 1xPBS-0.05% BSA to obtain a final cell concentration 500-1000 cells/µl, suitable for 10x Genomics scRNA-seq. An equal volume of hashed cell suspension from each of the two conditions was mixed and filtered with a 40-µm strainer. Cell concentration was verified by counting with a TC20™ Automated Cell Counter. Then, cells were partitioned into Gel Bead In Emulsions with a Target Cell Recovery of 10,000 total cells. Sequencing libraries were prepared using the v3.0 single-cell 3' mRNA kit (10x Genomics, Pleasanton, CA) with some adaptations for cell hashing, as indicated in TotalSeq™-A Antibodies and Cell Hashing with 10x Single Cell 3' Reagent Kit v3 Protocol from BioLegend. Briefly, 1 µl of 0.2 µM hashtag oligonucleotide (HTO) primer (GTGACTGGAGTTCAGACGTGTGC*T*C; *phosphorothioate bond) was added to the cDNA amplification reaction to amplify the hashtag oligos together with the full-length cDNAs. A SPRI selection clean-up was performed to separate mRNA-derived cDNA (>300 bp) from antibody-oligo-derived cDNA (<180 bp), as described in the above-mentioned protocol. 10x Genomics cDNA libraries were prepared following the 10x Genomics Single Cell 3' mRNA kit protocol, while HTO cDNAs were indexed by PCR as follows. Briefly, 5 µl of purified hashtag oligo cDNA were mixed with 2.5 µl of 10 µM Illumina TruSeq D70X_s primer (IDT) carrying a different i7 index for each sample, 2.5 µl of SI primer from 10x Genomics Single Cell 3' mRNA kit, 50 µl of 2xKAPA HiFi PCR Master Mix (KAPA Biosystem) and 40 µl of nuclease-free water. The reaction was carried out using the following thermal cycling conditions: 98°C for 2 min (initial denaturation), 12 cycles of 98°C for 20 sec, 64°C for 30 sec, 72°C for 20 sec, and a final extension at 72°C for 5 min. The HTO libraries were purified with 1.2X SPRI bead selection. Size distribution and concentration of cDNA and HTO libraries were verified on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies, Santa Clara, CA). Finally, sequencing of HTO and cDNA libraries was carried out on a NovaSeq6000 system (Illumina Inc., San Diego, CA) to obtain approximately >25,000 reads per cell.

Data processing of scRNA-seq.

To profile the cellular transcriptome, we processed the sequencing reads using the CellRanger software package (version 6.1.1) from 10X Genomics Inc. We mapped the reads against the human GRCh38 reference genome. For the libraries where multiple samples were pooled together, we also specified the HTO associated to each sample. All analyses presented in this manuscript were conducted using R version 4.0.5, along with specific analysis and data visualization packages such as Seurat R package (version 4.0.0) (*6*) SeuratObject package (version 4.0.1), and other packages specified in the subsequent sections. All plots were generated using ggplot2 package (version 3.3.3). For sample demultiplexing, we demultiplexed cell hashtags as described in Stoeckius et al. (*7*) for each library separately. Briefly, we normalized HTO counts using a centered log ratio and run "HTODemux()" function from Seurat package with the default parameters. Each barcode was classified to its sample-of-origin, and multiplets (barcodes assigned to more than one condition) or negatives (barcodes not assigned to any condition) were discarded for downstream analysis. For

quality control (QC) and cell annotation, we ensured that there were no remarkable differences on the main QC metrics (library size, library complexity, percentage of mitochondrial and ribosomal expression) among the different hashed and non-hashed libraries, and then performed common QC, normalization and further analysis following the guidelines provided by Leucken et al. (*8)*. We removed low-quality cells by filtering out barcodes with a very low number of Unique Molecular Identifiers (UMIs) (500) and genes (<250), or with a high percentage of mitochondrial expression (>20%), as it is indicative of lysed cells. Additionally, we considered removing barcodes with a large library size (>50,000) and complexity (>6,000) as they could be putative doublets. We eliminated genes that were detected in very few cells (<10). Finally, filtered data were normalized, log transformed and shared 3,000 highly variable features were used to compute Principal Component Analysis. To achieve successful cell-type annotation combining data from different donors and treatments, we removed the batch-effect with the Harmony integration method (*9*) using the library as a confounder variable. After integration, we created a k-nearest neighbors graph with the "FindNeighbors" function using the first 20 Principal Components, followed by the cell clustering with the Louvain clustering algorithm using the "FindClusters" function at different resolutions. To visualize scRNA-seq data in two-dimensional embedding, we run the Uniform Manifold Approximation and Projection algorithm. To obtain a fine-grained cell type annotation, we followed a top-down approach; first, cells were clustered into large clusters representing cell lineages (B lymphocytes, myeloid cells, T lymphocytes and NK cells), subsequently, cells from the lineages of interest were re-processed and clustered, or even sub-clustered, to define cell types and cell states. To annotate the clusters into specific cell types, we performed a Differential Expression Analysis (DEA) for all clusters to determine their marker genes using the normalized RNA counts, we examined the expression of canonical gene markers, and referred to gene markers from published annotated datasets. Low quality clusters based on poor QC metrics or platelet complexes showing dual gene markers were removed. We performed DEA for cell annotation using the Seurat function "FindMarkers" with the Wilcoxon signed-rank sum test. We defined genes to be relevant markers if the Log2 Fold Change (Log2FC) >0.25, with a FDR adjusted p-value <0.05, and if they were present in at least 25% of cells. For cell type compositional analysis and to estimate changes in the proportions of cell types due to albumin treatment, we compared the fraction of cell types relative to all cells within each lineage population, and computed a paired Wilcoxon signed-rank sum test for each cell type comparing albumin with vehicle treated samples for AD patients and healthy donors, independently. To validate the albumin effect on CD4 T cells using scRNA-seq data, we downloaded the list of genes associated with the BTM of interest and computed a signature-specific score for each cell using the Ucell package (*10*). The BTM signature score was then assessed on specific cell populations across treatment conditions. Statistical analysis between albumin and vehicle treatment conditions was performed using a paired Wilcoxon signed-rank sum test. The code to reproduce the full analysis is hosted at Github: https://github.com/LJimenezGracia/AD-ACLF_cirrhosis_HSAtreatment.

Bulk whole blood RNA-seq.

RNA was isolated from blood stored in Tempus tubes using the Tempus™ Spin RNA Isolation Kit (Applied Biosystems, Foster City, CA) as previously described (*5*). RNA quality was assessed using Agilent RNA 6000 Nano and Pico Chips (Agilent Technologies) and concentration via Qubit RNA HS Assay Kit (Thermo Fisher Scientific). Sequencing libraries were prepared using the TruSeq Stranded Total RNA with Ribo-Zero Globin kit and TruSeq RNA CD Index Plate (both from Illumina Inc.) following TruSeq Stranded Total RNA Sample Prep-guide (Part # 15031048 Rev. E). Starting from 500 ng of total RNA, rRNA and globin mRNA were depleted, and remaining RNA was purified, fragmented, and primed for cDNA synthesis. cDNA first strand was synthesized with SuperScript-II

Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA) for 10 min at 25°C, 15 min at 42°C, 15 min at 70°C and pause at 4°C. cDNA second strand was synthesized with Illumina reagents at 16°C for 1 hour. Then, A-tailing and adaptor ligation were performed. Finally, enrichment of libraries was achieved by PCR (30 s at 98°C; 15 cycles of 10 s at 98°C, 30 s at 60°C, 30 s at 72°C; 5 min at 72°C and pause at 4°C). Afterwards, libraries were visualized on an Agilent 2100 Bioanalyzer using Agilent High Sensitivity DNA kit (Agilent Technologies), quantified using Qubit dsDNA HS DNA Kit (Thermo Fisher Scientific) and sequenced in a NovaSeq-6000 (Illumina Inc.) by at least 100 million paired-end 100nt reads. The reads were aligned to the hg38 genome assembly and the transcriptome using STAR v2.5.3a6 and GENCODE 26 annotation (*11*). RNA-Seq by Expectation Maximization (RSEM) v1.3.0REF.8 was used to compute the expected read counts of each sample from the corresponding binary alignment map (known as BAM) files. Trimmed mean of M values normalization method and limma-voom transformation from rounded expected counts were used to normalize the nonbiological variability. Using the HUGO Genome Nomenclature Committee, which is a resource for approved human gene nomenclature, we included protein-coding genes, and among other gene locus types, immunoglobulin genes, and T-cell receptor genes, resulting in a total of 14,520 genes for analysis. Genes brought by sexual chromosomes and mitochondrial DNA were not included in the analysis. Differential expression between groups was assessed using the moderated t-statistic (*12*).

Data processing of bulk blood RNA-seq.

Bulk blood RNA-Seq data were used for different purposes, that are listed below. RNA-seq data have been deposited at the Gene Expression Omnibus (GEO accession number, GSE171741).

Assessing differential gene expression. DEGs were defined by the simultaneous presence of two criteria, i.e., an absolute FC greater than 1.5 and moderated P value of less than 0.05.

Analyzing gene-set enrichment. The QuSAGE method, as implemented in the QuSAGE package *(13)*, was used to conduct Gene Set Enrichment Analysis (GSEA) using the 346 blood transcription modules (BTMs) as gene sets. BTMs are gene sets related to blood cells (e.g., plasma cells, immunoglobulins (M156.1)) developed through large-scale network integration of publicly available human blood transcriptomes (*14)*. QuSAGE provides an activity score for each gene set in each pairwise comparison. A BTM was identified as DE when the QuSAGE activity score of this BTM had a False Discovery Rate (FDR) <0.05 or P <0.05, where appropriate.

Analyzing shared genes and shared BTMs. A DEG or a differentially expressed BTM was defined as shared between two comparisons when the DEG or differentially expressed BTM were concordant in sign.

Assessing immune-cell signatures with the use of SingleR software. The SingleR software (*15)* was used to compare bulk blood RNA-seq data from our study participants with a reference dataset containing 114 human bulk RNA-seq samples of sorted immune-cell populations from

4 HS (GSE107011) *(16)*. This reference dataset contained 114 human RNA-seq samples annotated to 10 main immune-cell types, including neutrophils, basophils, monocytes, dendritic cells, T cells, CD4 T cells, CD8 T cells, progenitors, B cells, and NK cells. The GSE107011 dataset also contained samples that were additionally annotated to 29 fine immune-cell types, including low-density neutrophils, low-density basophils, classical monocytes, intermediate monocytes, nonclassical monocytes, plasmacytoid dendritic cells, myeloid dendritic cells, naïve CD8 T cells, central memory CD8 T cells, effector memory CD8 T cells, terminal effector CD8 T cells, naïve CD4 T cells, Th1 cells, Th1/Th17 cells, Th17 cells, Th2 cells, T regulatory cells, follicular helper T cells, terminal effector CD4 T cells, Vd2 gamma delta T cells, non-Vd2 gamma delta T cells, MAIT cells, progenitor

cells, naïve B cells, non-switched memory B cells, switched memory B cells, exhausted B cells, plasmablasts, and NK cells. The SingleR pipeline compared our bulk blood RNA-seq data, first with the 10 main immune-cell types of reference and then with the 29-fine immune-cell types. In brief, the annotation in SingleR was performed for each whole blood transcriptome independently. The first step was to identify variable genes among each reference immune-cell types. Variable genes were defined as the top N genes that had a higher median expression in a cell-type compared to each other cell-type; here we used 70 variable genes per cell-type for the first round of analysis. Using only variable genes increased the ability to distinguish closely related cell-types. Then, a Spearman coefficient was calculated for whole blood RNA expression with each of the samples in the reference dataset. The correlation analysis was performed only on variable genes in the reference dataset. Next, multiple correlation coefficients per sample according to the named annotations of the reference dataset were aggregated to provide a single value per cell-type per sample. Of note, SingleR uses the 80^{th} percentile of correlation values, to prevent misclassification due to heterogeneity in the reference samples. Finally, in the fine-tuning step, SingleR reran the correlation analysis, but only for the top cell-types from the previous step. The lowest value cell-type was removed (or values more than 0.05 below the top value), and then this step was repeated until only two cell-types remained. The cell-type corresponding to the top value after the last run was assigned to the blood sample.

Bulk RNA-seq and real-time PCR in PBMCs

Isolation of total RNA from PBMCs was performed using the TRIzol reagent following the manufacturer's instructions. RNA concentrations were assessed in Microvolume UV-Vis Spectrophotometer (Nanodrop One, Thermo Fisher Scientific). Bulk RNA-seq was performed as described above. For real-time PCR analysis, cDNA synthesis from 500 ng of total RNA was performed using the High-Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA). Real-time PCR analysis of human IGHA1 (Hs00733892_m1, encoding IgA1), IGHM (Hs00941538_g1, encoding IgM) and IHG2 (Hs00390545_3, customized, encoding IgG2) was performed in QuantStudio™ 7 Pro Real-Time PCR System (Thermo Fisher Scientific) using RPS18 (Hs01375212_g1, encoding for RPS18) as endogenous control.

Neutrophil degranulation assay.

After 30 minutes of resting, neutrophils were seeded at a density of $3x10^6$ cells/mL and incubated with either HSA, recombinant human albumin (both at 15 mg/ml) or vehicle control in the absence or presence of phorbol 12-myristate 13-acetate (100 nM) for 2 hours at 37ºC in a 5% CO_2 incubator. At the end of the incubation period, supernatants were collected to measure degranulation using the Neutrophil MPO Activity Assay Kit (Cayman Chemical, Ann Arbor, MI). Briefly, 25 µl of neutrophil supernatant together with 25 µl of assay buffer were added to each well of the experimental plate. Then, 50 µl of 3,3',5,5'-tetramethylbenzidine (TMB), a substrate for horseradish peroxidase, were added to each well and absorbance was measured at minute one and minute five after the addition of TMB in a microplate reader (Infinite M PLEX Monochromator, TECAN, Männedorf, Switzerland). The assay was performed at room temperature.

Neutrophil phagocytosis assay.

After 30 minutes of resting, neutrophils were seeded at a density of $5x10^5$ cells/mL and incubated with either HSA, recombinant human albumin (both at 15 mg/ml) or vehicle control for 2 hours at 37ºC in a 5% CO_2 incubator. Thereafter, 50 µL of opsonized fluorescein conjugate zymosan bioparticles or

fluorescent-labeled *E. coli* (Thermo Fisher Scientific) were added to each well (ratio cells/bioparticles, 1:10) to a final volume of 200 µL and incubated at 37ºC for 60 min. Cells were then washed with sterile $DPBS^{--}$ and 100 µL trypan blue solution (diluted 1/10 in sterile $DPBS^{--}$) were added to quench fluorescence of extracellular bioparticles. Plates were finally centrifuged for 5 min at 400 g at room temperature and excess trypan blue was carefully aspirated. The fluorescent intensity of each well was read in a microplate reader (FLUOstar Optima, Ortenberg, Germany).

Neutrophil chemotaxis assay.

A chemotaxis-phagocytosis assay in microfluidic arenas (*17*) was used to test the role of albumin on the ability of neutrophils to migrate directionally towards and to phagocytose *Candida albicans* yeast. For this purpose, microfluidic arenas were fabricated in PDMS and glass at the BioMEMS Core. The arenas were primed with a suspension of *Candida* yeast at 107 yeast cells/ml. The channels outside the arenas were washed with media to remove any extra *Candida* outside the arenas. After the addition of neutrophils to the chemotaxis assay in the presence or absence of HSA, images of the arenas were recorded on a Nikon Ti-E microscope with a motorized stage and environmental chamber. The number of neutrophils recruited to each arena was counted form brightfield images over time. The total area of GFP-expressing *Candida* was quantified from fluorescence images.

Swarming assay.

A swarming assay was used to test the contribution of HSA to the ability of neutrophils from patients to contain the growth of *Candida* clusters. To this purpose, arrays of *Candida*-adherent spots (200 µm diameter) were printed utilizing a microarray printing platform (Picospotter PolyPico, Galway, Ireland) and a solution of poly-l-lysine (Sigma-Aldrich). The arrays were printed onto ultra-clean glass slides (Thermo Fisher Scientific). The printed slides were mounted into 16-well ProPlate wells (Grace Bio-labs, Bend, OR) and 50 µL of a suspension of *Candida* inoculums in PBS were added to each well and incubated with rocking for 5 min. Following incubation, the wells were washed with PBS to remove unbound yeast from the glass surface. The mounted slides were checked to ensure appropriate patterning of targets onto the spots with minimal non-specific binding before use. Isolated neutrophils were counted and re-suspended in IMDM + 20% FBS with the indicated concentration of albumin or vehicle, and then added to swarming chambers at 500,000 cells per well. Swarming was observed using a Nikon Ti-E microscope. Time-lapse imaging was conducted using a 10x Plan Fluor Ph1 DLL (NA=0.3) lens and endpoint images were taken with a 2x Plan Apo (NA=0.10) lens. All selected points were optimized for perfect focus before launching the experiment. Fungal growth area analysis was automated in ImageJ software.

SUPPLEMENTARY FIGURES

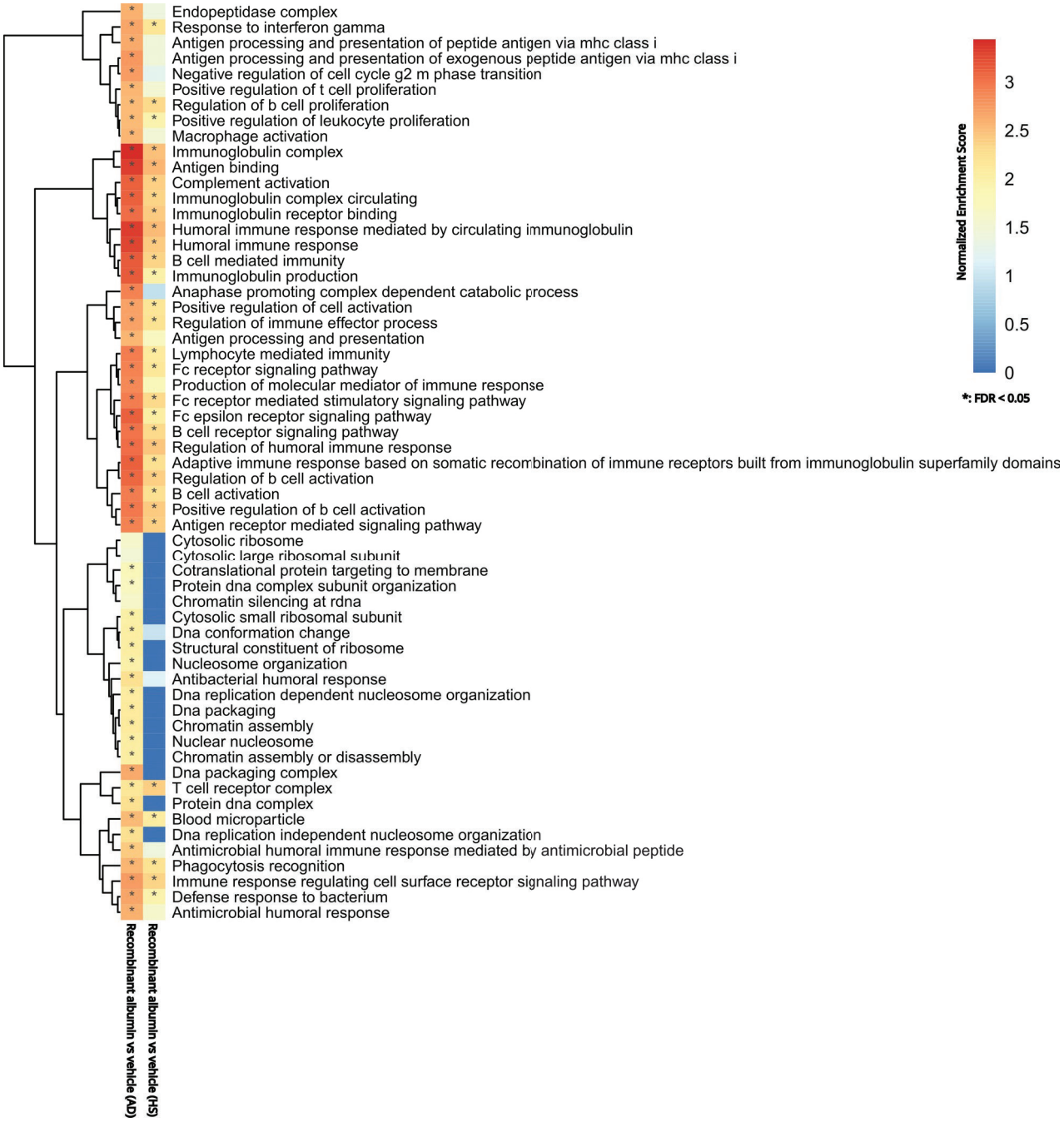


Figure S1. Changes in the transcriptional landscape of PBMCs isolated from HS incubated in vitro with recombinant albumin. RNA-seq data were submitted to GSEA to generate ranked lists of genes for the following pairwise comparisons: recombinant albumin versus vehicle in PBMCs from patients with AD cirrhosis and recombinant albumin versus vehicle in PBMCs from healthy subjects. Gene-enrichment analysis was then performed. Color gradient corresponds to increasing values of the normalized enrichment score (NES) of gene sets from the less upregulated (in blue) to the most upregulated (in red). The presence of stars in cells of the heat map indicates a false discovery rate (FDR) <0.05. For comparison, the transcriptional changes in PBMCs isolated from patients with AD cirrhosis incubated *in vitro* with recombinant albumin are shown.

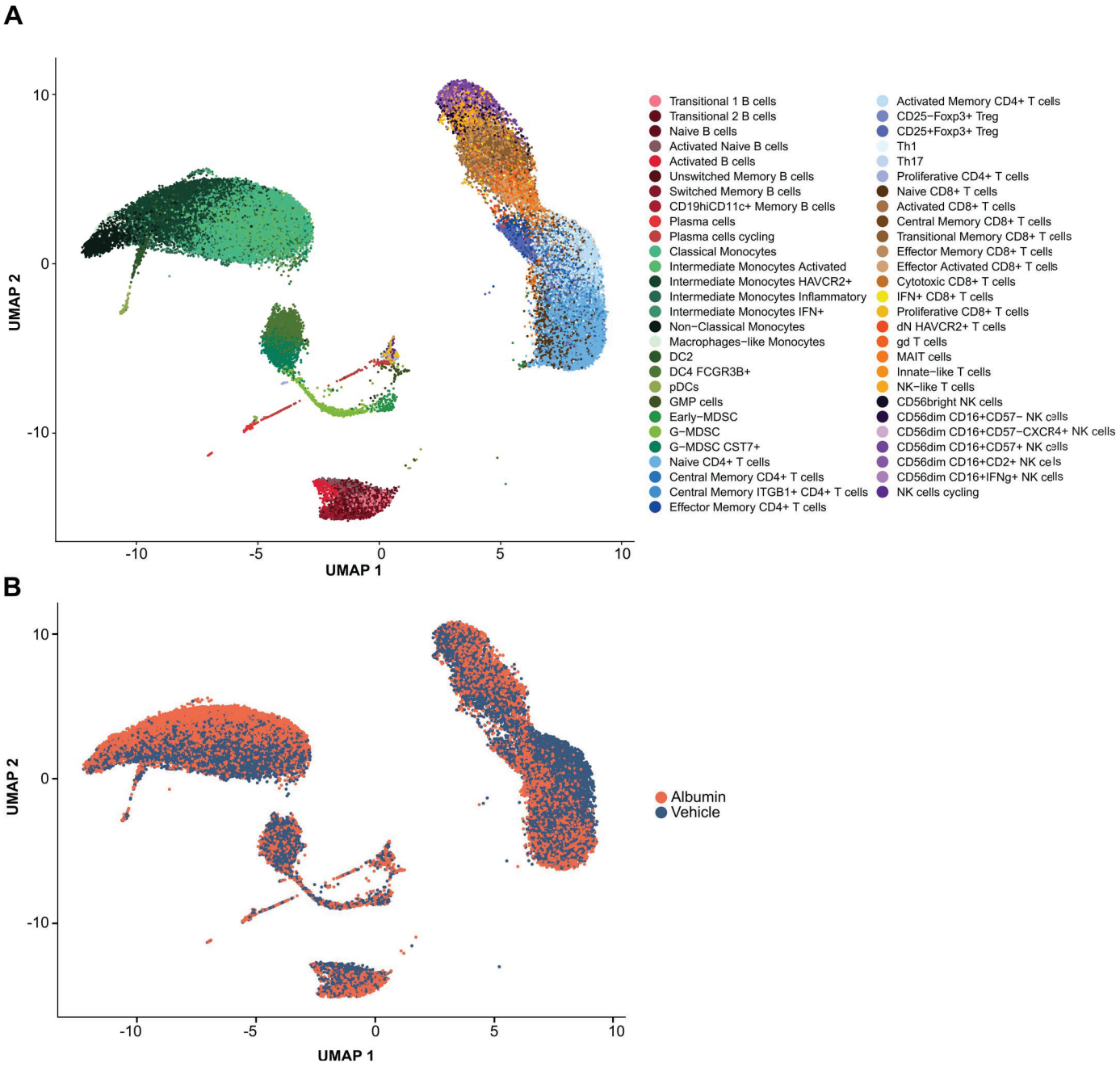


Figure S2. Overview of results of scRNA-seq experiments of PBMCs from patients with AD cirrhosis. Fresh PBMCs from 9 patients with AD cirrhosis were exposed *in vitro* to albumin (15 mg/ml) or vehicle during two hours before performing scRNA-seq. (A) Uniform manifold approximation and projection (UMAP) of 69,293 human PBMCs from AD patients exposed to albumin and vehicle, colored by cell types (a total of 66,064 PBMCs were further analyzed after excluding NK cells). (B) Overlay of albumin and vehicle exposure on the full PBMC UMAP.

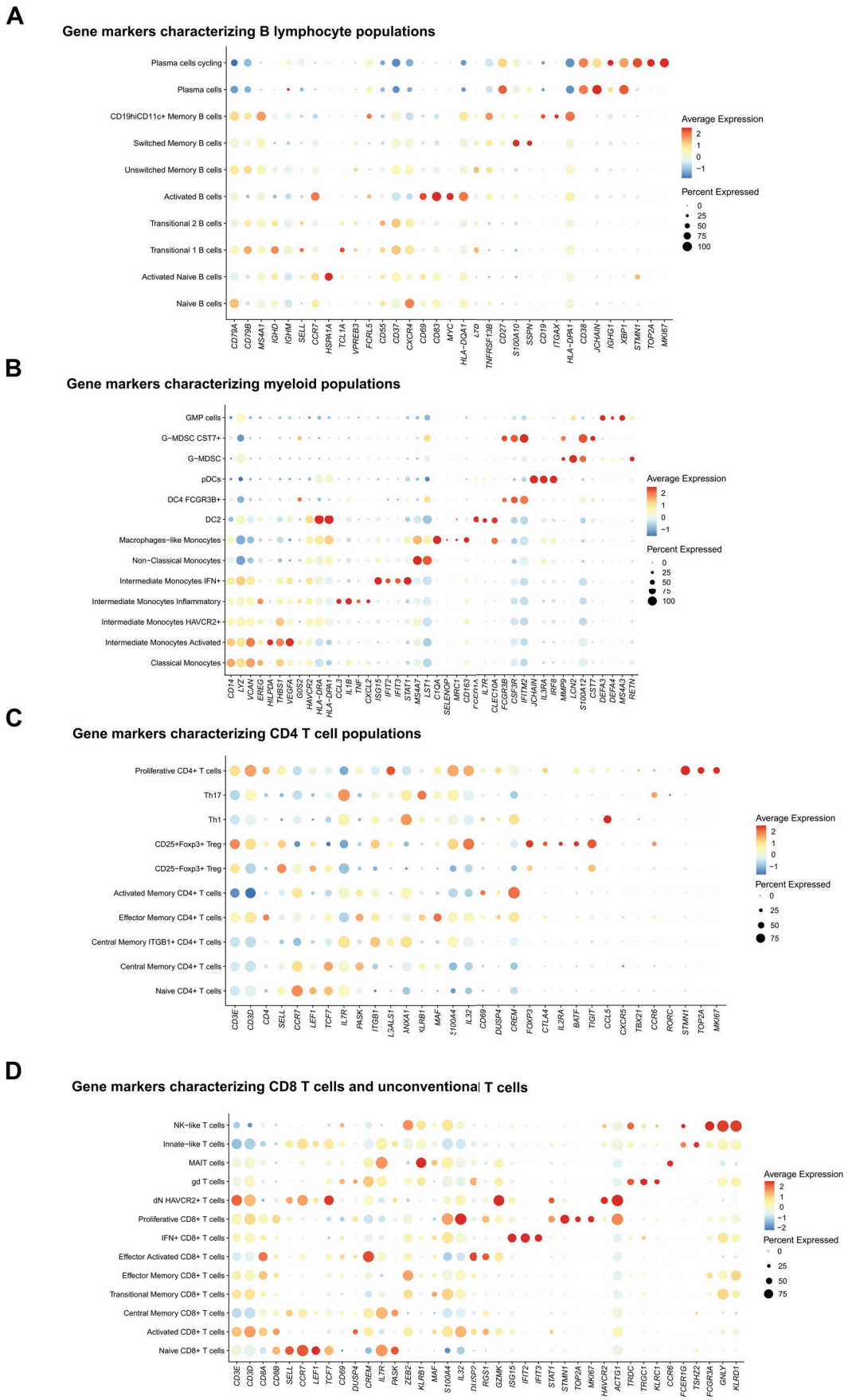


Figure S3. Top gene markers describing all the immune-cell populations identified in the scRNA-seq dataset. Dot plot showing the average expression for top genes (*x-axis*) across all cell populations (*y-axis*) for B lymphocytes (A), myeloid cells (B), $CD4^+$ (C) and $CD8^+$ and unconventional T cells (D). Dot size represents the percentage of cells in a cluster expressing each gene, and the color indicates the average expression level.

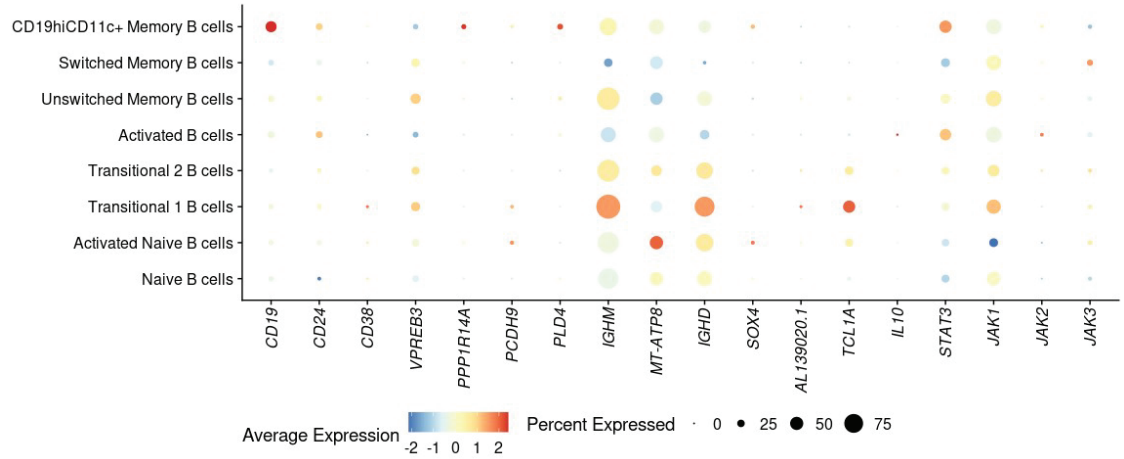


Figure S4. Identity of the B cell populations affected by albumin using scRNA-seq data. Expression of genes associated with transitional-like identity from Steward et al. *(18)* across B cell populations, confirming that albumin induces transcriptional changes resembling transitional-like B cell signatures.

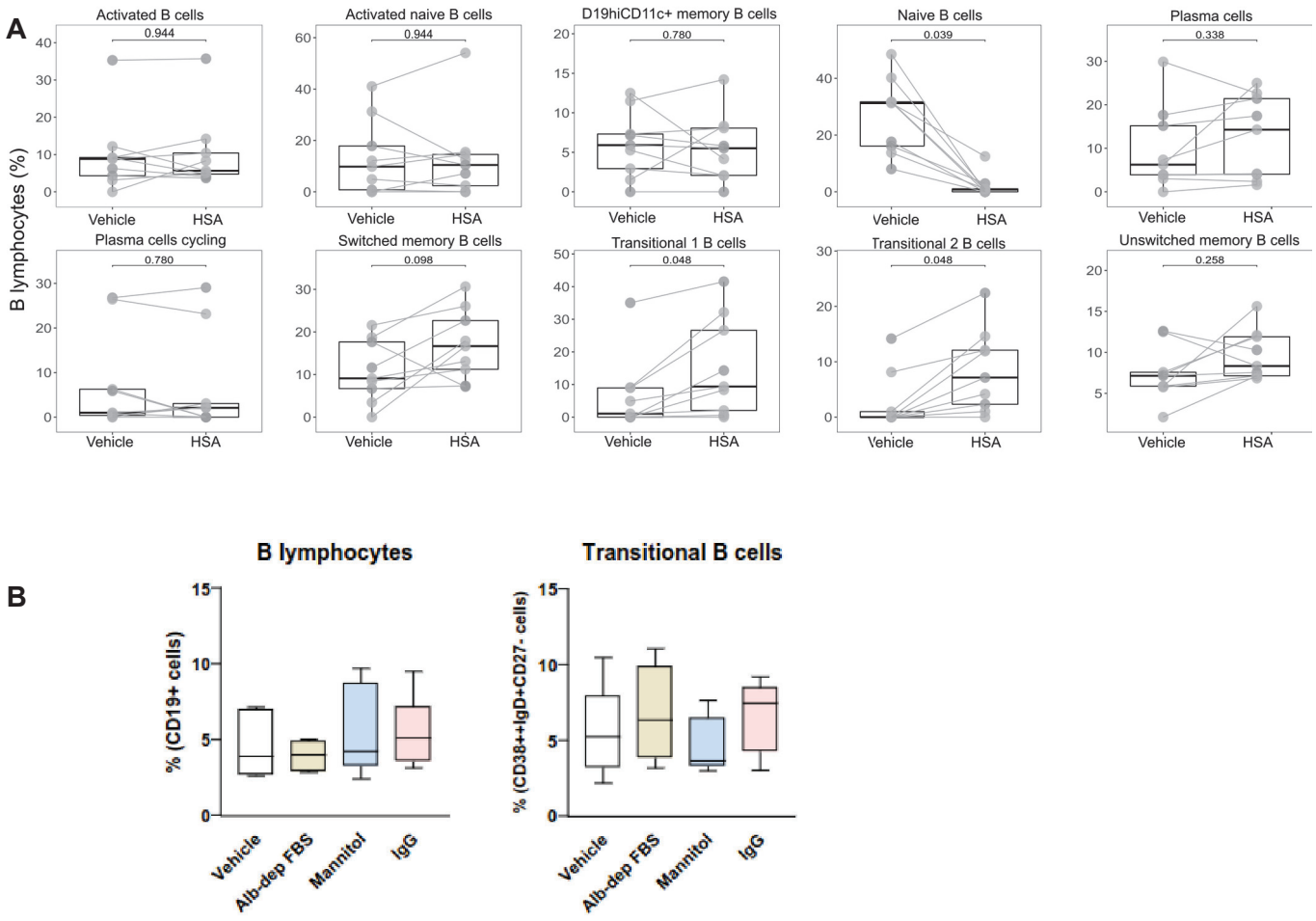


Figure S5. In vitro effects of human serum albumin (HSA) on B lymphocytes. (A) Boxplots for the abundance of B lymphocyte populations using scRNA-seq data. All comparisons between HSA and vehicle were tested for significance with a paired Wilcoxon signed-rank sum test and adjusted p-values are indicated. This figure has been designed with results obtained in cells from 9 patients with AD cirrhosis. **(B)** Flow cytometry analysis of the B cell and transitional B cell populations in response to vehicle control, 15% albumin-depleted FBS (alb-dep FBS), mannitol (15 mg/ml) and IgG (15 mg/ml).

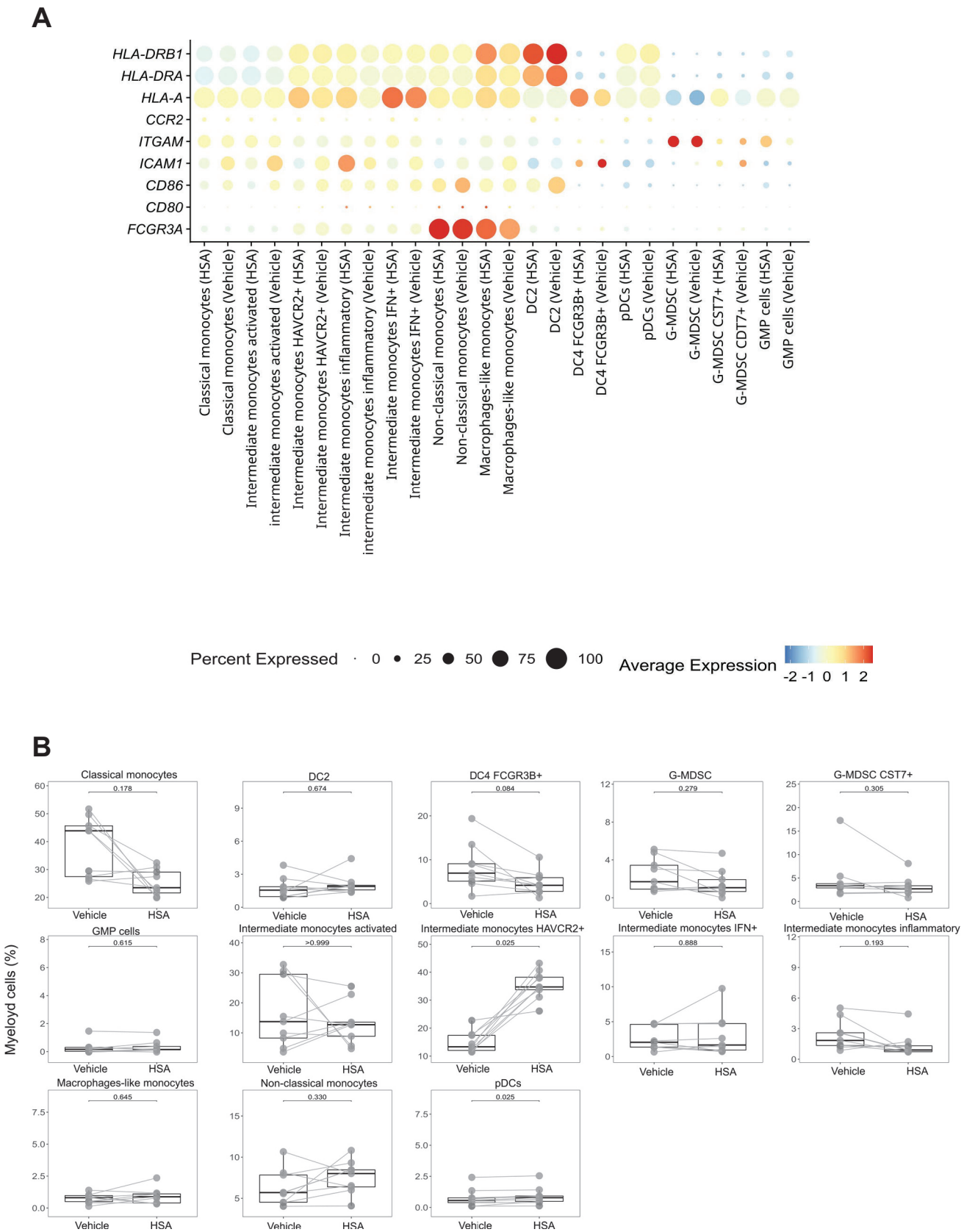


Figure S6. Effects of HSA on myeloid cells using scRNA-seq. (A) Expression of genes associated with monocyte activation in cells exposed *in vitro* to either vehicle or HSA. **(B)** Boxplots for the abundance of myeloid cells. All comparison of cell type abundance between HSA and vehicle were tested for significance with a paired Wilcoxon signed-rank sum test, and adjusted p-values are indicated. This figure has been designed with results obtained in cells from 9 patients with AD cirrhosis.

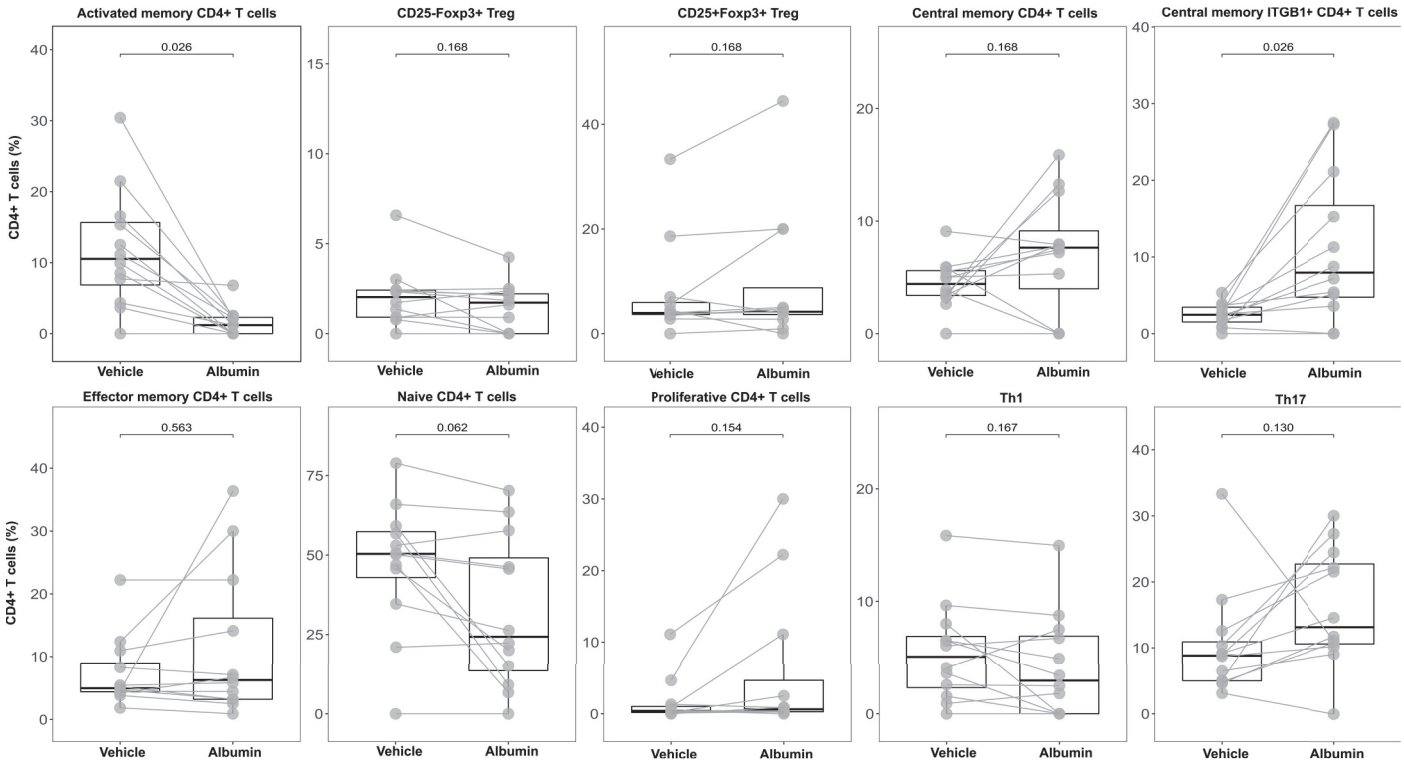


Figure S7. Analysis of the cell type proportions for patients' CD4$^+$ T cells exposed *in vitro* to albumin. This figure has been designed with results obtained in cells from 9 patients with AD cirrhosis. Boxplots for the abundance of CD4$^+$ T cells. All comparison of cell type abundance between albumin and vehicle were tested for significance with a paired Wilcoxon signed-rank sum test; adjusted P values are indicated.

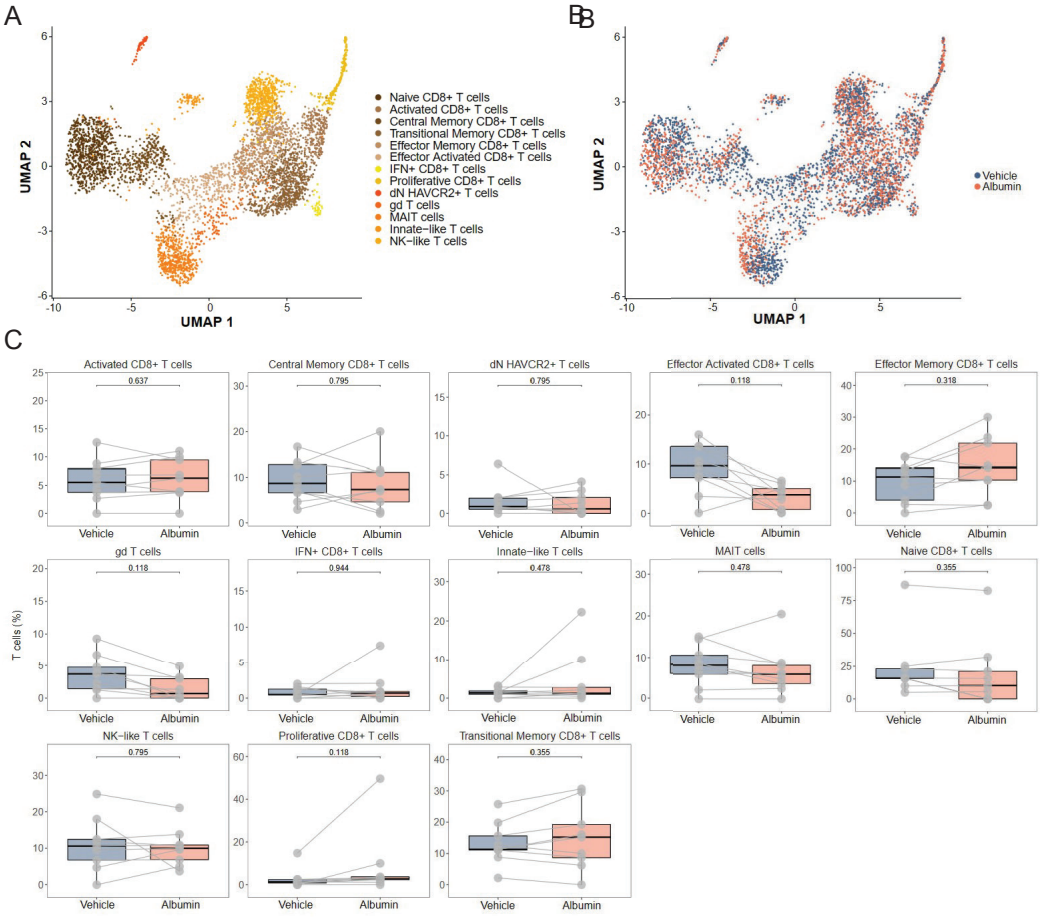


Figure S8. Analysis of patients' CD8$^+$ and unconventional T cells exposed *in vitro* to albumin. This figure has been designed with results obtained in cells from 9 patients with AD cirrhosis. **(A)** UMAP of 3483 CD8$^+$ T cells and 1292 unconventional T cells exposed to albumin and vehicle, colored by cell populations. **(B)** Overlay of albumin and vehicle exposure on the CD8 and unconventional T-cell UMAP. **(C)** Box plots for the abundance of CD8 and unconventional T cell types after albumin exposure. All comparison of cell type abundance between albumin and vehicle were tested for significance using a paired Wilcoxon signed-rank sum test; adjusted p-values are indicated.

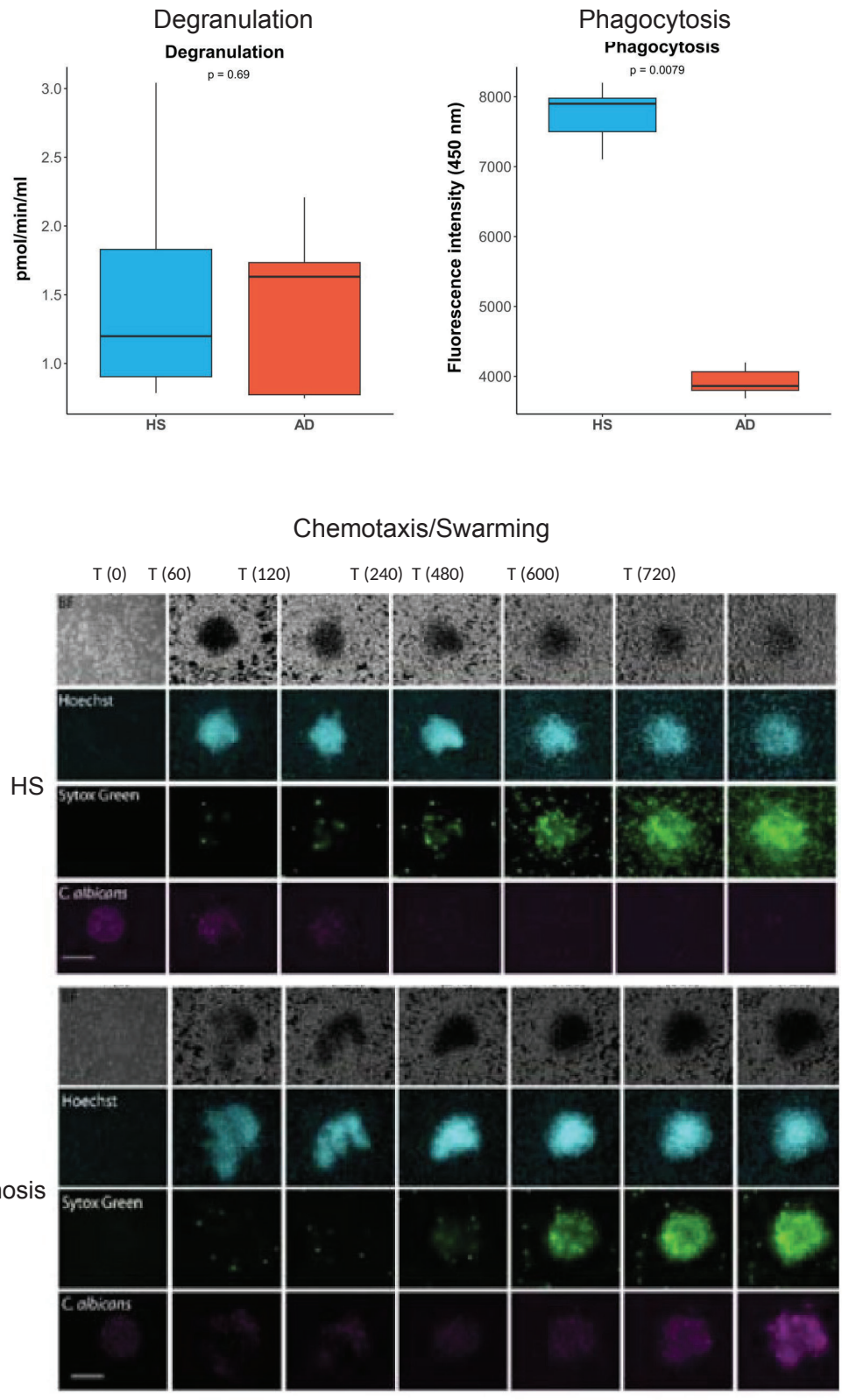


Figure S9. Neutrophils from patients with AD cirrhosis exhibit impaired anti-microbial functions. Neutrophil degranulation was assessed by measuring the MPO enzymatic activity in the cell supernatants. Phagocytosis was assessed by ingestion of FITC-conjugated zymosan bioparticles. Chemotaxis and swarming were assessed through the ability of neutrophils to delay the escape of *Candida albicans* hyphae from the neutrophil swarm.

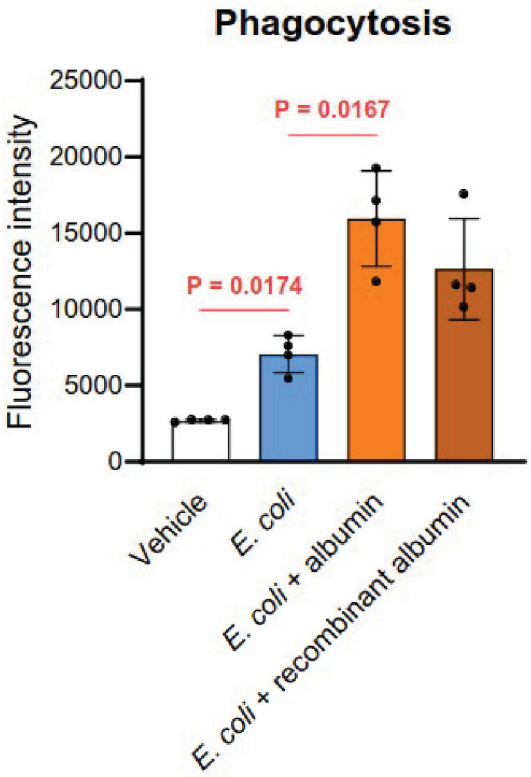


Figure S10. Phagocytic capacity assessed by incubating neutrophils with fluorescently labeled *E. coli* alone or in the presence of albumin and recombinant albumin for 60 min and compared to vehicle control.

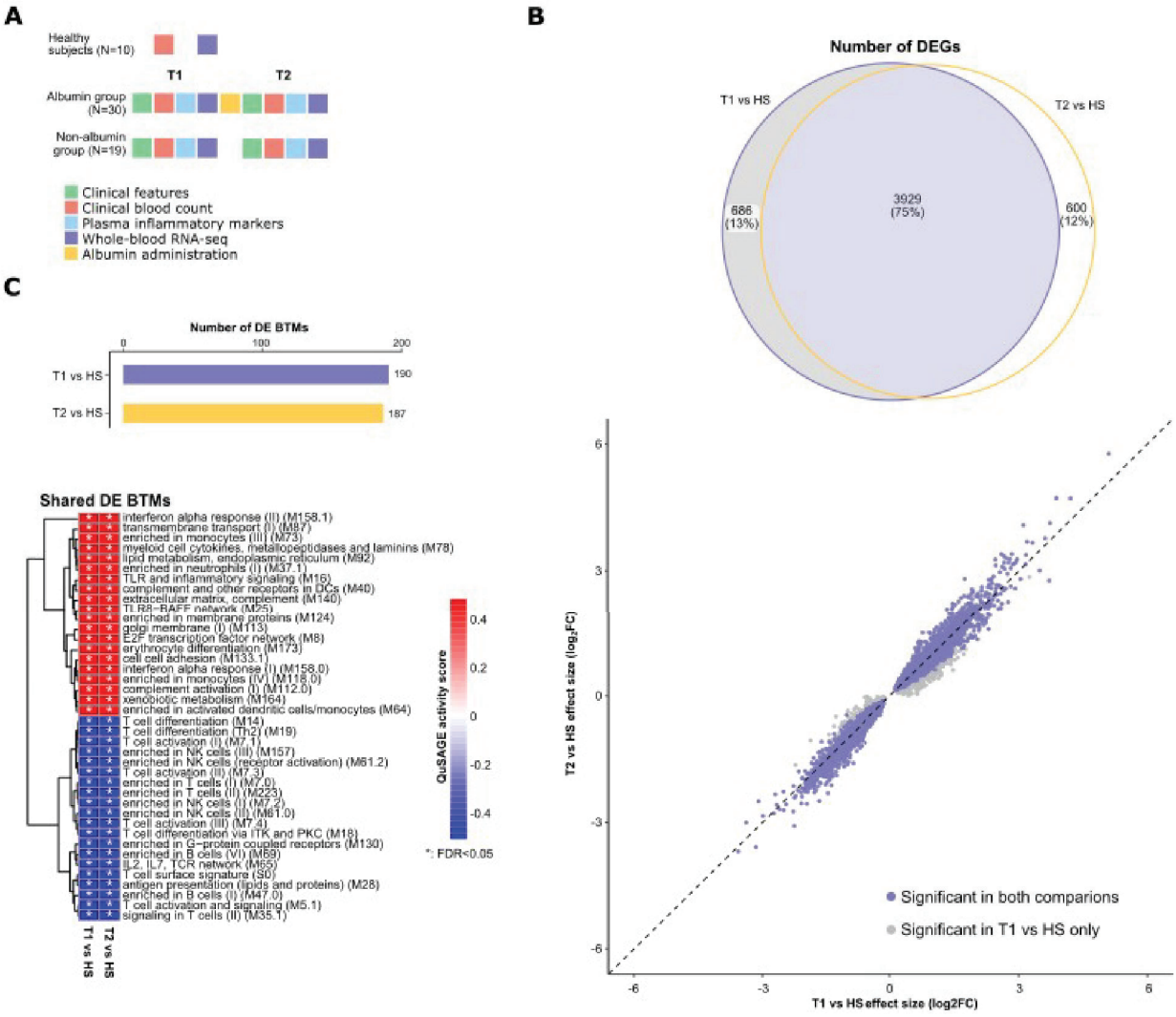


Figure S11. Whole-blood transcriptional characteristics of all patients at T1 and T2 stages. (A) Overview of the longitudinal collection of clinical data and blood specimens (including blood that has been used for RNA-sequencing [RNA-seq]) in 49 patients at time 1 (T1, at enrollment) when they had AD cirrhosis and high risk of developing ACLF and at time 2 (T2) when they had progressed to ACLF. Thirty patients had received intravenous albumin when progressing from T1 to T2. Nineteen patients had not received albumin during the progression to ACLF. Ten age-matched healthy subjects were also studied, but only once. **(B)** Top, Venn diagram showing the overlap of differentially expressed genes (DEGs) in two pairwise comparisons, T1 vs healthy subjects and T2 vs healthy subjects. Bottom, Differential expression effect-size (log2 FC) of T2 compared with T1 changes for the 4615 T1 DEGs. **(C)** Top, number of differentially expressed (DE) blood transcription modules (BTMs) in two pairwise comparisons (T1 vs healthy subjects and T2 vs healthy subjects). Bottom, heatmap showing representative significantly upregulated BTMs and downregulated BTMs that are shared in two comparisons, T1 vs healthy subjects and T2 vs healthy subjects. BTMs are hierarchically clustered based on significant QuSAGE activity scores obtained in the comparison of T1 with healthy subjects. Asterisks denote false-discovery rate (FDR) <0.05. Color represents QuSAGE activity score.

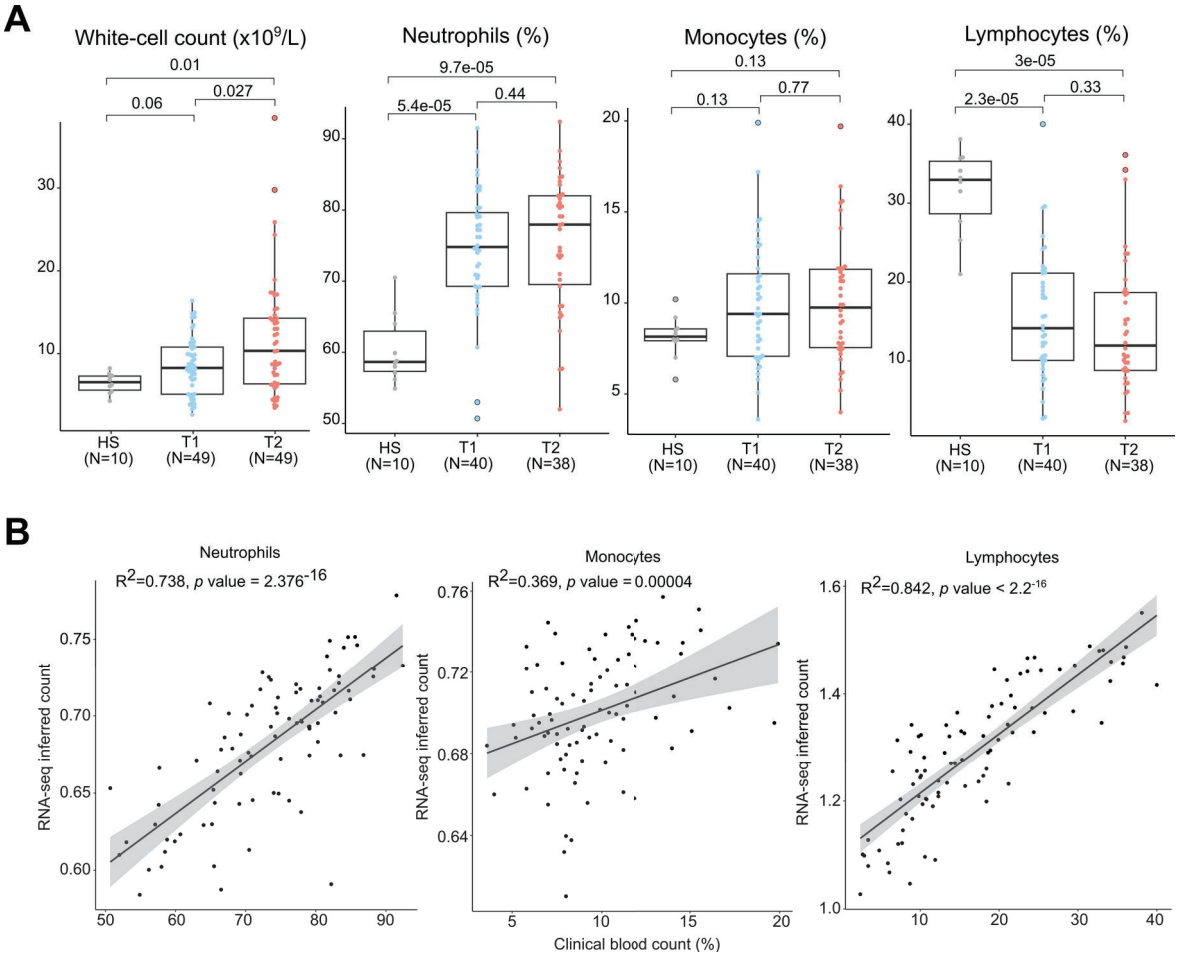
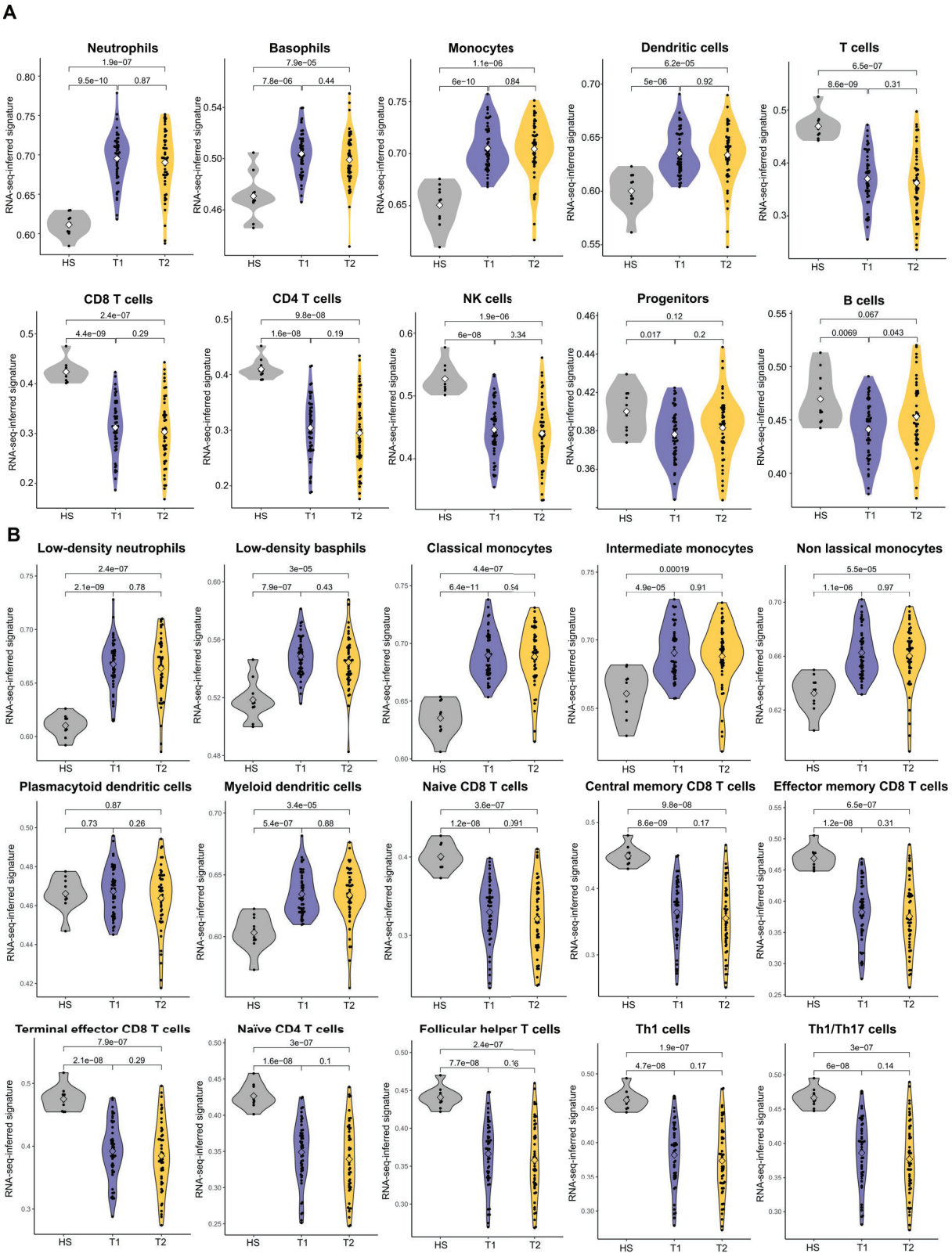


Figure S12. Clinical blood counts and RNA-seq-inferred blood counts. Longitudinal collection of blood specimens was used for clinical blood count measurements and RNA-sequencing [RNA-seq]) in 49 patients at time 1 (T1, at enrollment) when they had pre-ACLF and at time 2 (T2) when they had progressed to ACLF. Ten age-matched healthy subjects were also studied, but only once. (A) Box plots of clinical complete blood counts were obtained in healthy subjects, and in patients at T1 and T2. There were no missing data in healthy subjects and for white blood cell count in patients. In patients, there were missing data for differential counts of neutrophils, monocytes, and lymphocytes in 2 patients. The horizontal line in each box represents the median, the lower and upper boundaries of the boxes the interquartile range, and the length of the vertical lines are 1.5 times the interquartile range. The P values are from Kruskal–Wallis tests followed by Mann–Whitney U tests. (B) Neutrophil, lymphocyte, and monocyte counts in paired clinical complete blood counts as compared with the RNA-seq-inferred blood counts obtained with the use of peripheral blood (47 paired specimens from healthy subjects, pre-ACLF and ACLF). The shaded areas represent the 95% confidence intervals. To obtain this figure, we correlated the clinical differential blood counts for neutrophils, monocytes and lymphocytes from all study participants including healthy subjects with the corresponding RNA-seq–inferred counts for these white cells (calculated based on the results of the SingleR software of RNA deconvolution for 10 main immune-cell types; see **Figure S13A**).



Continued on next page.

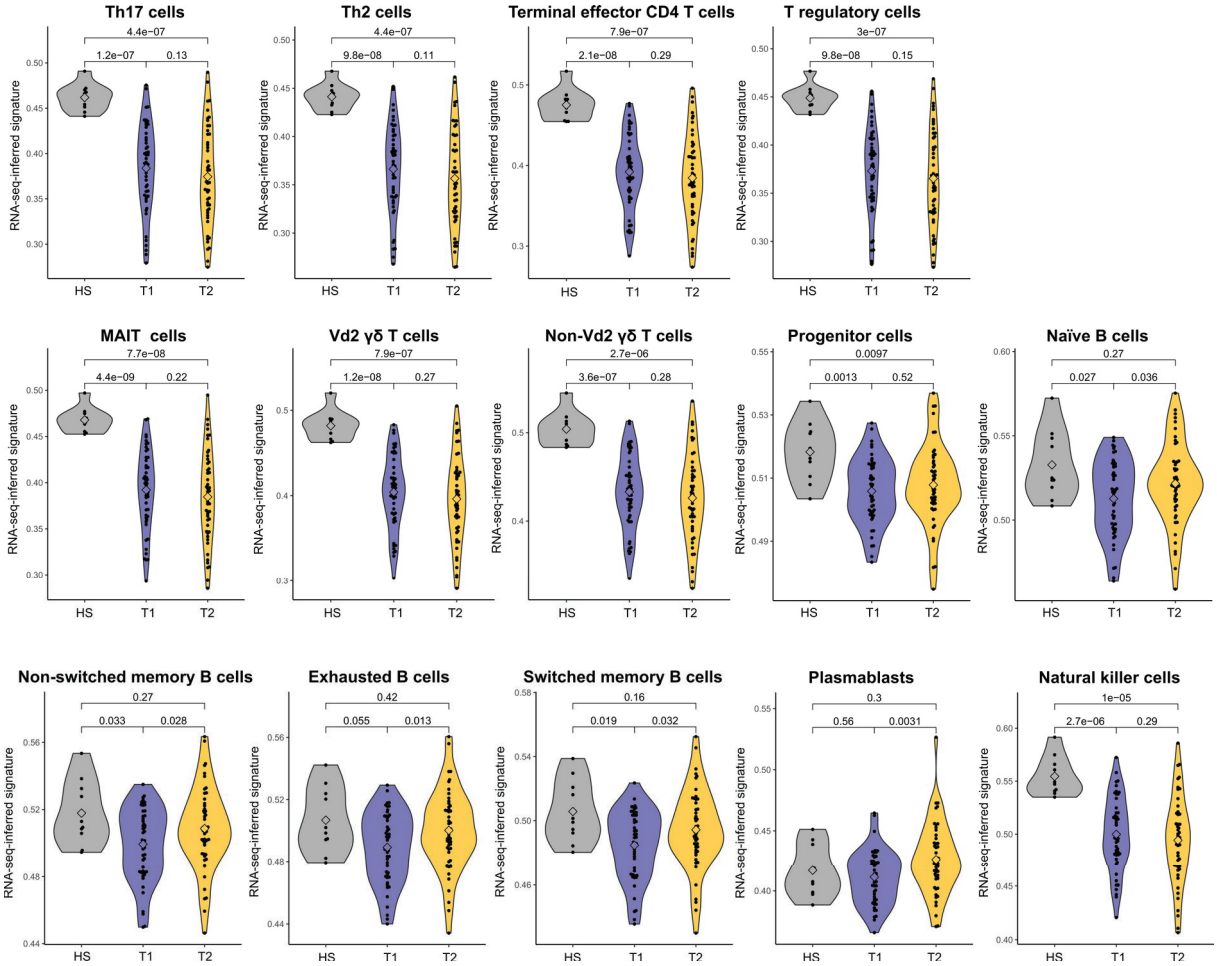


Figure S13. Violin plots of RNA-seq-inferred signatures of blood immune cells in all patients at T1 and T2 and healthy subjects. Longitudinal collection of blood specimens was used for RNA-sequencing [RNA-seq]) in 49 patients, at time 1 (T1, at enrollment) when they had pre-ACLF, and at time 2 (T2) when they had progressed to ACLF. Ten age-matched healthy subjects were also studied, but only once. Whole-blood RNA-seq data were analyzed with the use of the SingleR software to determine signatures for each of the 10 main immune-cell types **(A)** and each of the 29-fine immune-cell types **(B).** White rhomboid symbols indicate medians. The P values are from Kruskal–Wallis tests followed by Mann–Whitney U tests.

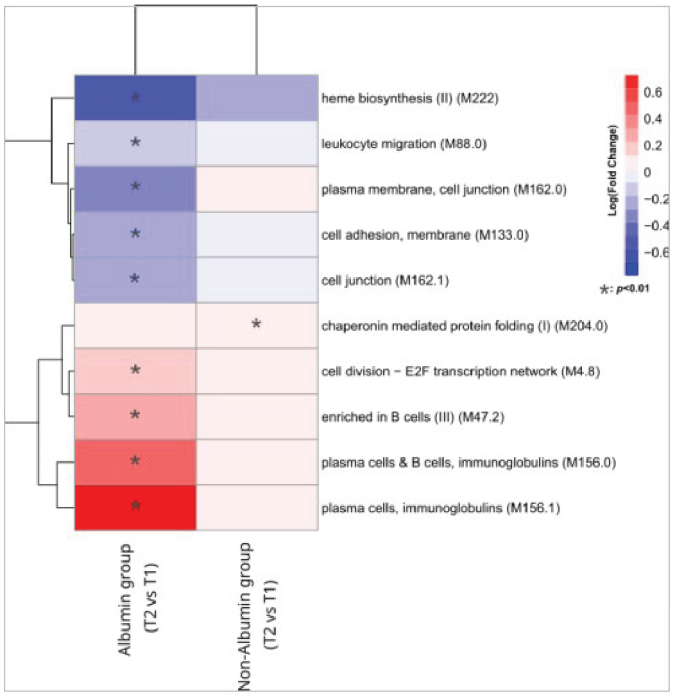


Figure S14. Analysis of BTMs after adjusting by a non-linear regression multivariate model the transcriptomic data by disease severity of each patient estimated by the MELD score.

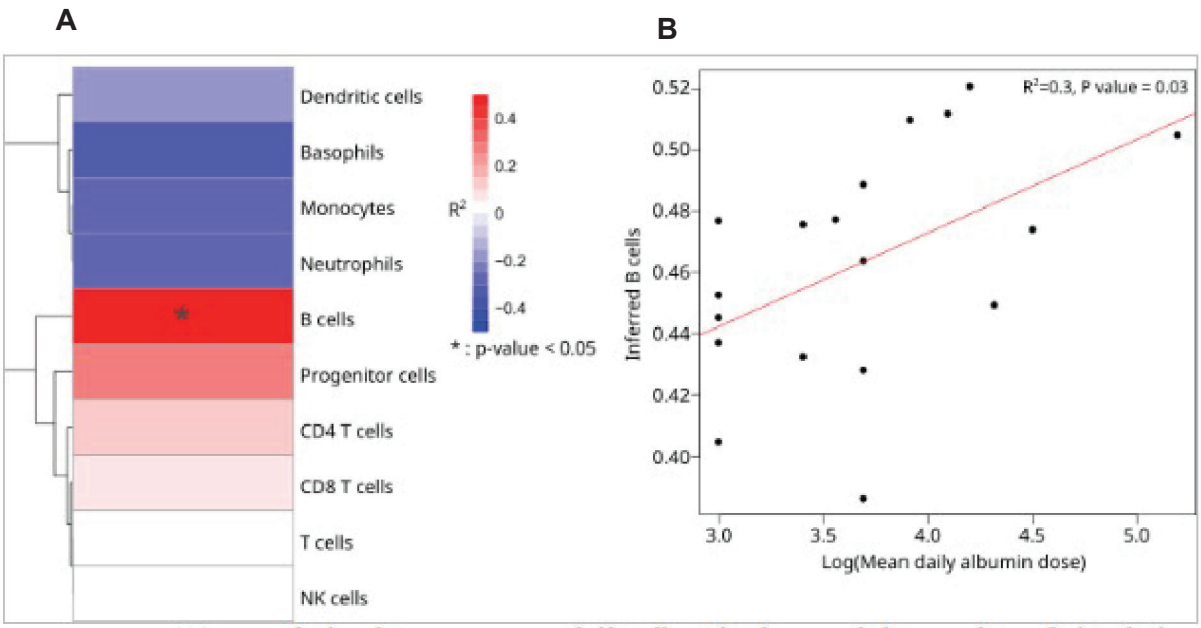


Figure S15. (A) Correlation between mean daily albumin dose and the number of circulating immune cells inferred from Monaco cell population estimations. **(B)** Correlation between mean daily albumin dose and the number of B cells by means of Monaco estimation.

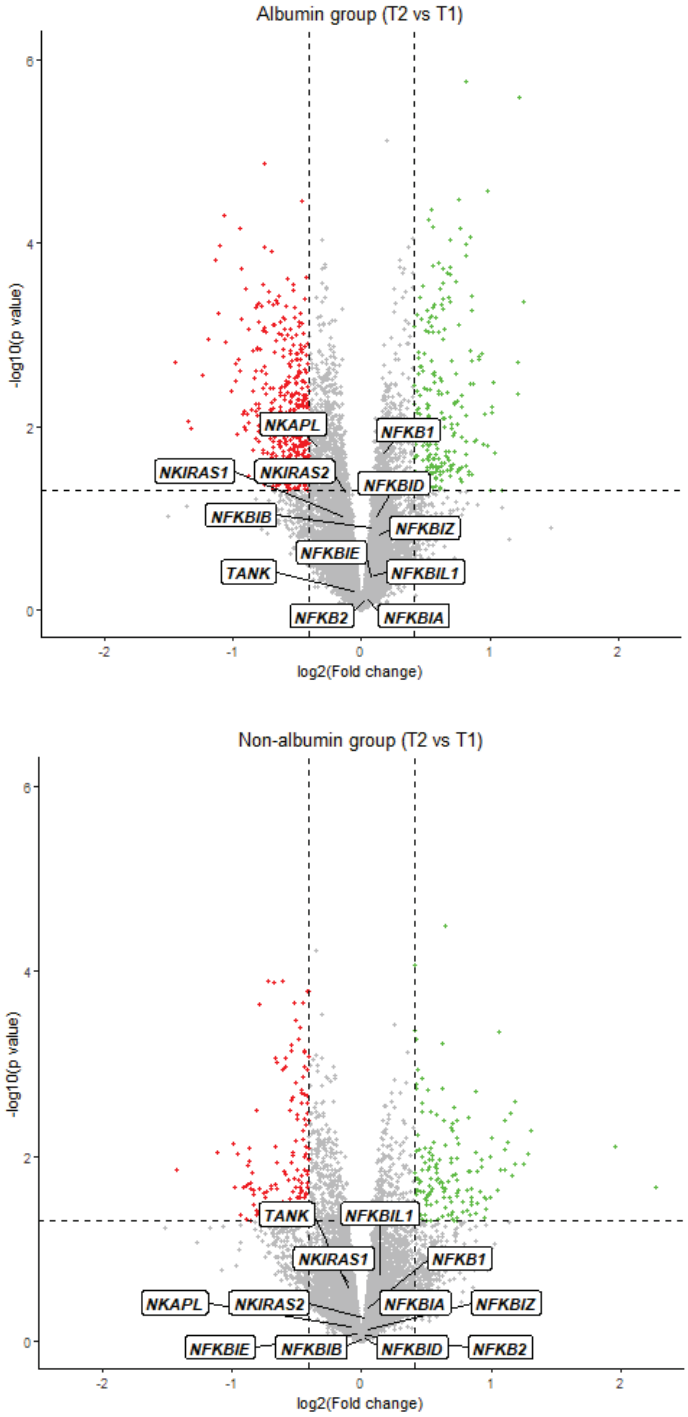
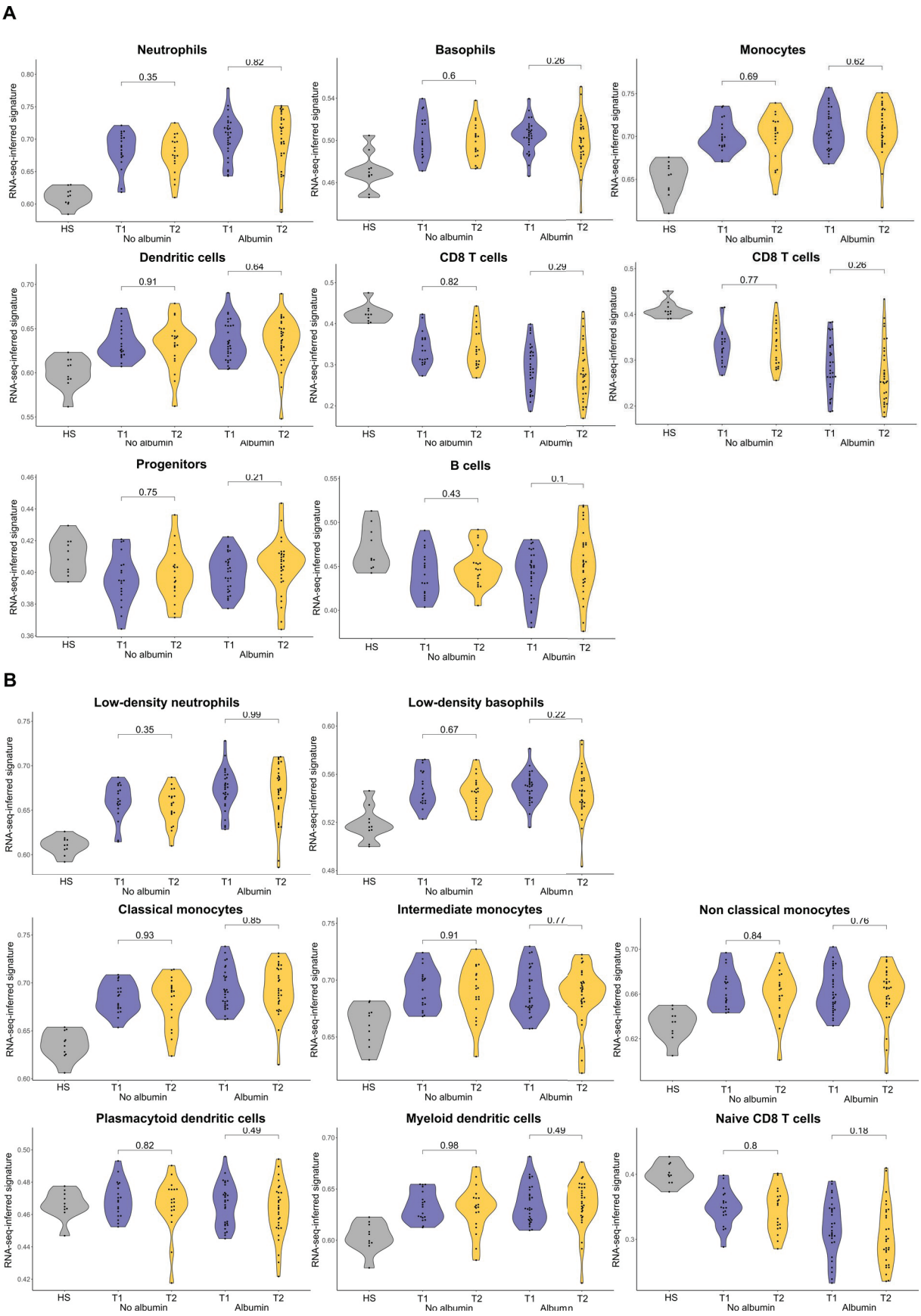
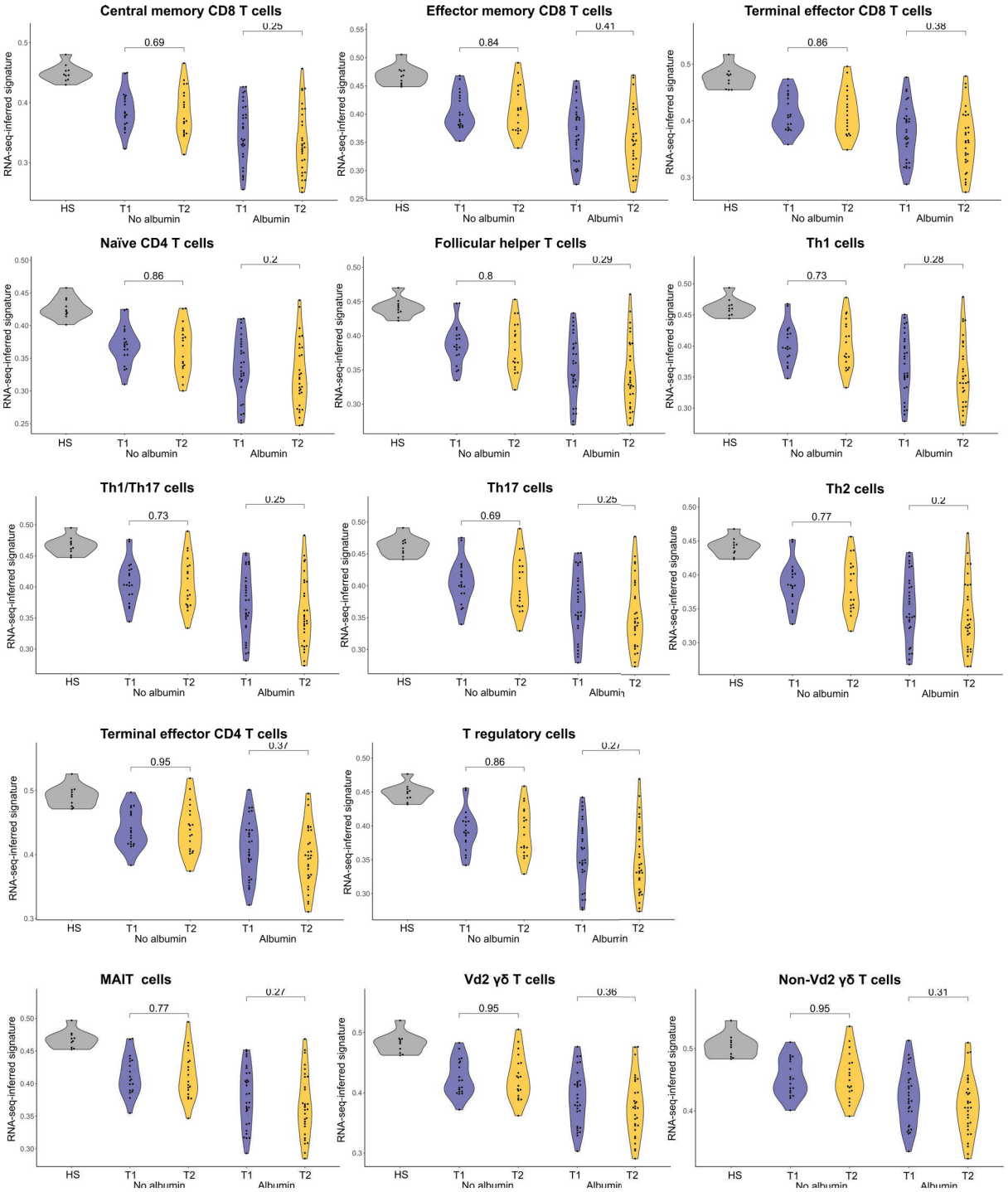


Figure S16. Volcano plots showing differential expression effect-size (log2 FC) in genes of the NF-κB family between T2 and T1 plotted against significance (-log10 p value), for the albumin group (top) and non-albumin group (bottom). In both volcano plots, gray points indicate genes with no significant difference in expression between T2 and T1 (with absolute FC of less than 1.5 and P>0.05, i.e., -log10 P <1.3). *NFKB1,* Nuclear Factor Kappa B Subunit 1; *NFKB2,* Nuclear Factor Kappa B Subunit 2; *NKAPL*: NFKB Activating Protein Like, *NKAP,* NFKB Activating Protein; *TANK*, TRAF Family Member Associated NFKB Activator; *NKRF*, NFKB Repressing Factor; *NFKBIA*, NFKB Inhibitor Alpha; *NFKBIB*, NFKB Inhibitor Beta; *NFKBIZ*, NFKB Inhibitor Zeta; *NFKBID*, NFKB Inhibitor Delta; *NFKBIE*, NFKB Inhibitor Epsilon; *NFKBIL1*, NFKB Inhibitor Like 1; *NKIRAS1*, NFKB Inhibitor Interacting Ras Like 1; *NKIRAS2* ,NFKB Inhibitor Interacting Ras Like 2.



Continued on next page.



Continued on next page.

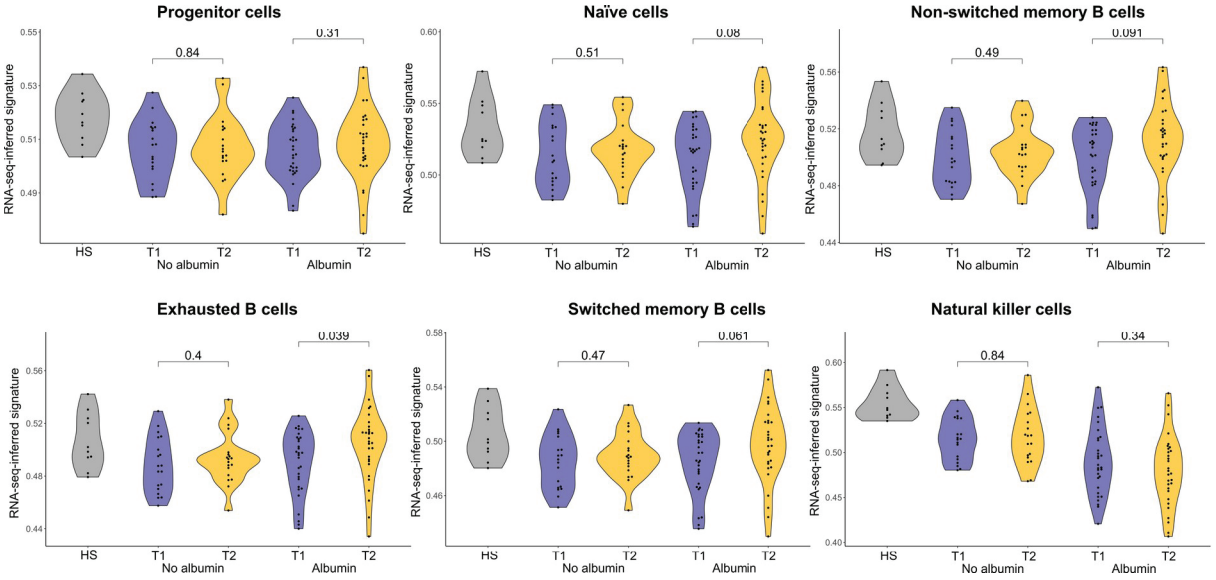


Figure S17. Violin plots of RNA-seq-inferred signatures of blood immune cells at T1 and T2 in patients of the albumin group and those of the non-albumin group. Longitudinal collection of blood specimens was used for RNA-sequencing [RNA-seq]) in 49 patients at time 1 (T1, at enrollment) when they had pre-ACLF and at time 2 (T2) when they had progressed to ACLF. Thirty patients had received intravenous albumin when progressing from T1 to T2. Nineteen patients had not received albumin during the progression to ACLF. Whole-blood RNA-seq data were analyzed with the use of the SingleR software to determine signatures for each of the 10 main immune-cell types **(A)** and each of the 29-fine immune-cell types **(B).** Of note, results of plasmablasts are not shown in panel B because these results are shown in Figure 5D, top. Results obtained in 10 healthy subjects are provided for reference. The P values are from Kruskal–Wallis tests followed by Mann–Whitney U tests.

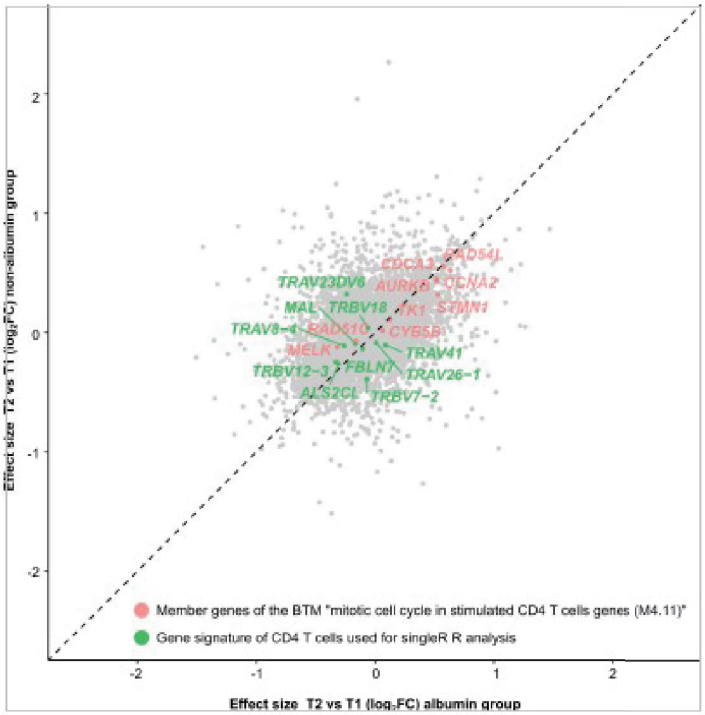


Figure S18. Scatter plots showing the lack of overlap between member genes for the BTM related to mitotic cell cycle in stimulated CD4 T cells (M4.11) and genes defining the signature CD4 T cells used in the SingleR pipeline. There were 9 genes (red color) and 10 genes (green color), respectively.

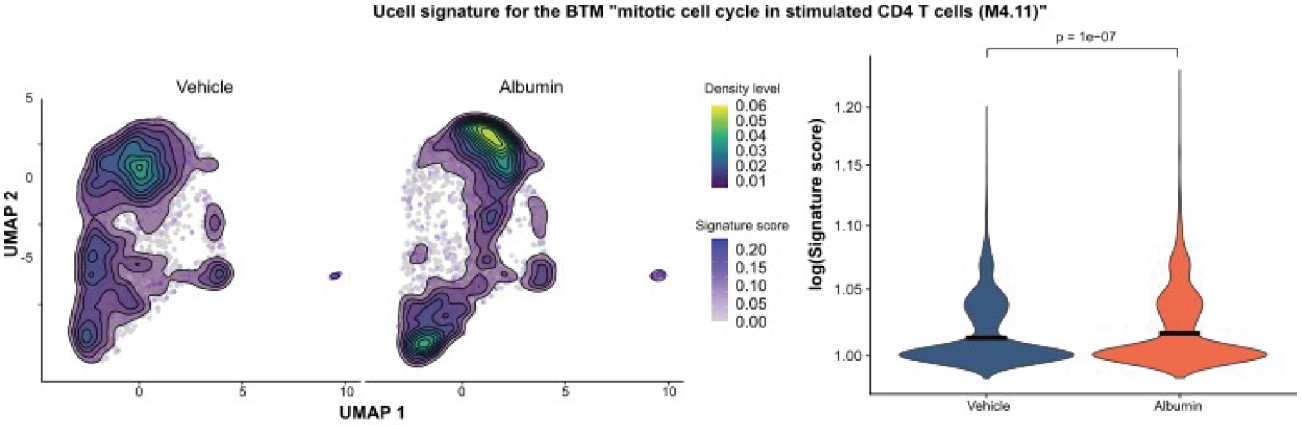


Figure S19. Left, representative density plot showing the signature score for the BTM "mitotic cell cycle in stimulated CD4 T cells (M4.11)" on the CD4 T cells UMAP across vehicle and albumin conditions. Right, comparison of log(signature score) for the BTM 4.11 between albumin and vehicle on the CD4 T cell compartment. Statistical analysis for BTM signature was performed using a Wilcoxon signed-rank sum test; p-value significance is indicated. All comparison of cell type abundance between albumin and vehicle were tested for significance by paired Wilcoxon signed-rank sum test, and significant adjusted p-values are indicated.

SUPPLEMENTARY REFERENCES

1. Trebicka J, Fernandez J, Papp M, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. J Hepatol 2020;73:842-54.
2. Moreau R, Jalan R, Ginès P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology 2013;144:1426-37.
3. Arroyo V, Moreau R, Jalan R. Acute-on-Chronic Liver Failure. N Engl J Med 2020;382:2137-45.
4. Clària J, Stauber RE, Coenraad MJ, et al. Systemic inflammation in decompensated cirrhosis: characterization and role in acute-on-chronic liver failure. Hepatology 2016;64:1249-64.
5. Weiss E, de la Grange P, Defaye M, et al. Characterization of Blood Immune Cells in Patients With Decompensated Cirrhosis Including ACLF. Front Immunol 2021;11:619039.
6. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573-3587.
7. Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19: 224.
8. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746.
9. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289-1296.
10. Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796-3798.
11. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766-D773.
12. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
13. Yaari G, Bolen CR, Thakar J, et al. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. Nucleic Acids Res. 2013;**41**:e170.
14. Hagan T, Gerritsen B, Tomalin LE, et al. Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses. Nat Immunol. 2022;23:1788-1798.
15. Aran AP, Looney AP, Liu L, et al. Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163-172.
16. Monaco G, Lee V, Xu W, et al. RNA-Seq Signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. 2019;26:1627-1640.
17. Ellett F, Jalali F, Marand AL, et al. Microfluidic arenas for war games between neutrophils and microbes. Lab Chip. 2019;19:1205-1216.
18. Stewart A, Ng JC, Wallis G, et al. Single-Cell Transcriptomic Analyses Define Distinct Peripheral B Cell Subsets and Discrete Development Pathways. Front Immunol. 2021;12:602539.

Discussion

In this thesis, I employed advanced experimental designs, validated a novel experimental protocol for sample fixation at the time of collection, developed computational approaches to analyze single-cell data from large-scale, decentralized studies, and assessed the effects of specific therapies on the immune system. The main focus of this thesis was to analyze human inflammation-associated profiles from circulating immune cells across multiple diseases and in healthy individuals using single-cell transcriptomics.

To fulfill these objectives, I evaluated the effects of sample fixation on compositional profiles and transcriptomic levels compared to current standards in the field across multiple tissues. Additionally, I created a reference atlas of inflammation from circulating immune cells in a large cohort of patients with a wide range of immune-mediated diseases, as well as healthy individuals. Then, I investigated how the immune response is specifically shaped in each disease condition and how these differences can be captured to identify inflammatory biomarkers. These insights were then exploited to train advanced machine learning models for the development of single-cell based diagnostics, a tool with the potential to advance towards clinical applications. Finally, I evaluated the immunomodulatory role of albumin therapy at the cellular and transcriptomic levels in diseased patients.

In this section, I will review the articles presented in this thesis, discuss their broader applications in the field, and address their limitations. I will first concentrate on the technical challenges of multicenter research projects, then transition to the biological insights, and finally discuss the scientific principles that guided this thesis, while envisioning its future directions.

1. Large-scale research projects: overcoming technical challenges

During my PhD, I have learnt the critical importance of being aware about the limitations and challenges of scRNA-seq experiments. Designing robust experimental protocols in advance is essential to ensure reliable and reproducible data analysis. This approach should consider all stages, including sample collection, preservation, multiplexing, data integration, and cell identity annotation, which will be explored further in this discussion.

1.1. Sample preservation for decentralized studies

The initial step in single-cell experiments involves the critical collection of liquid or solid biologic material from human donors, animal models, or even organoids. These samples must be processed immediately or preserved properly to maintain their viability and characteristics for downstream analysis. Although this process may appear straightforward in theory, it is complicated by the challenges of multi-center research, patient availability, sensitive specimens, and limited resources. In multi-center research studies, samples often need to be collected from

remote areas with limited processing infrastructure and transported to different institutions. Patient recruitment for longitudinal studies or biobanks can require optimal storage conditions to preserve sample integrity. Clinical sample collection depends on procedural requirements and hospital resources. Moreover, decentralized studies involving solid biopsies, where sampling and processing are disconnected, require effective preservation methods to prevent damage or loss of sensitive cells. In fact, mechanical stress and delays in processing can lead to cellular damage and RNA degradation, affecting downstream analysis due to an altered transcriptomic profile and underrepresentation of cells in downstream analysis[156,158,162,238,239]. Cryopreservation, the standard method for preserving samples, can cause alterations based on the processing time and changes in sample composition[158]. To mitigate these difficulties, we developed the FixNCut protocol, which uses the Lomant's reagent (dithio-bis-succinimidyl propionate; DSP) for sample fixation followed by tissue dissociation. As detailed in **Publication 1**, this method captures a snapshot of transcriptome and cellular composition at the time of sampling, while minimizing technical artifacts introduced during processing. Also, FixNCut is compatible with cell membrane labeling with antibodies or LMOs, which are commonly applied to increase cell numbers while reducing technical artifacts and costs.

First, we validated the applicability of the FixNCut protocol for single-cell experiments and microfluidics systems using human blood biopsies. FixNCut preserves RNA integrity, library complexity, cellular phenotyping, composition, and gene expression profiles similarly to fresh biopsies, which are considered the gold-standard, minimizing technical artifacts during sample handling and processing, as described by Baechler *et al*.[240] and Massoni-Badosa *et al.*[158].

Next, we evaluated the FixNCut protocol on solid tissues using lung and colon mouse biopsies. In this context, DSP-fixation enhanced transcriptome capture efficiency and library complexity. DSP penetrates cell membranes and tissues, preserving cellular structure by crosslinking amine groups on polypeptide chains rather than nucleic acids[241]. While its role in mitigating RNA degradation is not fully understood, DSP appears to inhibit endogenous RNAses, which may explain the enhanced transcriptomic profile obtained. This protective effect has already been observed in previous studies using cell lines[241]. In both tissues, the fixed samples showed a shift in cell type composition with large amounts of sensitive cells captured, such as epithelial and stromal cells, with the effect being more pronounced in the colon, a more delicate tissue[238]. The protocol preserves native gene expression, while reducing stress-response gene signatures associated with high temperatures or enzymatical tissue dissociation[156,162,242], in a cell population-specific manner. We hypothesize that this reduction might be even more significant under stressful conditions or improper sample handling. To ensure FixNCut's suitability for complex scenarios, such as long-term sample storage, we combined DSP-fixation with cryopreservation on mouse lung samples, yielding results similar to previous findings. Notably, cryopreserved samples showed upregulation of stress-related genes compared to fixed and

cryopreserved samples, demonstrating that this combination approach maximizes the benefits of both methods. After validating the FixNCut method on mouse models, we tested it on human biopsies from IBD patients. The results once again demonstrated improved data quality compared to the gold-standard of fresh and cryopreserved samples.

The FixNCut protocol proved highly versatile, extending its utility beyond the tissues, species and diseases we have explored. For instance, Aney *et al.* applied FixNCut in scRNA-seq experiments with healthy and pancreatitic mouse models, showcasing its effectiveness in sample preservation[243]. Their work achieved high-resolution insights into pancreatic cell types, comparable to snRNA-seq data from biopsied patients, and facilitated the discovery of novel immune cell states. FixNCut also reduced stress-response gene expression, enhancing data interpretation and reproducibility[243]. Furthermore, other collaborators are currently successfully using FixNCut for scRNA-seq of adult brain samples from mice and samples from Cnidaria, reinforcing the protocol's broad applicability for basic, translational, and clinical research.

Next, we verified FixNCut compatibility with single-cell applications requiring cell membrane labeling (e.g., sample multiplexing, FACS, or CITE-Seq). This was possible thanks to DSP's crosslinking being reversible, though recent studies suggest that de-crosslinking could lead to cell loss[244]. This represents an improvement over other fixatives that cross-link RNA, DNA and proteins indiscriminately causing protein unfolding and precipitation, hindering cell labeling before single-cell analysis[160]. However, antibody concentrations should be optimized to maintain signal strength in DSP-fixed cells. Although this manuscript did not explore other single-cell applications involving TCR and BCR cell profiling, based on the success of Nesterenko *et al.* in using CLInt-seq technology for V(D)J sequencing on DSP-fixed blood samples[245], we anticipate that FixNCut could be compatible to profile immune receptors in tissue.

Lastly, we validated DSP fixation as an alternative to formalin fixation for spatial proteomics with a human prostate paraffin-embedded tissue section using the Phenocycler. Based on these results, FixNCut protocol could also yield promising outcomes for spatial transcriptomics and proteomics analysis on tissue sections, as we have shown that DSP fixation preserves tissue morphology as well as RNA and protein integrity, which are critical for high-quality spatial molecular profiling.

This study has delivered actionable protocols for the single-cell genomics community. The protocol, titled **FixNCut v1.0,** is publicly available on the protocols.io website [https://dx.doi.org/10.17504/protocols.io.14egn3xjql5d/v1] and has been published in a protocol-centric journal, included in **Appendix A** (**FixNCut: a Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assay**) of this thesis.

1.2. Hybrid sample demultiplexing: combining cell-tagging and SNP-based deconvolution methods

In single-cell genomics, as the number of samples and cells analyzed grows, new experimental design strategies are needed to reduce batch effects, enhance data robustness, and minimize costs. One effective approach is pooling multiple samples into the same library for simultaneous processing. Cell-tagging methods, while widely used, present several drawbacks. They may not be suitable for samples with low input numbers due to potential cell loss from washing steps, and inadequate washing can leave antibodies or free oligos in suspension, increasing the background noise. Additionally, suboptimal hashtag binding can lead to poorly stained cells and unreliable sample identification, while using multiple hashtags can result in higher false doublet rates. Altogether, these issues can lead to false doublets or unlabelled cells, impairing accurate cell identification and reducing overall cell recovery rates[164,246,247]. SNP-based multiplexing offers high precision and up to 90% cell recovery, but its effectiveness can be compromised by poor transcriptome quality (e.g., low coverage or ambient RNA), or abnormal gene expression profiles from dying cells. Furthermore, it requires samples with diverse genetic backgrounds, limiting its application in most animal or longitudinal human studies[248,249]. Also, the need for individual patient genotype references increases costs.

At the time of generating our datasets, over ten deconvolution methods (e.g., *HTODemux, HashSolo,* or CellRanger for barcoded-based, and Vireo, Demuxlet, or Souporcell for SNP-based)[168,177,249–251] were available, along with various benchmarking studies[246,252]. However, no existing methods combined both cell-tagging and SNP-based algorithms to address these challenges. By integrating these two approaches, the hybrid strategy could leverage the strengths of the individual methods to counterbalance each technique's weaknesses. To fill this gap, I conceptualized and developed a novel strategy for sample demultiplexing, which was applied to all in-house human samples processed using the CellPlex strategy, and included in the **Publication 2** (see Methods)**.** This hybrid strategy allows samples to be stained and pooled together in absence of reference genotype data, with cells later deconvoluted based on the learnt genotype profiles. Pools of unknown singlet cells are subsequently traced back to their donor of origin based on the barcoded cell groups overlapping with each genotype-based cell group. This combined approach overcomes issues such as poorly stained cells or multiple hashtag contamination, improving cell recovery rates and accuracy in cell identification, all without patient genotype reference data.

To demonstrate the effectiveness of the hybrid approach applied to in-house generated samples included in **Publication 2**, I will present a case study. Specifically, the scRNA-seq experiment involved pooling 8 samples, including 6 COPD patients and 2 matched healthy controls, using the CellPlex strategy (CMO-tagged) in absence of genotype data. The CMO-tagged approach

successfully recovered 64.14% of singlets (38,921 out of 60,679) (Figure 12a). Then, I compared donor identification between the hashtag and genotyping methods, and found that 98.58% (38,368 out of 38,921) of singlets identified by CellRanger using CMO-tags were also recognized as singlets by Vireo (genotypes), confirming the accuracy of both methods. Interestingly, each unknown donor group identified through genotyping corresponded to a single CMO group with high correlation (0.73 to 0.77) (Figure 12b). This correlation enabled mapping of unknown donor pools to hashtags, revealing true donor identities of genotype-based clusters, and recovering 92.2% of cells with undetectable hashing (i.e., negatives: unassigned or blanks). Additionally, 31.9% of originally misclassified doublet cells, likely due to low hashing efficacy, background noise, or a high false doublet rate were recovered (Figure 12a). Overall, the hybrid strategy combining barcode- and genotype-based approach, achieved a cell recovery rate of 83.50% (50,664 out of 60,679) (Figure 12a).

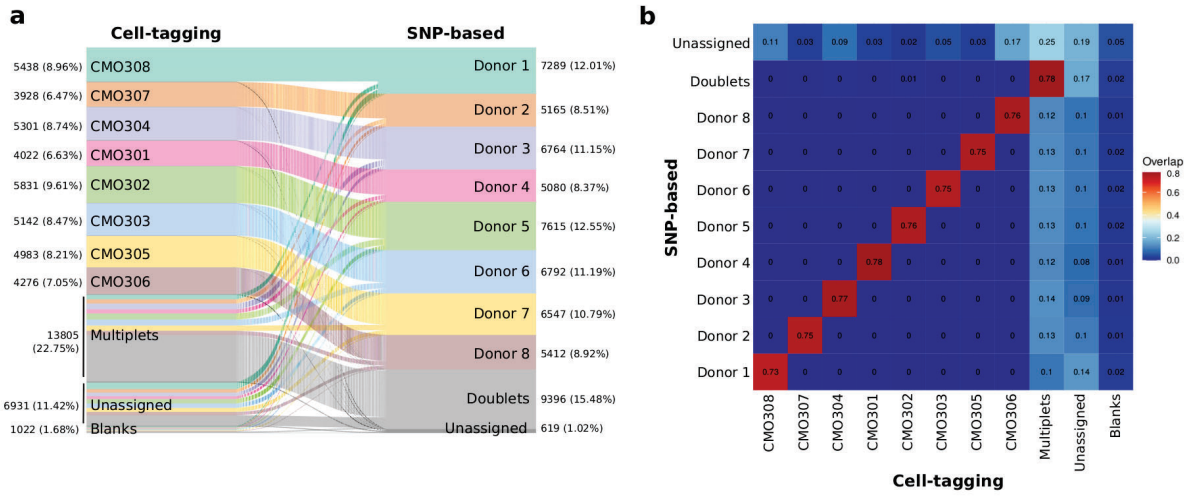


Figure 12. Hybrid demultiplexing strategy using cell-tagging and SNP-based algorithms. (**a**) Alluvial plot showing donor deconvolution results using cell-tagging (**left**) and SNP-based (**right**) approaches. Color indicates SNP-based unknown sample grouping. The plot also displays cell numbers and percentages for each category. (**b**) Correlation heatmap comparing cells classified by CMO-tagged groupings (using CellPlex) with unknown genotype groupings (using cellSNP-lite followed by Vireo).

To evaluate the quality and biological relevance of cells rescued by the hybrid strategy, I jointly analyzed the count depth and transcriptomic complexity of these cells (Figure 13a,b). Cells initially classified as doublets or unassigned based on barcodes showed high total genes and overall counts. In contrast, SNP-based classification revealed that real unclassified cells had low mRNA content and fewer genes, indicating poor-quality cells, whereas doublet-classified droplets exhibited complex transcriptomic profiles with high molecule and gene counts.

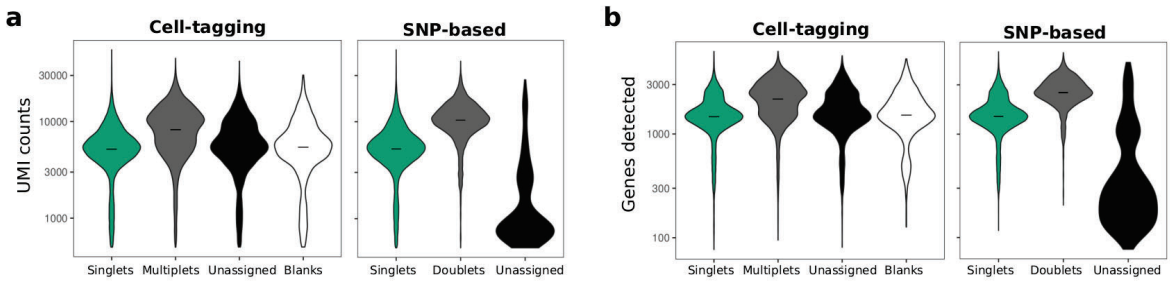


Figure 13. Comparison of deconvoluted cell quality between cell-tagging and SNP-based (or hybrid) algorithms. (**a**) Violin plots showing the UMI count distribution of deconvoluted cells for cell-tagging (**left**) and SNP-based (**right**) grouping. (**b**) Violin plots showing the distribution of detected genes in deconvoluted cells for cell-tagging (**left**) and SNP-based (**right**) grouping. In both panels, deconvoluted cells include singlets (from all individual donors identified), multiplets/doublets, unassigned cells or blanks.

During the writing of this thesis, new methods implementing similar hybrid strategies were released. Curion *et al.* developed a Nextflow pipeline called "hadge" to identify the best demultiplexing method by evaluating existing hashing and genetic-based methods and determining the optimal hybrid deconvolution strategy using Pearson correlation scores between the results from SNP-based and cell-tagging methods. They achieved excellent results, recovering 89.7% of barcoded-based negative cells, which remain high-quality for downstream analysis[252]. In addition, Li *et al*. published a benchmarking study confirming the benefits of hybrid demultiplexing; single-cell hashing groups consistently correlated with genotype clusters, accurately identifying sample origins, and achieved nearly 90% cell recovery from 8-pooled samples, in line with our results. They also demonstrated that hybrid demultiplexing improves performance in samples with low cell counts and reduces costs by minimizing the amount of hashtag/antibody stained cells per sample (i.e., staining only a small fraction)[253].

In conclusion, the combined use of barcoded-based and genetic deconvolution methods effectively addresses common errors in cell hashing demultiplexing, enhancing both cell recovery rates and sample identification accuracy. This dual approach is crucial for experiments where cell-tagging data might be affected by technical noise or tissue variability, enhancing flexibility, performance, and robustness while increasing the number of high-quality singlets for further analysis. Although the strategy was initially developed for this specific project, its successful application and benefits have motivated me to develop a pipeline for internal use.

1.3. Impact of data processing

Transforming the count matrices into meaningful biological insights requires a deep understanding of the data and often involves navigating cycles of trial and error to fully harness its potential. This section outlines the key steps and methods for data processing and cell identity annotation building on my experience with large, complex datasets during my thesis. I will also discuss challenges not detailed in **Publications 1, 2** and **3**, and how they were tackled.

Impact of sequencing depth

In **Publication 1**, while assessing the FixNCut protocol, I encountered the challenge of samples processed in different libraries with varying sequencing depths. To address this, we used the *CellRanger aggr* tool, which standardized sequencing depth. This ensured consistency across samples, allowing any observed differences to be attributed to the DSP-fixation effect rather than variations in sequencing depth. Sequencing depth is critical in single-cell analysis, as it directly affects data resolution, technical noise, gene detection sensitivity, and accurate gene expression quantification (i.e., with shallow sequencing favors highly expressed genes) and cell type distinction[254,255]. A minimum of 20,000 read pairs per cell is recommended for gene expression libraries[145], but the optimal depth depends on the experimental goals and sample type (e.g., capturing rare genes or maximizing the number of cells)[254,256]. Maintaining consistent depth across datasets is essential for reliable comparisons, but experimental limitations often prevent this, necessitating post-hoc computational adjustments like the approach we applied.

In **Publication 2**, while building the inflammation atlas, we faced a similar challenge, but some datasets lacked raw sequencing reads, making it impossible to standardize sequencing depth with the same method. However, by analyzing multiple libraries across conditions (i.e., a patient is sequenced in different 10X Genomics GEM channels) and elevated number of patients and cells, we reasoned that this would balance potential effects of uneven depth, minimizing its impact on downstream results.

In **Publication 3**, we did not encounter this problem. Here, samples from blood co-cultured with either albumin or vehicle from the same donor were multiplexed in the same 10X Genomics GEM channel, eliminating the batch effect on paired samples for condition assessment.

Influence of genomic references versions

While constructing the inflammation atlas presented in **Publication 2**, we recognized that the choice of genomic reference, the mapping approach, and the gene naming could introduce a batch effect at the beginning of the analysis. Variations in reference versions include differences in gene annotations and identifiers[257]. Also, the decision to include intronic regions during mapping could significantly impact gene expression quantification[145]. However, whether the observed effects were due to disease-specific factors or attributable to the mapping process, as well as the extent to which these variations impact downstream analysis, remains unclear. To investigate this, I evaluated the impact of different genomic reference versions (GRCh37/Ensembl87 or GRCh38/Ensembl98), mapping choices (including or excluding introns), and gene naming conventions (Ensembl identifier or symbol) using a single 5' scRNA-seq sample from a BRCA patient (Table 1), included in **Publication 2,** before scaling up to large-scale studies.

Table 1. Main library metrics for the 5' scRNA-seq BRCA sample. Metrics from the report provided by CellRanger (v6.0) after mapping using different genomic references, with or without the inclusion of introns.

	Total cells captured	Total genes detected	Median genes / cell	Median UMI counts / cell
GRCh37/Ensmbl87	6,091	20,058	1,399	3,681
GRCh38/Ensmbl98 without introns	6,107	21,996	1,445	3,774
GRCh38/Ensmbl98 with introns	6,267	25,175	1,760	4,289

First, I compared the latest human genomic reference (GRCh38) used for mapping with CellRanger with a prior reference (GRCh37), both without considering introns. The newer version detected more total genes, with higher median reads and UMI counts per cell (Table 1). This improvement is due to CellRanger algorithms which, by refining expression detection, may capture more cells or exclude some. After standardizing gene names using Ensembl identifiers, we detected 28,374 unique genes, with 16,265 (57,32%) shared between both versions, while Ensembl98 recovered 44% more exclusive genes. Most shared genes between versions (98.7%) had official Human Genome Organisation (HUGO) names [National Human Genome Research Institute (NHGRI); https://www.genenames.org/; Feb 2024], while only 25.5% and 6.3% of version-exclusive genes for Ensembl98 and Ensembl87, respectively, did. Understanding gene function is crucial for elucidating cellular behavior, intracellular pathways, and cell-to-cell interactions. Therefore, we aim to uncover the roles of both HUGO shared and version-exclusive genes. Common genes were predominantly protein-coding (94%), while exclusive genes were spread across non-coding RNA, pseudogenes, and other categories. Next, I compared gene quantification levels for shared genes to assess any impact. Due to data sparsity and high dropout rates, I used the Mean Squared Residues (MSR) metric to calculate the average squared differences between values from both versions, with values close to 0 indicating high similarity. We obtained an MSR of 0.035. However, considering only genes detected with at least 1 count increased the MSR to 0.302, highlighting discrepancies between data mapped with different references that could impact downstream analysis.

Next, using the latest GRCh38/Ensembl98 version, I assessed the impact of including introns while mapping. Including introns substantially increased the total number of genes detected, the counts assigned to each gene, along with larger median reads and UMI counts per cell, and captured more cells (Table 1). The total variety of harmonized genes captured was 26,759, with 86.88% shared between both conditions, indicating the inclusion of introns did not significantly bias the genes recovered that much. The MSR filtered was 0.557, which is higher than previous comparisons. This suggests that including introns captures more gene information, such as nascent transcripts like pre-mRNA, which strongly affects gene quantification.

In summary, since the same sequenced sample was used across different mapping strategies, sequencing depth cannot be a factor in the observed differences. Different reference genomic versions significantly impact the genes recovered, leading to inconsistencies in gene names and annotation. Also, the annotation file version might result in genes with multiple aliases or deprecated symbols. Additionally, including introns affects gene expression quantification, making downstream corrections more challenging. Considering these results, to avoid batch effects from discrepancies in gene annotations or quantification for datasets used in **Publication 2**, we unified the mapping process across all datasets. Although including introns offered greater resolution, it was not feasible for some external datasets due to raw sequencing reads not being publicly available, and since those were mapped without intron inclusion, could exacerbate differences. Consequently, we mapped all external data using available FASTQ data with the same human genomic reference (excluding introns) and CellRanger version as for the in-house data. This approach, along with using gene Ensembl identifiers instead of symbols, minimizes discrepancies. Many publicly available datasets provide only count matrices with gene symbols, which represent a challenge when analyzing them alongside other datasets. To address this, in cases where sequencing reads were unavailable, we used the HUGO database to align gene identifiers, improving mappability and ensuring consistent identifiers across datasets. Given the minimal differences observed in gene expression quantification, the impact was considered small. Additionally, scRNA-seq analysis often reduces data magnitude and sparsity by removing genes expressed at minimal levels in few cells and with HVG selection. This data processing step mitigates the impact in large-scale studies where common pre-processing was not feasible.

Addressing batch effects: pre- and post- cell identity annotation

During the construction of the inflammation atlas in **Publication 2**, which included both in-house and public datasets to achieve a diverse cohort of patients and immune-driven diseases, we observed significant batch effects. Those arose from the use of different 10X Genomics chemistry and assays, along with inconsistent disease representation across chemistries, varying numbers of patients, and differences in patient age, sex, and disease status. The lack of clinical details in public data further complicated stratification. These issues affected data processing steps like clustering, raising concerns about reproducibility and validity of biological findings.

Addressing batch effects before cell identity annotation involves integrating data from multiple datasets to identify common transcriptional profiles linked to well-known cell types. We used scVI[211], a variational autoencoder approach, to mitigate batch effects caused by technical factors, mainly the 10X chemistry. scVI was preferred over other methods due to its effectiveness in complex scenarios, such as large-atlas projects with millions of cells and missing annotation information[213]. We also applied a tailored feature selection step to reduce the influence of specific patient or study factors (e.g., mitochondrial/ribosomal content, TCR/BCR and MHC genes) and

retained only HVGs present in multiple studies. This process led to the accurate annotation of clusters into 64 cell populations, which were then grouped into 15 major cell populations.

In contrast, correcting batch effects after atlas creation focuses on making data comparable to detect subtle differences in gene expression, signatures, and cellular interactions between conditions[213]. In our project, after annotating the data, we aimed to address batch effects while preserving disease-related biological heterogeneity, making gene expression comparable, and handling missing or noisy observations. To tackle these challenges, we used scANVI, an extension of the scVI model[207]. scANVI leverages scVI pre-trained embedding space and cell annotations to refine batch correction, producing a new low-dimensional embedding[207]. Its negative binomial posterior model allows for the generation of a normalized, corrected count matrix, where batch effects are mitigated by sampling and averaging each cell's expression as if it originated from each batch chemistry covariate. Due to the high computational resources needed for this step, we could not analyze the entire gene universe. To avoid missing crucial genes for delineating the inflammatory response, we employed a combinational gene selection approach. After removing unwanted genes, we combined HVGs, genes differentially expressed between healthy and diseased samples (corrected for chemistry, sex and age at broad cell-type annotation), and a curated list of immune-specific genes.

In the single-cell field, there is ongoing debate about using batch-corrected data for cross-batch comparisons[258]. Batch correction can sometimes inadvertently remove biological differences between conditions, particularly when a specific condition introduces a systematic difference across all cells (e.g., knock-out samples), which may be misinterpreted as a batch effect during alignment. Additionally, some data integration methods assume that differences between samples are orthogonal to within-sample variation, making these methods questionable when comparing samples with similar transcriptional profiles from different batches[259]. In our project, which compares cellular profiles of pathological conditions with healthy ones, we anticipated that biological differences would uniquely affect certain cell populations across diseases. Therefore, we did not expect batch correction to eliminate the inherent biological heterogeneity. To verify this, we evaluated different data integration methods, as done by a recent benchmarking study[213], and then compared inflammatory signatures computed using data from the corrected count matrix with those from the uncorrected count matrix (Figure 14). Finally, we evaluated the performance of assigning each cell to its correct disease label and observed improved accuracy when using corrected counts rather than log-normalized counts. These comparisons confirmed that we successfully corrected the batch effect introduced by the experimental design while preserving biological variability. Consequently, the corrected count matrix was used for subsequent analyses, enabling exploration of the inflammatory profile across diseases and enhancing biological insights and data comparability. However, for comparisons within the same study, the uncorrected count matrix was used to maintain consistency with the original data.

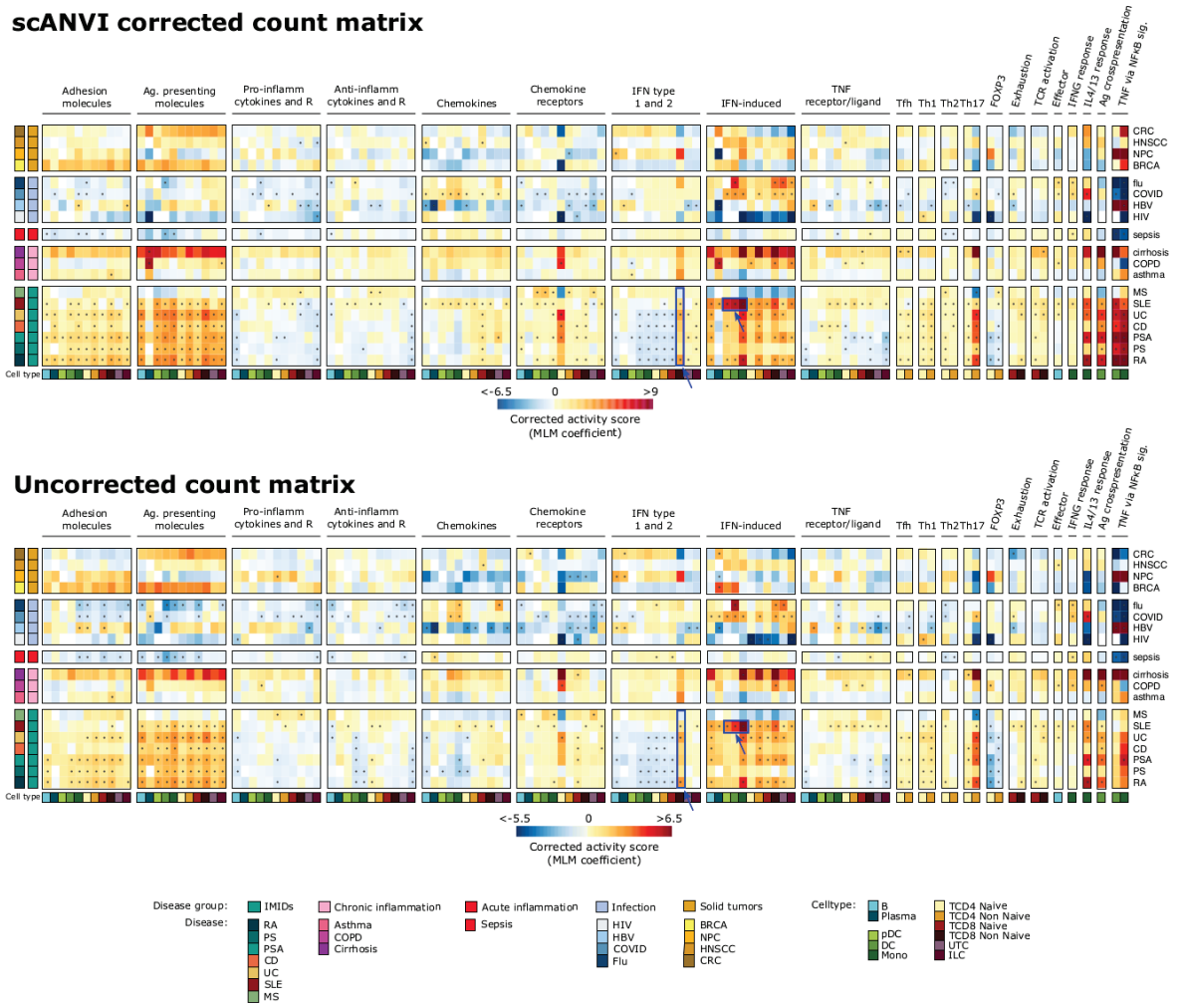


Figure 14. Heatmap displaying inflammatory signature activity scores across diseases and cell types. Data are derived from the patient and cell-type pseudobulk based on the scANVI corrected count matrix (**top**) or the uncorrected count matrix (**bottom**). The *x-axis* represents the SPECTRA cell type-specific factor associated with a given function. The *y-axis* represents the diseases grouped by disease group. Significant signature activity differences between disease and healthy controls (HC) are marked with a dot (·) (p-value Adj < 0.05). Arrows highlight significant IFN response and IFN Type 1 and 2 activation (signatures highlighted in the manuscript) with both data inputs.

Learning the importance of benchmarking potential confounders and evaluating available methods and pipelines while developing new strategies to address challenges presented in scRNA-seq experiments, is particularly important in large-scale to produce reliable and reproducible data while obtaining meaningful biological insights. These lessons have been applied to the inflammation atlas project (**Publication 2**), addressing technological limitations associated with integrating extensive datasets and effectively leveraging the underlying molecular and cellular mechanisms orchestrating inflammation.

2. Characterizing disease inflammatory profile

As our understanding of inflammation advances, it enhances our insight into a wide range of diseases. Since each disease targets specific tissues and exhibits unique characteristics, studying and comparing multiple diseases requires a broad approach. Blood proves to be a valuable, minimally invasive medium for dynamically monitoring inflammation by measuring cytokine levels, cellular extravasation, antigen presentation, and cell activation[129,130]. Single-cell sequencing has been widely applied to investigate several inflammatory diseases, providing insights into changes in cell type composition and helping to identify disease-driving mechanisms that could serve as therapeutic targets[260,261]. Despite these efforts, a map of immune cell states, capturing the spectrum of immune plasticity across inflammatory diseases remains lacking.

In **Publication 2**, we developed a comprehensive resource for studying the inflammation landscape of circulating immune cells in both disease and health at single-cell resolution. We analyzed the transcriptomic profile of over 6.5 million peripheral blood mononuclear cells (PBMC) from 1,057 patients across 19 diseases, categorized into immune-mediated inflammatory diseases, acute and chronic inflammation, infections, solid tumors, and healthy controls. By assessing the immune cell composition and inflammatory molecule expression in PBMCs, we identified commonalities and differences across diseases, potential biomarkers for specific diseases, and proposed a framework to stratify patients based on their inflammatory profile. This study demonstrates the potential for future precision medicine tools to improve the diagnosis and management of patients with unknown inflammatory profiles in clinical settings.

2.1. Inflammation signatures across diseases and cell types

To capture inflammatory profiles across diseases and cell types -such as immune cell activation, cytokine-related responses, cellular migration and extravasation, antigen presentation, and signaling pathways- we initially proposed a biology-driven approach based on the assessment of signatures that encompass inflammatory-related functions using curated molecules (Sup. Table 4 - Publication 2). In addition to those general processes guiding inflammation, we also included a list of cell-type-specific signatures derived from single-cell studies via Cytopus[262]. However, limiting the analysis to what is already known may restrict the discovery of novel insights. Therefore, we applied a “semi” data-driven approach: starting with known biology, we modified the gene list of each signature using data-driven methods (i.e., Spectra[263]), which provided a list of cell-type-specific factors, representing refined gene sets. This strategy enables the addition or removal of genes that may influence inflammation, even if they are not directly linked to a well-defined immune function, as well as the exclusion of genes that are either not captured or not expressed in this dataset without considering disease context. Most importantly, Spectra operates at the cell-type level, tailoring each gene set signature to a specific cell population.

These results were then used to assess inflammation activity scores, highlighting which inflammatory signatures are up- or down-regulated in specific conditions and cell types.

To our knowledge, this is the first study that explores the inflammatory profile across diseases stratified by cell populations. The feasibility to apply linear models at cell population level is supported by the use of single-cell resolution datasets from a cohort of hundreds of patients included in the atlas. This approach is particularly relevant, as cell composition varies with disease conditions, which can significantly impact results and introduce bias towards the most common cell population. By training the classifier by major cell types, we accurately captured changes in the inflammatory profile, minimizing the risk of missing important signals unique to rare cell types. In general, we observed an increased inflammation activity in disease compared to healthy status, with specific depletions that highlight the importance of inflammation in driving immune-related diseases. By capturing well-described inflammatory signals, such as the upregulation of IFN-induced signature in SLE across all immune cell types -particularly within the myeloid compartment-[264,265] or the activation of the NF-κB signaling pathway in myeloid cells[266] across all IMIDs, we provide strong biological validation that our approach is robust enough to capture inflammation signals that are not only disease-specific but also linked to distinct cell types. For example, the upregulation patterns observed for IFN types 1 and 2 across all non-naive CD8 T cell populations in IMIDs were driven by specific effector-memory CD8 T cell subsets. Additionally, we found that the *FGFBP2* gene*,* a key gene defining this signature, was overexpressed across all CD8 effector subtypes. Interestingly, this gene has been identified in IBD as a marker of CD8 T cells in areas of epithelial damage in response to DC signaling[267] and is used to characterize a specific CD8 effector T cell population[268]. Thus, working with single-cell data allowed us to deepen these signatures and gene sets that compose them in detail, and their role in specific subpopulations or in distinct cellular states, offering a higher level of granularity.

This strategy enables us to uncover previously undescribed inflammation signals and distinguish between diseases and cell types based on their unique inflammatory profiles. We propose extending this approach to explore inflammation signatures across a wider range of diseases and biological contexts, providing valuable insights into the overall inflammatory status.

2.2. Biomarker discovery through interpretable modeling

To investigate inflammatory patterns underlying diseases beyond the aforementioned signals, alternative approaches are needed. In this study, we aimed to identify disease-discriminatory genes that effectively differentiate one disease from all other conditions, while considering gene-gene interactions with nonlinear relationships at cell-type level. Machine learning (ML) methods, such as Gradient Boosted Decision Trees (GBDTs), are effective at handling complex tasks with noisy data and nonlinear feature dependencies[269], making them well-suited for single-cell analysis. We applied XGBoost[270], a data-driven ML method, to learn inflammatory

patterns across diseases. Although we validated the model performance in a left-out set, containing unseen patients, the trained models were not inherently interpretable. To address this, we used post-hoc methods like SHapley Additive exPlanation (SHAP)[235] to improve model interpretability and extract biological insights, aiding the identification of potential clinical biomarkers.

By applying SHAP to the GBDT-trained models, we successfully ranked genes based on their discriminative power between diseases at cell type level, minimizing the influence of cell-specific expression profiles and allowing a more precise examination of gene importance for differentiating diseases. To provide a measure of how our biomarkers generalize to unseen studies, we included XGBoost trained on the whole set of genes from external data, demonstrating that SHAP-based gene ranking provides more effective feature selection than random gene sets for predicting disease labels for the majority of cell types. Further validation of this approach was done by finding previously identified biomarkers, such *STAT3* for RA disease specifically in CD4 T cells[271], to also be relevant for disease discrimination when considering interactions with other genes.

This machine learning strategy using GBDT along with SHAP yielded explainable results and identified a list of genes with high discriminative power between diseases. Notably, we identified *CYBA* as an important marker in monocytes for classifying diseases that affect barrier tissues. Here, high levels of *CYBA* expression led the model to classify cells towards intestinal (UC and CD) and pulmonary inflammatory diseases (asthma and COPD), whereas low levels were linked to skin-related diseases (PS and PsA). Using scRNA-seq from an external cohort, we observed *CYBA* overexpression in monocytes/macrophages in patients with IBD (UC and CD) compared to healthy individuals, not only in PBMCs[272], but also in colon tissue biopsies[268]. However, considering that disease-relevant genes were identified by SHAP using a non-linear method that takes gene interactions into account, validating these genes as definitive biomarkers for clinical diagnoses remains challenging and requires further gene-specific evaluation in independent, large patient cohorts.

2.3. Patient stratification based on inflammatory profiles

In clinical practice, some diseases can be easily diagnosed using a single biomarker, such as measuring levels of a biochemical molecule[273], analyzing proportions of specific cell populations[274], or assessing gene mutations[275]. However, the majority of inflammation-driven diseases are far more complex, arising from a combination of multiple factors. In this context, the inflammation immune cell spectrum captured in the atlas, combined with state-of-the-art machine learning techniques, enabled us to use cell embeddings to classify patients according to their disease of origin. To accomplish this, we first performed data integration to obtain batch corrected embeddings, averaged embedded dimensions considering cells belonging to each cell

type independently, and then trained a classifier for each cell type to predict the patient's condition. We then used the reference embedding to map unseen patient's cells, transfer annotations, and obtain a common embedded space, which was subsequently averaged by cell type. Finally, we predicted inflammatory conditions of unseen patients using the trained models and aggregating the results from all cell type-specific classifiers with a majority voting strategy.

We validated our framework using five-fold cross-validation on the Inflammation Atlas data (Main) by testing on random sampled patients from studies included in the atlas that were removed before performing cell annotations (Validation), yielding very good classification performances. These findings suggest that gene expression profiles from blood biopsies, encoded in a low-dimensional space, efficiently capture the underlying inflammation profile of diseases and distinguish differences across them. Notably, disease prediction varied across cell types, with certain cells better capturing inflammatory profiles for specific diseases. This supports our approach of training a classifier per cell type and combining results via majority voting instead of using a single classifier approach. However, when evaluating our pipeline on unseen patients from external studies, classification performances decreased significantly. This decline is not caused by residual batch effects, which were mitigated by adjusting the integration approach and optimizing data integration parameters, but can attributed to a limited sample size, poor stratification (e.g., neglecting covariates like gender, age or comorbidities), and the underrepresentation of certain diseases. Expanding the patient cohort in a more balanced way, with adequate representation across diseases and conditions, could improve data integration, reduce noise, and enhance the biological signal.

In conclusion, this framework captures the inflammation profile across various diseases and demonstrates potential to be used as a precision diagnostic tool for early disease prediction using blood biopsies. By bridging experimental and computational research with clinical application, this approach supports the advancement of personalized medicine tailored to individual patient needs.

3. The immunomodulatory effects of albumin in acute decompensated cirrhosis

The underlying molecular and cellular mechanisms of therapeutic approaches that modulate the immune system in the context of immune-driven diseases can be complex. However, single-cell sequencing technologies enable a detailed examination of treatment effects at individual cell level, helping us to understand how these changes contribute to overall disease outcomes.

In **Publication 3**, we investigated the immunomodulatory action of human serum albumin (HSA) in patients with acute decompensated (AD) cirrhosis, characterized by severe lymphopenia and defective neutrophil antimicrobial function. Single-cell and bulk RNA-sequencing of PBMC from AD patients incubated with albumin revealed changes in gene expression levels and in lymphocytes, particularly B cells, compared to healthy volunteers (HV). Additionally, neutrophils exposed to albumin also exhibited increased chemotaxis, degranulation, and phagocytosis. Finally, validation in patients receiving albumin revealed a similar immunomodulation profile. Overall, we demonstrated that albumin modules the immune system by promoting transcriptional reprogramming, modifying immune cell composition, and revitalizing neutrophil antimicrobial functions in patients with AD cirrhosis.

We began by characterizing the immune cell landscape in patients with AD cirrhosis, confirming neutrophilia, lymphopenia (affecting B cells, CD8+T cells, and NK cells), and monocytosis[276]. To investigate the specific effect of HSA on immune cells, we performed *in vitro* experiments on patient PBMCs. Bulk RNA-seq data revealed transcriptional changes under HSA treatment, notably upregulation of signatures linked to B cell immunity and antimicrobial responses. Next, scRNA-seq provided deeper insights into cell populations, showing that HSA reprograms naive B cells towards a transitional-like profile, findings validated by flow cytometry. Due to limitations of scRNA-seq for neutrophil analysis[179], we employed functional assays, confirming that HSA restored impaired neutrophil activity in cultures from AD patients. Building on these *in vitro* analysis, we extended our research to a longitudinal study of AD patients under HSA therapy that develop ACLF, where clinical outcomes aligned with our *in vitro* observations. Interestingly, since the aim of the study was to elucidate the mechanism of action of albumin, it was crucial to design experiments that controlled for potential confounders (e.g., *in vitro* studies to avoid the effect of co-administered treatments) and to conduct parallel experiments to discard alternative explanations (e.g., osmotic stress effects, induction of NF-kβ signaling, extracellular acidification)[277,278]. Overall, this exemplifies a robust research workflow that proceeds from disease characterization to mechanistic *in vitro* studies, using diverse methodologies, and ultimately to translational applications in clinical settings.

Despite the strengths mentioned above, this study faced some limitations due to the controversies surrounding the clinical pathophysiology of the disease (e.g., clinical characteristics, biochemical parameters, cytokine levels, albumin dosage, and treatment duration), which affected subsequent patient stratification. In addition, a limitation of scRNA-seq experiments is the need to prioritize either high-resolution annotation of populations to capture rare cell types, which often results in low cell numbers and high dropout rates. On the other hand, working at a lower resolution with larger cell populations allows for better assessment of transcriptional differences. This experimental choice depends on the specific study objectives, with the advantage that both strategies can be combined to maximize insights.

4. Core scientific principles

Core scientific principles extend beyond experimental rigor, highlighting the importance of collaboration, transparency, and data sharing in current research. Throughout my PhD thesis, I have actively embraced these principles by adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) data guidelines and promoting open dissemination. In this section, I will detail how I implemented these practices to facilitate data sharing and ensure replicability across the scientific community, and contribute to the broader dissemination of knowledge.

4.1. Fostering Collaboration

The Human Cell Atlas (HCA) [https://www.humancellatlas.org/] is a global initiative aimed at comprehensive 3D reference maps of cells in the human body, across key organs and tissues to enhance our understanding of human biology in both health and disease and improve disease diagnosis, monitoring, and treatment. This ambitious large-scale project exemplifies the power of collaborative efforts across a wide range of disciplines, accelerating research through the combined expertise of clinicians, wet lab researchers and computational analysts.

Being part of the HCA consortium has given me valuable opportunities to collaborate with international research groups, highlighting the power of interdisciplinary teamwork. My involvement in building the inflammation atlas, as detailed in **Publication 2**, allowed me to experience firsthand the importance of coordinated teamwork across all stages, from the bench to bioinformatics, in advancing scientific discovery.

During my PhD, I also collaborated with the Colorectal Cancer Lab, led by Dr. Eduard Batlle at the Institute for Research in Biomedicine (IRB), contributing to two publications detailed in **Appendix B** titled "**Metastatic recurrence in colorectal cancer arises from residual *EMP1*+ cells**" and **Appendix C** titled **"*Mex3a* marks drug-tolerant persistent colorectal cancer cells that mediate relapse after chemotherapy"**.

4.2. FAIR data and Open Code

In a collaborative research environment involving multiple researchers and groups, organization and reproducibility are crucial. Following FAIR principles and Open Code practices ensures transparency, long-term usability, promotes data and code reuse, and saves time and costs. Effective project management tools, such as those for tracking progress and assigning tasks, also enhance clarity and structure. These practices ensure research is well-organized, and impactful. Following this, I will detail how protocols, code, and data were managed in line with these principles during my thesis.

- **Protocols**: Standardized lab protocols enable accurate replication of experiments across teams and institutions. To support the single-cell genomics community, the FixNCut protocol presented in **Publication 1** has been made publicly accessible via protocols.io [https://dx.doi.org/10.17504/protocols.io.14egn3xjql5d/v1] and published in a protocol journal, detailed in **Appendix A** (**FixNCut: a Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assay**) of this thesis.

- **Code**: An end-to-end pipeline for processing data, analyzing results, and generating figures is crucial for research reproducibility, and sharing this code with clear documentation fosters openness and transparency. This can be achieved with tools like GitHub, which also support version control and manage contribution. The code from this thesis is stored in GitHub repositories, with links provided in their respective publications.

- **Raw Data**: While sharing raw data is valuable, privacy and consent concerns, especially with human genetic data, often limit full public access. To balance these concerns with the need for transparency, genomic data generated during this thesis were anonymized and deposited in the Sequence Read Archive (SRA) [https://www.ncbi.nlm.nih.gov/sra], NCBI Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo/], or European Genome Archive (EGA) [https://ega-archive.org], depending on the privacy levels established by the respective ethical committees.

- **Processed Objects:** Sharing processed data or curated datasets enhances reproducibility and data reuse when raw data sharing is restricted. The processed data used in this thesis are available on Zenodo [https://zenodo.org/], offering a controlled access point that respects privacy and consent constraints.

- Sharing **preprints** speed up the dissemination of research, enabling preliminary findings to reach the community quickly. **Open access articles** then ensure that peer-reviewed findings are available to a broader audience. In this PhD thesis, we have disseminated preliminary findings as preprints and ensured the articles are open access.

4.3. Widespread Science Dissemination

Bridging the gap between specialized research and the public understanding requires making science both approachable and comprehensible. During my PhD, I engaged in multiple dissemination activities to share discoveries with the scientific community and the broader public, such as delivering talks and presentations to share findings and insights with diverse audiences, inspiring interest in STEM (Science, Technology, Engineering, and Mathematics) careers in young students, organizing and participating in workshops to facilitate knowledge exchange, publishing research articles in preprint repositories and open-access journals to ensure broad accessibility, and utilizing digital platforms and tools for efficient code and data sharing.

5. Shaping the future

The future of scientific research and technology is unpredictable, but by analyzing current trends and potential developments in both technological and computational domains, we can envision how these advancements could be applied and enhance the single-cell genomics field.

On one hand, with decreasing costs of sequencing, many researchers may begin generating their own data. To take full advantage of this opportunity, it is essential that samples are processed in a standardized manner to minimize bias and reduce batch effects. In this context, developing protocols similar to FixNCut, which standardize sample preservation and processing, will be crucial to ensure data reliability and reduce variability across datasets. This is especially important in large-scale studies, such as the Inflammation Atlas, to ensure high data quality and consistency while enabling meaningful comparisons across datasets and research centers.

On the other, as the volume and complexity of data grows, integrating data from different modalities and multiple sources requires new approaches and the development of innovative methods. Here, there is a duality in the use of simpler, explainable models, and more complex models, often opaque, like deep learning. Explainable models, such as linear models, are easily interpretable, making them valuable in biological contexts where understanding the mechanisms driving diseases or identifying actionable targets is crucial. For instance, in tasks such as automatic cell annotation using a reference with few condition labels, simpler methods often perform just as well, if not better. However, these models might struggle to capture the complexity of living organisms, where gene interactions are multifactorial and nonlinear. In contrast, deep learning methods are designed to handle high-dimensional data and, despite their lack interpretability and tendency to overfit, can achieve higher predictive accuracy, making them suitable for clinical decision-making. For instance, deep learning (DL) models from other fields, such as transformers[236], are now being applied to model millions of cells from various sources, improving data integration and gene representation learning. The choice between methods depends on the specific task to perform, ensuring a balance between predictive power and actionable understanding.

Finally, promoting collaborative research requires efficient data analysis, storage, and sharing in public repositories. Some collaborative initiatives and companies, like CELLxGENE [https://cellxgene.cziscience.com/] or immuneAI [https://www.immunai.com/platform/], are developing databases to manage single-cell data from millions of cells and thousands of patients. As seen in the inflammation atlas, ensuring homogeneous data formats and the availability of clinical metadata is crucial for effective reuse. However, these initiatives are still evolving and could be improved by integrating common tasks within their pipeline to facilitate usability.

Conclusions

The main conclusions of this PhD thesis are the following:

- FixNCut protocol preserves single-cell transcriptional profiles and tissue cellular composition at the sampling time, allowing for subsequent tissue dissociation while minimizing stress-related artifacts from downstream processing steps.

- Large-scale studies benefit from experimental designs with sample multiplexing to reduce technical variability and from accurate data processing, including computational methods to integrate data and minimize batch effects.

- Comprehensive blood profiling captures the full inflammatory spectrum of circulating immune cells across immune-driven diseases and healthy conditions.

- Assessing inflammation-related molecular signatures reveals the inflammation status across diseases and helps explore disease-driving mechanisms, capturing both commonalities and unique characteristics.

- Interpretable machine learning algorithms identify disease-discriminatory inflammatory genes at the cell type level, which could be used as clinical biomarkers.

- The inflammatory landscape underlying blood samples can be used to project patients onto the atlas embedding, enabling their classification into specific disease conditions. This new framework could transform current atlas applications into diagnostic tools for clinical use.

- Albumin treatment reprograms the transcriptional landscape and cellular composition of immune cells in the blood of patients with acute decompensated cirrhosis, enhancing B cell compartment and neutrophil antimicrobial functions, which leads to a reduced incidence of infections. Specifically, scRNA-seq analysis demonstrated changes in gene expression in B lymphocytes, CD4+ T cells, and myeloid cells of albumin-cultured PBMCs.

References

1. Murphy, K. & Weaver, C. *Janeway's Immunobiology*. (Garland Science, 2016).

2. Chaplin, D. D. Overview of the immune response. *J. Allergy Clin. Immunol.* **125**, S3–S23 (2010).

3. Carpenter, S. & O'Neill, L. A. J. From periphery to center stage: 50 years of advancements in innate immunity. *Cell* **187**, 2030–2051 (2024).

4. Aristizábal, B. & González, Á. Innate immune system. in *Autoimmunity: From Bench to Bedside [Internet]* (El Rosario University Press, 2013).

5. Turvey, S. E. & Broide, D. H. Innate immunity. *J. Allergy Clin. Immunol.* **125**, S24–S32 (2010).

6. Rehaume, L. M. & Hancock, R. E. W. Neutrophil-derived defensins as modulators of innate immune function. *Crit. Rev. Immunol.* **28**, 185–200 (2008).

7. Fulkerson, P. C. & Rothenberg, M. E. Targeting eosinophils in allergy, inflammation and beyond. *Nat. Rev. Drug Discov.* **12**, 117–129 (2013).

8. Abraham, S. N. & St John, A. L. Mast cell-orchestrated immunity to pathogens. *Nat. Rev. Immunol.* **10**, 440–452 (2010).

9. Silva, M. T. When two is better than one: macrophages and neutrophils work in concert in innate immunity as complementary and cooperative partners of a myeloid phagocyte system. *J. Leukoc. Biol.* **87**, 93–106 (2010).

10. Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998).

11. Walker, J. A., Barlow, J. L. & McKenzie, A. N. J. Innate lymphoid cells — how did we miss them? *Nat. Rev. Immunol.* **13**, 75–87 (2013).

12. Amersfoort, J., Eelen, G. & Carmeliet, P. Immunomodulation by endothelial cells — partnering up with the immune system? *Nat. Rev. Immunol.* **22**, 576–588 (2022).

13. Semple, J. W. & Freedman, J. Platelets and innate immunity. *Cell. Mol. Life Sci. CMLS* **67**, 499–511 (2010).

14. Morán, G. A. G., Parra-Medina, R., Cardona, A. G., Quintero-Ronderos, P. & Rodríguez, É. G. Cytokines, chemokines and growth factors. in *Autoimmunity: From Bench to Bedside [Internet]* (El Rosario University Press, 2013).

15. van Loo, G. & Bertrand, M. J. M. Death by TNF: a road to inflammation. *Nat. Rev. Immunol.* **23**, 289–303 (2023).

16. Akdis, M. *et al.* Interleukins (from IL-1 to IL-38), interferons, transforming growth factor β, and TNF-α: Receptors, functions, and roles in diseases. *J. Allergy Clin. Immunol.* **138**, 984–1010 (2016).

17. Kaneko, N., Kurata, M., Yamamoto, T., Morikawa, S. & Masumoto, J. The role of interleukin-1 in general pathology. *Inflamm. Regen.* **39**, 12 (2019).

18. Tanaka, T., Narazaki, M. & Kishimoto, T. IL-6 in Inflammation, Immunity, and Disease. *Cold Spring Harb. Perspect. Biol.* **6**, a016295 (2014).

19. Liu, J. *et al.* Interleukin-12: an update on its immunological activities, signaling and regulation of gene expression. *Curr. Immunol. Rev.* **1**, 119–137 (2005).

20. Waldmann, T. A., Miljkovic, M. D. & Conlon, K. C. Interleukin-15 (dys)regulation of lymphoid homeostasis: Implications for therapy of autoimmunity and cancer. *J. Exp. Med.* **217**, e20191062 (2019).

21. Nakanishi, K., Yoshimoto, T., Tsutsui, H. & Okamura, H. Interleukin-18 Regulates Both Th1 and

Th2 Responses. *Annu. Rev. Immunol.* **19**, 423–474 (2001).

22. Saraiva, M. & O'Garra, A. The regulation of IL-10 production by immune cells. *Nat. Rev. Immunol.* **10**, 170–181 (2010).

23. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. *Febs J.* **285**, 2944–2971 (2018).

24. Griffith, J. W., Sokol, C. L. & Luster, A. D. Chemokines and Chemokine Receptors: Positioning Cells for Host Defense and Immunity. *Annu. Rev. Immunol.* **32**, 659–702 (2014).

25. Aldinucci, D. & Colombatti, A. The Inflammatory Chemokine CCL5 and Cancer Progression. *Mediators Inflamm.* **2014**, 292376 (2014).

26. Pozzi, S. & Satchi-Fainaro, R. The role of CCL2/CCR2 axis in cancer and inflammation: The next frontier in nanomedicine. *Adv. Drug Deliv. Rev.* **209**, 115318 (2024).

27. Bickel, M. The role of interleukin-8 in inflammation and mechanisms of regulation. *J. Periodontol.* **64**, 456–460 (1993).

28. Gold, L. I. The role for transforming growth factor-beta (TGF-beta) in human cancer. *Crit. Rev. Oncog.* **10**, 303–360 (1999).

29. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.* **2**, 1–9 (2017).

30. Hu, X., Li, J., Fu, M., Zhao, X. & Wang, W. The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduct. Target. Ther.* **6**, 1–33 (2021).

31. Moens, U., Kostenko, S. & Sveinbjørnsson, B. The Role of Mitogen-Activated Protein Kinase-Activated Protein Kinases (MAPKAPKs) in Inflammation. *Genes* **4**, 101–133 (2013).

32. Ji, L. *et al.* The crucial regulatory role of type I interferon in inflammatory diseases. *Cell Biosci.* **13**, 230 (2023).

33. Strowig, T., Henao-Mejia, J., Elinav, E. & Flavell, R. Inflammasomes in health and disease. *Nature* **481**, 278–286 (2012).

34. Parra-Medina, R., Quintero-Ronderos, P. & Rodríguez, É. G. The complement system. in *Autoimmunity: From Bench to Bedside [Internet]* (El Rosario University Press, 2013).

35. Bonilla, F. A. & Oettgen, H. C. Adaptive immunity. *J. Allergy Clin. Immunol.* **125**, S33–S40 (2010).

36. Lam, N., Lee, Y. & Farber, D. L. A guide to adaptive immune memory. *Nat. Rev. Immunol.* 1–20 (2024) doi:10.1038/s41577-024-01040-6.

37. Cano, R. L. E. & Lopera, H. D. E. Introduction to T and B lymphocytes. in *Autoimmunity: From Bench to Bedside [Internet]* (El Rosario University Press, 2013).

38. Cyster, J. G. & Allen, C. D. C. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524–540 (2019).

39. Higgins, B. W., McHeyzer-Williams, L. J. & McHeyzer-Williams, M. G. Programming Isotype-specific Plasma Cell Function. *Trends Immunol.* **40**, 345–357 (2019).

40. Inoue, T. & Kurosaki, T. Memory B cells. *Nat. Rev. Immunol.* **24**, 5–17 (2024).

41. Bousso, P. T-cell activation by dendritic cells in the lymph node: lessons from the movies. *Nat. Rev. Immunol.* **8**, 675–684 (2008).

42. Bromley, S. K. *et al.* The immunological synapse and CD28-CD80 interactions. *Nat. Immunol.* **2**, 1159–1166 (2001).

43. Brownlie, R. J. & Zamoyska, R. T cell receptor signalling networks: branched, diversified and bounded. *Nat. Rev. Immunol.* **13**, 257–269 (2013).

44. Hwang, J.-R., Byeon, Y., Kim, D. & Park, S.-G. Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development. *Exp. Mol. Med.* **52**, 750–761 (2020).

45. Saini, A., Ghoneim, H. E., Lio, C.-W. J., Collins, P. L. & Oltz, E. M. Gene Regulatory Circuits in Innate and Adaptive Immune Cells. *Annu. Rev. Immunol.* **40**, 387–411 (2022).

46. Boyman, O. & Sprent, J. The role of interleukin-2 during homeostasis and activation of the immune system. *Nat. Rev. Immunol.* **12**, 180–190 (2012).

47. Koh, C.-H., Lee, S., Kwak, M., Kim, B.-S. & Chung, Y. CD8 T-cell subsets: heterogeneity, functions, and therapeutic potential. *Exp. Mol. Med.* **55**, 2287–2299 (2023).

48. Luckheeram, R. V., Zhou, R., Verma, A. D. & Xia, B. CD4+T Cells: Differentiation and Functions. *J. Immunol. Res.* **2012**, 925135 (2012).

49. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of Effector CD4 T Cell Populations*. *Annu. Rev. Immunol.* **28**, 445–489 (2010).

50. Youngblood, B. *et al.* Effector CD8 T cells dedifferentiate into long-lived memory cells. *Nature* **552**, 404–409 (2017).

51. Wherry, E. J. T cell exhaustion. *Nat. Immunol.* **12**, 492–499 (2011).

52. Zhou, Y., McLane, M. & Levitt, R. C. Th2 cytokines and asthma — Interleukin-9 as a therapeutic target for asthma. *Respir. Res.* **2**, 80–84 (2001).

53. Zeng, R. *et al.* Synergy of IL-21 and IL-15 in regulating CD8+ T cell expansion and function. *J. Exp. Med.* **201**, 139–148 (2005).

54. Mills, K. H. G. IL-17 and IL-17-producing cells in protection versus pathology. *Nat. Rev. Immunol.* **23**, 38–54 (2023).

55. Zheng, Y. *et al.* Interleukin-22, a TH17 cytokine, mediates IL-23-induced dermal inflammation and acanthosis. *Nature* **445**, 648–651 (2007).

56. Scott, T. E. *et al.* IL-4 and IL-13 induce equivalent expression of traditional M2 markers and modulation of reactive oxygen species in human macrophages. *Sci. Rep.* **13**, 19589 (2023).

57. Cook, S. A. Understanding interleukin 11 as a disease gene and therapeutic target. *Biochem. J.* **480**, 1987–2008 (2023).

58. Lei, Y. & Takahama, Y. XCL1 and XCR1 in the immune system. *Microbes Infect.* **14**, 262–267 (2012).

59. Gao, S.-H., Liu, S.-Z., Wang, G.-Z. & Zhou, G.-B. CXCL13 in Cancer and Other Diseases: Biological Functions, Clinical Significance, and Therapeutic Opportunities. *Life* **11**, 1282 (2021).

60. Burke, J. D. & Young, H. A. IFN-γ: A cytokine at the right time, is in the right place. *Semin. Immunol.* **43**, 101280 (2019).

61. Li, K. *et al.* The involvement of TNF-α and TNF-β as proinflammatory cytokines in lymphocyte-mediated adaptive immunity of Nile tilapia by initiating apoptosis. *Dev. Comp. Immunol.* **115**, 103884 (2021).

62. Schenten, D. & Medzhitov, R. Chapter 3 - The Control of Adaptive Immune Responses by the Innate Immune System. in *Advances in Immunology* (ed. Alt, F. W.) vol. 109 87–124 (Academic Press, 2011).

63. Netea, M. G. *et al.* A guiding map for inflammation. *Nat. Immunol.* **18**, 826–831 (2017).

64. Ortega-Gómez, A., Perretti, M. & Soehnlein, O. Resolution of inflammation: an integrated view.

EMBO Mol. Med. **5**, 661–674 (2013).

65. Fullerton, J. N. & Gilroy, D. W. Resolution of inflammation: a new therapeutic frontier. *Nat. Rev. Drug Discov.* **15**, 551–567 (2016).

66. Cowburn, A. S., Deighton, J., Walmsley, S. R. & Chilvers, E. R. The survival effect of TNF-alpha in human neutrophils is mediated via NF-kappa B-dependent IL-8 release. *Eur. J. Immunol.* **34**, 1733–1743 (2004).

67. Fox, S., Leitch, A. E., Duffin, R., Haslett, C. & Rossi, A. G. Neutrophil Apoptosis: Relevance to the Innate Immune Response and Inflammatory Disease. *J. Innate Immun.* **2**, 216–227 (2010).

68. Doran, A. C., Yurdagul, A. & Tabas, I. Efferocytosis in health and disease. *Nat. Rev. Immunol.* **20**, 254–267 (2020).

69. Vignali, D. A. A., Collison, L. W. & Workman, C. J. How regulatory T cells work. *Nat. Rev. Immunol.* **8**, 523–532 (2008).

70. Veglia, F., Sanseviero, E. & Gabrilovich, D. I. Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. *Nat. Rev. Immunol.* **21**, 485–498 (2021).

71. Duan, L., Rao, X. & Sigdel, K. R. Regulation of Inflammation in Autoimmune Disease. *J. Immunol. Res.* **2019**, 7403796 (2019).

72. Furman, D. *et al.* Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).

73. Mangani, D., Yang, D. & Anderson, A. C. Learning from the nexus of autoimmunity and cancer. *Immunity* **56**, 256–271 (2023).

74. Monteleone, G., Moscardelli, A., Colella, A., Marafini, I. & Salvatori, S. Immune-mediated inflammatory diseases: Common and different pathogenic and clinical features. *Autoimmun. Rev.* **22**, 103410 (2023).

75. Smolen, J. S. *et al.* Rheumatoid arthritis. *Nat. Rev. Dis. Primer* **4**, 1–23 (2018).

76. Alunno, A., Carubbi, F., Giacomelli, R. & Gerli, R. Cytokines in the pathogenesis of rheumatoid arthritis: new players and therapeutic targets. *BMC Rheumatol.* **1**, 3 (2017).

77. Yap, H.-Y. *et al.* Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. *Cells* **7**, 161 (2018).

78. Greb, J. E. *et al.* Psoriasis. *Nat. Rev. Dis. Primer* **2**, 1–17 (2016).

79. Sieminska, I., Pieniawska, M. & Grzywa, T. M. The Immunology of Psoriasis—Current Concepts in Pathogenesis. *Clin. Rev. Allergy Immunol.* **66**, 164–191 (2024).

80. Rendon, A. & Schäkel, K. Psoriasis Pathogenesis and Treatment. *Int. J. Mol. Sci.* **20**, 1475 (2019).

81. FitzGerald, O. *et al.* Psoriatic arthritis. *Nat. Rev. Dis. Primer* **7**, 1–17 (2021).

82. Azuaga, A. B., Ramírez, J. & Cañete, J. D. Psoriatic Arthritis: Pathogenesis and Targeted Therapies. *Int. J. Mol. Sci.* **24**, 4901 (2023).

83. Veale, D. J. & Fearon, U. The pathogenesis of psoriatic arthritis. *The Lancet* **391**, 2273–2284 (2018).

84. Guan, Q. A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease. *J. Immunol. Res.* **2019**, 7247238 (2019).

85. Roda, G. *et al.* Crohn's disease. *Nat. Rev. Dis. Primer* **6**, 1–19 (2020).

86. Petagna, L. *et al.* Pathophysiology of Crohn's disease inflammation and recurrence. *Biol. Direct*

15, 23 (2020).

87. Li, N. & Shi, R.-H. Updated review on immune factors in pathogenesis of Crohn's disease. *World J. Gastroenterol.* **24**, 15–22 (2018).

88. Kobayashi, T. *et al.* Ulcerative colitis. *Nat. Rev. Dis. Primer* **6**, 1–20 (2020).

89. Kaur, A. & Goggolidou, P. Ulcerative colitis: understanding its cellular pathology could provide insights into novel therapies. *J. Inflamm.* **17**, 15 (2020).

90. Filippi, M. *et al.* Multiple sclerosis. *Nat. Rev. Dis. Primer* **4**, 1–27 (2018).

91. Baecher-Allan, C., Kaskow, B. J. & Weiner, H. L. Multiple Sclerosis: Mechanisms and Immunotherapy. *Neuron* **97**, 742–768 (2018).

92. Soldan, S. S. & Lieberman, P. M. Epstein–Barr virus and multiple sclerosis. *Nat. Rev. Microbiol.* **21**, 51–64 (2023).

93. Kaul, A. *et al.* Systemic lupus erythematosus. *Nat. Rev. Dis. Primer* **2**, 1–21 (2016).

94. Crow, M. K. Pathogenesis of systemic lupus erythematosus: risks, mechanisms and therapeutic targets. *Ann. Rheum. Dis.* **82**, 999–1014 (2023).

95. Interferon pathway in SLE: one key to unlocking the mystery of the disease | Lupus Science & Medicine. https://lupus.bmj.com/content/6/1/e000270.

96. Hotchkiss, R. S. *et al.* Sepsis and septic shock. *Nat. Rev. Dis. Primer* **2**, 1–21 (2016).

97. van der Poll, T., van de Veerdonk, F. L., Scicluna, B. P. & Netea, M. G. The immunopathology of sepsis and potential therapeutic targets. *Nat. Rev. Immunol.* **17**, 407–420 (2017).

98. Lambrecht, B. N. & Hammad, H. The immunology of asthma. *Nat. Immunol.* **16**, 45–56 (2015).

99. Holgate, S. T. *et al.* Asthma. *Nat. Rev. Dis. Primer* **1**, 1–22 (2015).

100. Barnes, P. J. *et al.* Chronic obstructive pulmonary disease. *Nat. Rev. Dis. Primer* **1**, 1–21 (2015).

101. Domej, W., Oettl, K. & Renner, W. Oxidative stress and free radicals in COPD – implications and relevance for treatment. *Int. J. Chron. Obstruct. Pulmon. Dis.* **9**, 1207–1224 (2014).

102. Gadgil, A. & Duncan, S. R. Role of T-lymphocytes and pro-inflammatory mediators in the pathogenesis of chronic obstructive pulmonary disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* **3**, 531–541 (2008).

103. Ginès, P. *et al.* Liver cirrhosis. *The Lancet* **398**, 1359–1376 (2021).

104. Albillos, A. *et al.* Cirrhosis-associated immune dysfunction. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 112–134 (2022).

105. Moreau, R. *et al.* Acute-on-Chronic Liver Failure Is a Distinct Syndrome That Develops in Patients With Acute Decompensation of Cirrhosis. *Gastroenterology* **144**, 1426-1437.e9 (2013).

106. Arroyo, V., Moreau, R. & Jalan, R. Acute-on-Chronic Liver Failure. *N. Engl. J. Med.* **382**, 2137–2145 (2020).

107. Amin, A. & Mookerjee, R. P. Acute-on-chronic liver failure: definition, prognosis and management. *Frontline Gastroenterol.* **11**, 458–467 (2020).

108. Jalan, R. & Williams, R. Acute-on-Chronic Liver Failure: Pathophysiological Basis of Therapeutic Options. *Blood Purif.* **20**, 252–261 (2002).

109. Martin-Mateos, R., Alvarez-Mon, M. & Albillos, A. Dysfunctional Immune Response in Acute-on-Chronic Liver Failure: It Takes Two to Tango. *Front. Immunol.* **10**, (2019).

110. Arroyo, V. *et al.* Acute-on-chronic liver failure in cirrhosis. *Nat. Rev. Dis. Primer* **2**, 1–18 (2016).

111. Hernaez, R., Solà, E., Moreau, R. & Ginès, P. Acute-on-chronic liver failure: an update. *Gut* **66**, 541–553 (2017).

112. Casulleras, M. *et al.* Albumin internalizes and inhibits endosomal TLR signaling in leukocytes from patients with decompensated cirrhosis. *Sci. Transl. Med.* **12**, eaax5135 (2020).

113. Iracheta-Vellve, A. *et al.* Interleukin-1 inhibition facilitates recovery from liver injury and promotes regeneration of hepatocytes in alcoholic hepatitis in mice. *Liver Int.* **37**, 968–973 (2017).

114. Ki, S. H. *et al.* Interleukin-22 treatment ameliorates alcoholic liver injury in a murine model of chronic-binge ethanol feeding: Role of signal transducer and activator of transcription 3. *Hepatology* **52**, 1291–1300 (2010).

115. Yannaki, E. *et al.* G-CSF–primed hematopoietic stem cells or G-CSF per se accelerate recovery and improve survival after liver injury, predominantly by promoting endogenous repair programs. *Exp. Hematol.* **33**, 108–119 (2005).

116. Rouse, B. T. & Sehrawat, S. Immunity and immunopathology to viruses: what decides the outcome? *Nat. Rev. Immunol.* **10**, 514–526 (2010).

117. Pichlmair, A. & Sousa, C. R. e. Innate Recognition of Viruses. *Immunity* **27**, 370–383 (2007).

118. Mogensen, T. H. & Paludan, S. R. Molecular Pathways in Virus-Induced Cytokine Production. *Microbiol. Mol. Biol. Rev.* **65**, 131–150 (2001).

119. Krammer, F. *et al.* Influenza. *Nat. Rev. Dis. Primer* **4**, 1–21 (2018).

120. Jamison, D. A. *et al.* A comprehensive SARS-CoV-2 and COVID-19 review, Part 1: Intracellular overdrive for SARS-CoV-2 infection. *Eur. J. Hum. Genet.* **30**, 889–898 (2022).

121. Yuen, M.-F. *et al.* Hepatitis B virus infection. *Nat. Rev. Dis. Primer* **4**, 1–20 (2018).

122. Bekker, L.-G. *et al.* HIV infection. *Nat. Rev. Dis. Primer* **9**, 1–21 (2023).

123. National Institutes of Health, National Cancer Institute. SEER Training Modules, Cancer Classification. https://training.seer.cancer.gov/disease/categories/classification.html.

124. Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A. & Fares, Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target. Ther.* **5**, 1–17 (2020).

125. Wang, Q. *et al.* Role of tumor microenvironment in cancer progression and therapeutic strategy. *Cancer Med.* **12**, 11149–11165 (2023).

126. Zhao, H. *et al.* Inflammation and tumor progression: signaling pathways and targeted intervention. *Signal Transduct. Target. Ther.* **6**, 1–46 (2021).

127. Rodríguez, J. *et al.* When Tissue is an Issue the Liquid Biopsy is Nonissue: A Review. *Oncol. Ther.* **9**, 89–110 (2021).

128. Boukovala, M., Westphalen, C. B. & Probst, V. Liquid biopsy into the clinics: Current evidence and future perspectives. *J. Liq. Biopsy* **4**, 100146 (2024).

129. Wang, H. *et al.* Liquid biopsy for human cancer: cancer screening, monitoring, and treatment. *MedComm* **5**, e564 (2024).

130. De Donno, C. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* 1–10 (2023) doi:10.1038/s41592-023-02035-2.

131. Lustberg, M. B., Stover, D. G. & Chalmers, J. J. Implementing liquid biopsies in clinical trials: State of affairs, opportunities and challenges. *Cancer J. Sudbury Mass* **24**, 61–64 (2018).

132. Rahman, M. M. *et al.* In-vitro assays for immuno-oncology drug efficacy assessment and screening for personalized cancer therapy: scopes and challenges. *Expert Rev. Clin. Immunol.*

20, 821–838 (2024).

133. Adile, A. A., Bakhshinyan, D., Venugopal, C. & Singh, S. K. In Vitro Assays for Screening Small Molecules. in *Brain Tumor Stem Cells: Methods and Protocols* (eds. Singh, S. K. & Venugopal, C.) 189–196 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-8805-1_16.

134. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).

135. Li, X. & Wang, C.-Y. From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci.* **13**, 1–6 (2021).

136. Efremova, M., Vento-Tormo, R., Park, J.-E., Teichmann, S. A. & James, K. R. Immunology in the Era of Single-Cell Technologies. *Annu. Rev. Immunol.* **38**, 727–757 (2020).

137. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

138. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* **17**, 220–232 (2018).

139. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

140. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020).

141. Ramsköld, D. *et al.* Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

142. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

143. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).

144. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

145. 10X Genomics - Official 10X Genomics Support. *10X Genomics* https://www.10xgenomics.com/.

146. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776–779 (2014).

147. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.* **11**, 4025 (2020).

148. Huang, D. *et al.* Advances in single-cell RNA sequencing and its applications in cancer research. *J. Hematol. Oncol.J Hematol Oncol* **16**, 98 (2023).

149. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).

150. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

151. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

152. Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat.*

Biotechnol. **38**, 747–755 (2020).

153. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).

154. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

155. Boonlayangoor, P., Telischi, M., Boonlayangoor, S., Sinclair, T. F. & Millhouse, E. W. Cryopreservation of human granulocytes: study of granulocyte function and ultrastructure. *Blood* **56**, 237–245 (1980).

156. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).

157. Machado, L., Relaix, F. & Mourikis, P. Stress relief: emerging methods to mitigate dissociation-induced artefacts. *Trends Cell Biol.* **31**, 888–897 (2021).

158. Massoni-Badosa, R. *et al.* Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol.* **21**, 112 (2020).

159. Chen, D. *et al.* Cryopreservation Preserves Cell-Type Composition and Gene Expression Profiles in Bone Marrow Aspirates From Multiple Myeloma Patients. *Front. Genet.* **12**, (2021).

160. Wang, X., Yu, L. & Wu, A. R. The effect of methanol fixation on single-cell RNA sequencing data. *BMC Genomics* **22**, 420 (2021).

161. Phan, H. V. *et al.* High-throughput RNA sequencing of paraformaldehyde-fixed single cells. *Nat. Commun.* **12**, 5636 (2021).

162. O'Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 210 (2019).

163. Zhang, Y. *et al.* Sample-multiplexing approaches for single-cell sequencing. *Cell. Mol. Life Sci.* **79**, 466 (2022).

164. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).

165. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

166. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

167. Cell Ranger's Cell Multiplexing Algorithm - Official 10x Genomics Support. *10x Genomics* https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-3p-cellplex-algorithm.

168. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).

169. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).

170. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol.* **22**, 301 (2021).

171. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).

172. Poirion, O. B., Zhu, X., Ching, T. & Garmire, L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. *Front. Genet.* **7**, (2016).

173. Luecken, M. D. & Theis, F. J. Current best practices in single‐cell RNA‐seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

174. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).

175. Andrews, T. S., Kiselev, V. Y., McCarthy, D. & Hemberg, M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.* **16**, 1–9 (2021).

176. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).

177. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

178. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

179. Grieshaber-Bouyer, R. *et al.* The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. *Nat. Commun.* **12**, 2856 (2021).

180. Xi, N. M. & Li, J. J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR Protoc.* **2**, 100699 (2021).

181. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

182. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4 (2019).

183. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

184. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

185. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

186. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).

187. Bellman, R. & Bellman, R. E. *Dynamic Programming*. (Princeton University Press, 1957).

188. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).

189. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2**, 239–250 (2016).

190. Shlens, J. A Tutorial on Principal Component Analysis. Preprint at https://doi.org/10.48550/arXiv.1404.1100 (2014).

191. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

192. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

193. Hetzel, L., Fischer, D. S., Günnemann, S. & Theis, F. J. Graph representation learning for single-cell biology. *Curr. Opin. Syst. Biol.* **28**, 100347 (2021).

194. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

195. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).

196. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).

197. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1**, 80–83 (1945).

198. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

199. Zhang, J. M., Kamath, G. M. & Tse, D. N. Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq. *Cell Syst.* **9**, 383-392.e6 (2019).

200. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* **16**, 1007–1015 (2019).

201. Andreatta, M., Berenstein, A. J. & Carmona, S. J. scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets. *Bioinformatics* **38**, 2642–2644 (2022).

202. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).

203. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).

204. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).

205. Kang, J. B. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).

206. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).

207. Xu, C. *et al.* Probabilistic harmonization and annotation of single‐cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).

208. Massoni-Badosa, R. *et al.* An atlas of cells in the human tonsil. *Immunity* **57**, 379-399.e18 (2024).

209. Hie, B. L., Kim, S., Rando, T. A., Bryson, B. & Berger, B. Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nat. Protoc.* **19**, 2283–2297 (2024).

210. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

211. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

212. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).

213. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

214. Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y. & Hemberg, M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.* **49**, e42 (2021).

215. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).

216. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).

217. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

218. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

219. Junttila, S., Smolander, J. & Elo, L. L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Brief. Bioinform.* **23**, bbac286 (2022).

220. Murphy, A. E. & Skene, N. G. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* **13**, 7851 (2022).

221. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles | PNAS. https://www.pnas.org/doi/10.1073/pnas.0506580102.

222. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

223. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).

224. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data | Database | Oxford Academic. https://academic.oup.com/database/article/doi/10.1093/database/baz046/5427041.

225. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

226. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).

227. Zhang, Y. *et al.* Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.* **18**, 2953–2961 (2020).

228. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).

229. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).

230. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

231. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

232. Raimundo, F., Meng-Papaxanthos, L., Vallot, C. & Vert, J.-P. Machine learning for single-cell genomics data analysis. *Curr. Opin. Syst. Biol.* **26**, 64–71 (2021).

233. Erfanian, N. *et al.* Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomed. Pharmacother.* **165**, 115077 (2023).

234. Wagle, M. M., Long, S., Chen, C., Liu, C. & Yang, P. Interpretable deep learning in single-cell omics. *Bioinformatics* **40**, btae374 (2024).

235. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions.

236. Szałata, A. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).

237. Köhler, N. D., Büttner, M., Andriamanga, N. & Theis, F. J. Deep learning does not outperform classical machine learning for cell-type annotation. 653907 Preprint at https://doi.org/10.1101/653907 (2021).

238. Uniken Venema, W. T. C. *et al.* Gut mucosa dissociation protocols influence cell type proportions and single-cell gene expression levels. *Sci. Rep.* **12**, 9897 (2022).

239. Deleersnijder, D. *et al.* Current Methodological Challenges of Single-Cell and Single-Nucleus RNA-Sequencing in Glomerular Diseases. *J. Am. Soc. Nephrol. JASN* **32**, 1838–1852 (2021).

240. Baechler, E. C. *et al.* Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. *Genes Immun.* **5**, 347–353 (2004).

241. Attar, M. *et al.* A practical solution for preserving single cells for RNA sequencing. *Sci. Rep.* **8**, 2151 (2018).

242. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

243. Aney, K. J. *et al.* Novel Approach for Pancreas Transcriptomics Reveals the Cellular Landscape in Homeostasis and Acute Pancreatitis. *Gastroenterology* **166**, 1100–1113 (2024).

244. Mutisheva, I., Robatel, S., Bäriswyl, L. & Schenk, M. An Innovative Approach to Tissue Processing and Cell Sorting of Fixed Cells for Subsequent Single-Cell RNA Sequencing. *Int. J. Mol. Sci.* **23**, 10233 (2022).

245. Nesterenko, P. A. *et al.* Droplet-based mRNA sequencing of fixed and permeabilized cells by CLInt-seq allows for antigen-specific TCR cloning. *Proc. Natl. Acad. Sci.* **118**, e2021190118 (2021).

246. Howitt, G. *et al.* Benchmarking single-cell hashtag oligo demultiplexing methods. *NAR Genomics Bioinforma.* **5**, lqad086 (2023).

247. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).

248. Cardiello, J. F. *et al.* Evaluation of genetic demultiplexing of single-cell sequencing data from model species. *Life Sci. Alliance* **6**, (2023).

249. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

250. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).

251. Bernstein, N. J. *et al.* Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. *Cell Syst.* **11**, 95-101.e5 (2020).

252. Curion, F. *et al.* hadge: a comprehensive pipeline for donor deconvolution in single-cell studies. *Genome Biol.* **25**, 109 (2024).

253. Li, L. *et al.* A hybrid demultiplexing strategy that improves performance and robustness of cell hashing. *Brief. Bioinform.* **25**, bbae254 (2024).

254. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

255. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

256. Rizzetto, S. *et al.* Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Sci. Rep.* **7**, 12781 (2017).

257. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).

258. Amezquita, R., Lun, A., Hicks, S. & Gottardo, R. *Orchestrating Single-Cell Analysis with Bioconductor*.

259. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

260. Dann, E. *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* **55**, 1998–2008 (2023).

261. Cui, A. *et al.* Dictionary of immune responses to cytokines at single-cell resolution. *Nature* **625**, 377–384 (2024).

262. Walle, T. wallet-maker/cytopus: Cytopus v1.30. Zenodo https://doi.org/10.5281/zenodo.8366975 (2023).

263. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* 1–12 (2023) doi:10.1038/s41587-023-01940-3.

264. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).

265. Rönnblom, L. & Leonard, D. Interferon pathway in SLE: one key to unlocking the mystery of the disease. *Lupus Sci. Med.* **6**, e000270 (2019).

266. Sun, S.-C., Chang, J.-H. & Jin, J. Regulation of NF-κB in Autoimmunity. *Trends Immunol.* **34**, 282–289 (2013).

267. Thomas, T. *et al.* A longitudinal single-cell therapeutic atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. 2023.05.05.539635 Preprint at https://doi.org/10.1101/2023.05.05.539635 (2023).

268. Garrido-Trigo, A. *et al.* Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat. Commun.* **14**, 4506 (2023).

269. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).

270. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://dl.acm.org/doi/10.1145/2939672.2939785.

271. Anderson, A. E. *et al.* Expression of STAT3-regulated genes in circulating CD4+ T cells discriminates rheumatoid arthritis independently of clinical parameters in early arthritis. *Rheumatol. Oxf. Engl.* **58**, 1250–1258 (2019).

272. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).

273. Barry, M. J. Clinical practice. Prostate-specific-antigen testing for early diagnosis of prostate cancer. *N. Engl. J. Med.* **344**, 1373–1377 (2001).

274. Maina, E. K. *et al.* CD4+ T cell counts in initiation of antiretroviral therapy in HIV infected asymptomatic individuals; controversies and inconsistencies. *Immunol. Lett.* **168**, 279–284 (2015).

275. Chambuso, R., Mthembu, M., Kaambo, E., Robertson, B. & Ramesar, R. Immunogenomic Biomarkers and Validation in Lynch Syndrome. *Cells* **12**, 491 (2023).

276. Weiss, E. *et al.* Characterization of Blood Immune Cells in Patients With Decompensated

Cirrhosis Including ACLF. *Front. Immunol.* **11**, (2021).

277. Bihari, S., Bannard-Smith, J. & Bellomo, R. Albumin as a drug: its biological effects beyond volume expansion. *Crit. Care Resusc. J. Australas. Acad. Crit. Care Med.* **22**, 257–265 (2020).

278. Drumm, K., Bauer, B., Freudinger, R. & Gekle, M. Albumin Induces NF-κB Expression in Human Proximal Tubule-Derived Cells (IHKE-1). *Cell. Physiol. Biochem.* **12**, 187–196 (2002).

Appendices

Appendix A. FixNCut: a Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assay[4]

Shuoshuo Wang, **Laura Jiménez-Gracia**, Antonella Arruda de Amaral, Ioannis S. Vlachos, Jasmine Plummer, Holger Heyn, and Luciano G. Martelotto. **FixNCut: a Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assay**, *Bio-protocol* 14(17): e5063 (2024).

DOI: https://doi.org/10.21769/BioProtoc.5063

Impact Factor 2023: 1.0 (Q4)

[4] In this methodological paper, I contributed to writing and editing of the manuscript.

bio-protocol

Open Access

Published: Sep 05, 2024

FixNCut: A Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assays

Shuoshuo Wang[1, 2], Laura Jiménez-Gracia[3, 4], Antonella Arruda de Amaral[1, 2], Ioannis S. Vlachos[1, 2], Jasmine Plummer[6, 7, 8, 9], Holger Heyn[3, 4, 5] and Luciano G. Martelotto[10, 11, §, *]

[1]Spatial Technologies, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain
[4]Universitat de Barcelona (UB), Barcelona, Spain
[5]Omniscope, Barcelona, Spain
[6]Center for Spatial Omics, St Jude Children's Research Hospital, Memphis, TN, USA
[7]Department of Developmental Neurobiology, St Jude Children's Research Hospital, Memphis, TN, USA
[8]Department of Cellular and Molecular Biology, St Jude Children's Research Hospital, Memphis, TN, USA
[9]Comprehensive Cancer Center, St Jude Children's Research Hospital, Memphis, TN, USA
[10]Adelaide Centre for Epigenetics (ACE), University of Adelaide, Adelaide, South Australia, Australia
[11]South Australian immunoGENomics Cancer Institute (SAiGENCI), University of Adelaide, Adelaide, South Australia, Australia
[§]Technical contact
*For correspondence: luciano.martelotto@adelaide.edu.au

Abstract

The quality of standard single-cell experiments often depends on the immediate processing of cells or tissues post-harvest to preserve fragile and vulnerable cell populations, unless the samples are adequately fixed and stored. Despite the recent rise in popularity of probe-based and aldehyde-fixed RNA assays, these methods face limitations in species and target availability and are not suitable for immunoprofiling or assessing chromatin accessibility. Recently, a reversible fixation strategy known as FixNCut has been successfully deployed to separate sampling from downstream applications in a reproducible and robust manner, avoiding stress or necrosis-related artifacts. In this article, we present an optimized and robust practical guide to the FixNCut protocol to aid the end-to-end adaptation of this versatile method. This protocol not only decouples tissue or cell harvesting from single-cell assays but also enables a flexible and decentralized workflow that unlocks the potential for single-cell analysis as well as unconventional study designs that were previously considered unfeasible.

Key features

- Reversible fixation: Preserves cellular and molecular structures with the option to later reverse the fixation for downstream applications, maintaining cell integrity
- Compatibility with single-cell assays: Supports single-cell genomic assays such as scRNA-seq and ATAC-seq, essential for high-resolution analysis of cell function and gene expression
- Flexibility in sample handling: Allows immediate fixation post-collection, decoupling sample processing from analysis, beneficial in settings where immediate processing is impractical

Cite as: Wang, S. et al. (2024). FixNCut: A Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assays. Bio-protocol 14(17): e5063. DOI: 10.21769/BioProtoc.5063.

Copyright: © 2024 The Authors; exclusive licensee Bio-protocol LLC.
This is an open access article under the CC BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/).

bio-protocol

Open Access

Published: Sep 05, 2024

- Preservation of RNA and DNA integrity: Effectively preserves RNA and DNA, reducing degradation to ensure accurate transcriptomic and genomic profiling
- Suitability for various biological samples: Applicable to a wide range of biological samples, including tissues and cell suspensions, whether freshly isolated or post-dissociated
- Enables multi-center studies: Facilitates collaborative research across multiple centers by allowing sample fixation at the point of collection, enhancing research scale and diversity
- Avoidance of artifacts: Minimizes stress or necrosis-related artifacts, preserving the natural cellular physiology for accurate genomic and transcriptomic analysis

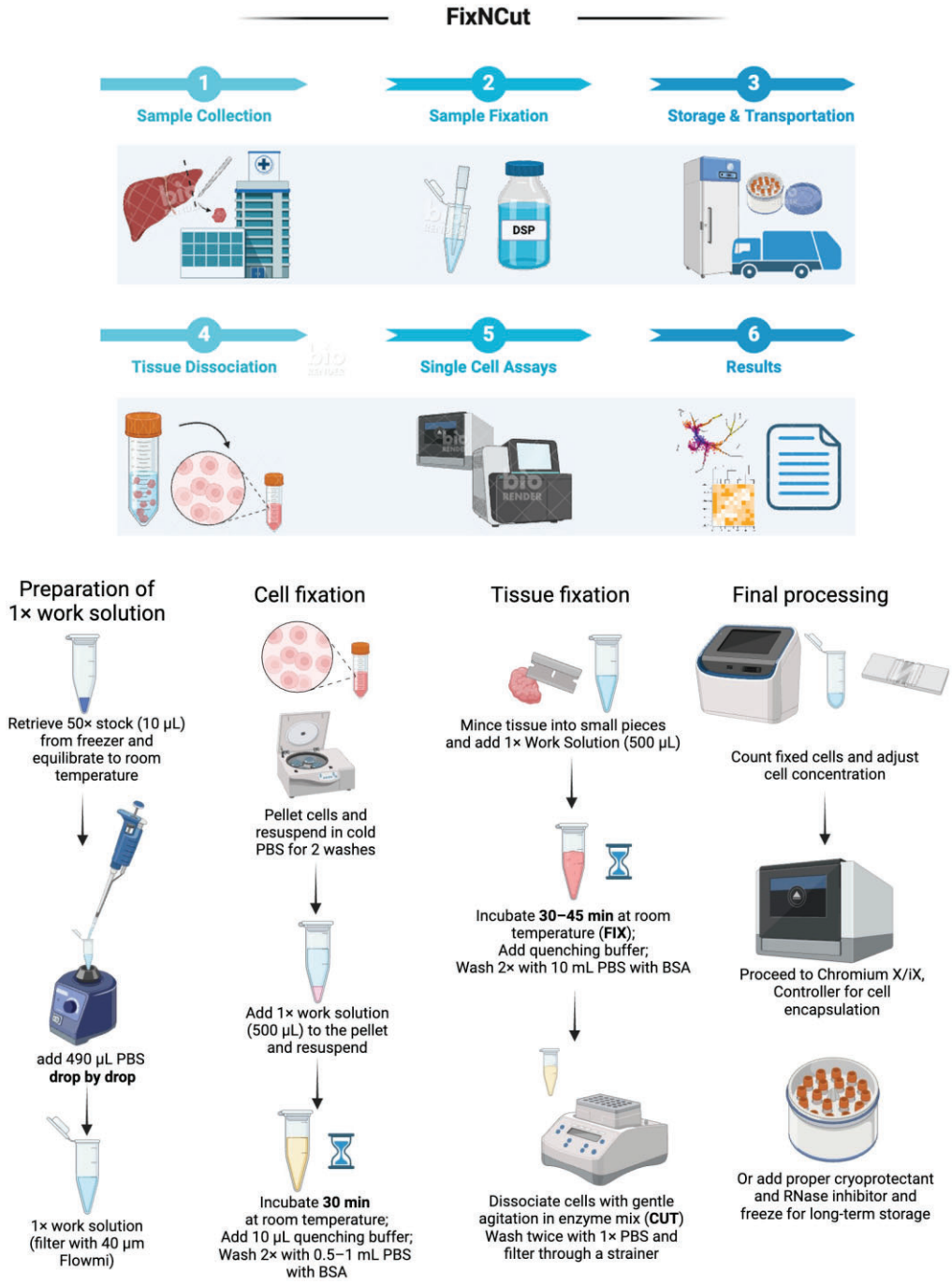
Keywords: Single cell, Tissue, Fixation, Lomant's reagent, DSP, Sample preservation, Single-cell sequencing

This protocol is used in: Genome Biol (2024), DOI: 10.1186/s13059-024-03219-5

Copyright: © 2024 The Authors; exclusive licensee Bio-protocol LLC.

This is an open access article under the CC BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/).

Graphical overview



Background

Single-cell techniques have revolutionized biological and clinical research by quantitatively capturing the genomic and transcriptomic state of individual cells at unprecedented resolution and scope [1]. This approach moves beyond averaging the bulk population, which consists of various cell types. However, the technical and logistical challenges

of acquiring or generating high-viability single-cell suspensions have limited the application of single-cell genomics and transcriptomics.

Multi-center clinical studies often involve sample collection from remote areas lacking processing infrastructure [2,3]; additionally, longitudinal and cohort recruitment of patients occurs over the years, subject to patient availability and logistics constraints. Moreover, even in advanced clinical centers, the urgency of tissue harvesting during surgical procedures, biopsies, or autopsies can depend on factors such as the type of surgery, the purpose, and the specific protocol followed, as well as the transfer procedures between surgery and histopathology [4,5]. In basic research using organoids, embryos, or rare specimens, the limited cell number poses a significant bottleneck for conducting scientifically meaningful and economically feasible single-cell analysis.

In many experimental setups typical for clinical research, the freeze-thaw process remains the predominant option for sample collection en masse. However, freezing will physically damage the cell membrane, impact the post-thaw viability, and disrupt intracellular compartments where endogenous RNases are sequestered, leading to RNA degradation. In addition, cryopreservation may lead to the loss of certain cell types and induce cellular stress response [6]. Moreover, variably staggered freezing and thawing through aliquoting cells for parallel orthogonal assays, for example, single-cell mass cytometry (CyTOF), single-cell RNA-Seq, single-cell ATAC-Seq, and additional multiomic assays, can lead to inconsistency, introducing technical artifacts and batch-wise noise to the data. Recently, the fixed RNA (flex) assay has been gaining popularity due to its robustness; however, so far probes are only available for humans and mice and do not target immunologically and clinically paramount transcripts such as joining and variable regions in B and T-cell receptor (BCR and TCR), Killer cell immunoglobulin-like receptors (KIRs), and human leukocyte antigen (HLA), nor polyadenylated non-coding RNA.

Notably, many important immune cells—such as neutrophils, dendritic cells, monocytes, macrophages, and lymphocytes (B cell and T cells), as well as cells with higher intrinsic mechanical integrity such as epithelial cells—are considered fragile for single-cell applications due to their susceptibility to mechanical damage during sample preparation, including cell sorting or tissue dissociation [7–9]. Loss of these cells often leads to underrepresentation in downstream analysis. Therefore, identifying a method allowing the stabilization of fragile cells without perturbing RNA integrity will enhance the recovery of these important populations.

The ideal fixative must possess several characteristics: first, it should be small and able to penetrate the cell membrane and tissues. Second, it should preserve the structure and integrity of cells. Third, it should not destroy RNA integrity and should, ideally, inhibit RNase activity at least temporarily. Currently, one of the most common options for single-cell assays is using either standard or modified methanol fixation [10–14]. Methanol works by dehydrating samples and denaturing proteins in a gentle manner, similar to histology fixation. However, methanol as a standalone fixative has raised concerns regarding biased results or ambient RNA leakage in single-cell RNA-Seq [6]. A possible mechanistic explanation is the irreversible intracellular compartment disruption resulting in the loss of normal lipid and protein structure during dehydration–rehydration cycles. In addition, incomplete reverse transcription of mRNAs with more complex secondary structures was suggested as a major caveat [15].

In contrast, dithio-bis-succinimidyl propionate (DSP), known as Lomant's reagent [16], has been traditionally used for histological tissue fixation to stabilize cellular integrity and structures through covalent crosslinking of free amine groups found at the N-terminus of polypeptide chains (diagrammatic molecular basis is depicted in Figure 1 of Akaki et al. [17]). It has been repurposed for single-cell RNA sequencing by maintaining RNA integrity and yield in bulk RNA extractions [18–20]. The cell and membrane permeability have been proven to be particularly useful in detecting rare cell populations such as tissue-resident immune cells. Certain T-cell populations can exist in very low numbers upon activation, making them difficult to isolate and analyze with conventional single-cell methods. A protocol called CLInt-Seq (crosslinker regulated intracellular phenotype sequencing) was developed to overcome the hurdle by crosslinking intracellular cytokines [21]. This technique combines and improves upon existing techniques to collect and genetically sequence rare T cells. The reversibility of DSP-induced crosslinking is a key feature allowing cells and tissue to be further processed or dissociated for downstream applications. However, a recent study concluded that de-crosslinking the DSP-fixed samples is optional, and additional handling steps of de-crosslinking may contribute counterproductively to cell loss [22].

Recently, we developed the further optimized FixNCut protocol and demonstrated its robustness and benefits systematically by comparing fresh and fixed lung, colon, and pancreas samples from different species even under cryopreservation for an extended period [23,24]. Here, we provide an end-to-end solution in the form of a practical guide on the implementation of this method, to remove practical barriers by streamlining the transport of samples

and scheduling of shared instruments for downstream single-cell isolation and processing.

Materials and reagents

Biological materials

1. Tissue samples and cell suspensions can both be used. For cells, the protocol is set up for up to 2×10^6 cells. For more cells, scale up the fixation accordingly. Tissues larger than 3 mm in diameter or edge length need to be partitioned into smaller pieces to facilitate fixative penetration

Reagents

1. DSP (Dithio-bis-succinimidyl propionate), also known as Lomant's reagent (Thermo Fisher, catalog number: 22586)
2. DMSO, anhydrous (Thermo Fisher, Molecular Probes, catalog number: D12345)
3. Liberase™ research grade, 10 mg (Roche, catalog number: 5401119001)
4. Protector RNase inhibitor (40 U/μL) (Roche, catalog number: RNAINH-RO)
5. UltraPure™ 1 M Tris-HCI buffer, pH 7.5 (Thermo Fisher, Invitrogen, catalog number: 15567027)
6. 10× PBS buffer, pH 7.4 (Invitrogen, catalog number: AM9624 or AM9625)
7. Ambion nuclease-free water (Invitrogen, catalog number: AM9932)
8. Bovine serum albumin (BSA) solution, sterile filtered and cell-culture tested (Sigma Aldrich, catalog number: A1595)
9. Tris base, UltraPure Tris buffer (powder format) (Thermo Fisher, Invitrogen, catalog number: 15504020)

Solutions

1. Fixation concentrate stock solution (1 mL, for 100 standard assays) (see Recipes)
2. PBS with 1% BSA (see Recipes)
3. 1 M Tris pH 7.5 (optional, if not using ready-made buffer) (see Recipes)

Recipes

1. **Fixation concentrate stock solution (1 mL, for 100 standard assays)**

Reagent	Final concentration	Amount
DSP (powder)	n/a	50 mg
DMSO	n/a	1 mL
Total	50 mg/mL (w/v)	1 mL

2. **PBS with 1% BSA**

Reagent	Final concentration	Amount
10× PBS	n/a	50 mg
BSA 10%	n/a	1 mL
Total	n/a	1 mL

3. **1 M Tris pH 7.5 (optional, if not using ready-made buffer)**
Start with 30 mL of water and adjust pH to 7.5 by adding 5 M HCl to a final volume of 50 mL.

Reagent	Final concentration	Amount
Tris base	n/a	6.05 g
Nuclease-free water	n/a	50 mL

Total	1 M	50 mL

Laboratory supplies

1. DNA LoBind tubes 1.5 mL (Eppendorf, catalog number: 022431021)
2. DNA LoBind tubes 2.0 mL (Eppendorf, catalog number: 022431048)
3. Flowmi® cell strainers, porosity 40 µm, for 1000 µL pipette tips (Sigma, Scienceware, catalog number: BAH136800040)
4. pluriStrainer mini 70 µm (pluriSelect, catalog number: 43-10070-40)
5. Falcon conical centrifuge tubes (Corning, catalog number: 352070 for 50 mL, 352095 for 15 mL)
6. Sterile serological pipettes with pipettor
7. Rainin Pipet-Lite LTS pipette tips (Rainin, catalog number: 30389240, 30389213, 30389226)

Equipment

1. Centrifuge 5810/5810R (Eppendorf, catalog number: EP022628188) with rotor S-4-104
2. Rotor F-35-6-30 (Eppendorf, catalog number: EP5427716009)
3. S-4-104 rotor adapters for 50 × 1.5/2 mL tubes (Eppendorf, catalog number: 58-257-40009)
4. Standard heavy-duty vortex mixer (VWR or Fisherbrand, catalog number: 97043-562)
5. ThermoMixer C (Eppendorf, catalog number: 05-412-503) or thermoblock with minimal temperature fluctuation
6. Ice bucket with cool blocks or Cool Rack CFT30 (Corning, catalog number: CLS432052)
7. Automated cell counter, e.g., Luna-FX7 (Logos Biosystems) or hemocytometer
8. CoolCell freezing container for 12 × 1 mL or 2 mL cryogenic vials (Corning, catalog number: 432000) or Mr. Frosty freezing container (Thermo Scientific, catalog number: 5100-0001)
9. Tissue-Tek cold plate (VWR Scientific, catalog number: 25608-942)

Procedure

A. 50× fixation concentrate stock preparation (50 mg/mL)

1. DSP equilibration
   Thoroughly equilibrate the DSP package to room temperature for 30 min prior to first opening. **Critical:** Only open the package **after** equilibration.
   Note: The NHS-ester of DSP is susceptible to hydrolysis upon contact with the humid atmosphere and is therefore moisture sensitive. Thorough equilibration is crucial in preventing condensation and premature inactivation of DSP.
2. 50× working solution stock preparation
   Dissolve 50 mg of DSP in 1 mL of high-quality anhydrous DMSO for best performance. Final concentration: 50 mg/mL.
3. Working stock storage
   a. For short-term storage, dispense 10 µL from the stock into 1.5 mL or 2 mL Eppendorf safe-lock microtubes and store at -80 °C in a sealed container or bag with desiccants.
   b. Alternatively, 10 µL aliquots can be stored in cryogenic vials with compression O-rings to prevent moisture absorption. Avoid freeze-thaw cycles by limiting the size of aliquots (up to 100 µL). Any opened and unused stock must be discarded.

B. Working solution preparation (1 mg/mL)

Note: ***Time-sensitive step.*** *Prepare fresh working solution immediately before the fixation process.*

1. Using a P200 pipette, slowly and drop-by-drop add 490 µL of PBS to an Eppendorf tube containing 10 µL of 50× stock **while vortexing** (Video 1).
   Caution: Choose "touch mode" and maximal speed (3,500 rpm or no less than 2,500 rpm) on the vortex instead of the "on mode". Using a P1000 pipette is acceptable too, but adding the PBS too fast will cause DSP to precipitate, which perturbs the effectiveness of fixation. It is not concerning to see tiny precipitates on the wall of the tube initially; however, in case of substantial precipitates, repeat by slowing down the PBS addition or start with a new DSP aliquot.

Video 1. Working solution preparation

2. Filter the working solution once using a 40 µm Flowmi strainer and transfer it to a new 1.5 or 2 mL tube to remove larger precipitates.
3. **Critical:** Working solution standing for more than **10 min** should not be used and must be discarded. If sample preparation takes longer than 10 min, prepare the working solution during the process to avoid leaving it standing.

C. Fixation

Note: For this section, follow instructions based on your sample type.

1. Fixation of tissue samples
   a. For larger tissue samples or organoids (3 mm or larger in diameter/edge length), finely cut or mince the tissue with a sterile razor blade on a cold plate to facilitate working solution penetration. For effective fixation, tissue samples should be no thicker than 3 mm, with 1–2 mm being preferable for uniformity. This is because fixatives penetrate through passive diffusion, and the higher molecular weight of DSP (14.5-fold paraformaldehyde and 4-fold glutaraldehyde) significantly reduces its penetration speed and depth in tissue.
   Note: Fine mincing without separating the tissue minimizes sample loss during washing and transferring into a new tube.
   b. Add the minced tissue to a new Eppendorf tube containing 500 µL of working solution and incubate at ambient temperature for 30–45 min.
   c. Gently invert the tube every 15 min. For certain tissue types, if the tube becomes overly bloody or

cloudy, consider replenishing once with fresh working solution.

2. Fixation of cells
   a. Wash the cells in cold, sterile RNase-free PBS.
      i. Centrifuge the cells using an appropriate speed depending on type and fragility: usually between 150× *g* for 3 min (large cells) and 600× *g* for 10 min (small cells). Keep the cell pellet and remove the supernatant. Resuspend in cold, sterile RNase-free PBS.
      ii. Repeat this step for a minimum of two washes.
      Note: Remnant media containing FBS (see General notes, Tip 4), storage additives, or sheath fluid components may interfere with the crosslinking.
   b. Pellet the cells, resuspend in 500 µL of working solution, and incubate for 30 min at ambient temperature.
   c. Gently invert the tube once at the 15-min mark.

D. Quenching of reactive DSP

Note: For this section, follow instructions based on your sample type.

1. For cells
   a. Add 10 µL of 1 M Tris-HCl pH 7.5. Then, mix thoroughly on a vortex for 2–3 s followed by a 15 min incubation at room temperature.
   Note: As an amine-reactive crosslinker, excessive reactive DSP must be quenched with Tris-HCl buffer before proceeding. The quantity is stoichiometrically optimized and must be adjusted accordingly if more fixative was used for larger tissues or a higher number of cells.
   b. Centrifuge at 500× *g* for 5 min at ambient temperature. Remove supernatant.
   c. Resuspend the cell pellet in 1,000 µL of PBS. Mix by vortexing for 2–3 s.
   d. Centrifuge at 500× *g* for 5 min at ambient temperature. Remove the supernatant.
   e. Repeat steps D1c–D1d for a total of two washes.
   f. Resuspend the cells in 0.5–1 mL of cold PBS with 1% BSA.
   Note: For maximal RNA integrity, add 0.2–1 unit/µL RNase inhibitor to the final resuspension buffer.
   g. Filter the final resuspension through an appropriately sized filter for the cell type. Commonly used mesh sizes for single-cell suspensions are between 40 and 70 µm.
   h. Count cells, bring concentration to 1,000–1,500 cells per microliter, and proceed to encapsulation.
2. For tissues
   a. Add 10 µL of 1 M Tris-HCl pH 7.5. Then, mix thoroughly on a vortex for 2–3 s followed by a 15 min incubation at room temperature.
   b. Centrifuge at 500× *g* for 20 s or for 5–10 s in a mini spinner and remove supernatant.
   c. Add 1 mL of 200 µg/mL Liberase in PBS.
   d. Incubate the tissue with digestion buffer at 37 °C for 30 min with agitation at 800 rpm on a ThermoMixer C with heated lid.
   Note: The length of digestion may need to be optimized according to tissue type.
   e. Pipette to mix the sample 5–10 times every 15 min to facilitate digestion.
   f. After digestion is completed, filter the sample through a 70 µm filter and into a 15 mL falcon tube to remove coarse debris.
   g. Add 10 mL of ice-cold PBS and mix well.
   h. Centrifuge the sample at 500× *g* for 5 min in a pre-cooled centrifuge with swinging bucket rotor.
   i. Discard the supernatant and resuspend the pellet again in 10 mL of ice-cold PBS with 1% BSA. Mix well.
   j. Repeat steps D2g–D2h for three washes.
   k. Filter cells with an appropriate filter depending on cell size and resuspend the cells in 0.5–1 mL cold PBS with 1% BSA. Count cells, bring the concentration to 1,000–1,500 cells per microliter, and proceed to encapsulation.
   Note: For maximal RNA integrity, add 0.2–1 unit/µL RNase inhibitor.

E. Cryopreservation of samples (Optional)

1. Resuspend the cell pellet in 500 µL of fresh PBS or media.
2. Count and record the cell number.
3. Add an appropriate volume of chilled cryopreservation medium to obtain a cell concentration of 1–2 × 10^6 cells per milliliter.
4. Dispense cell suspension aliquots of 1–2 mL into pre-cooled cryovials and place the cryovials inside a pre-cooled cell freezing container, e.g., CoolCell FTS30, to ensure gradual freezing.
5. Place the cell freezing container in a -80 °C freezer for ≥4 h. After 4 h, transfer the cryovials to liquid nitrogen for long-term storage.
6. Thaw in a 37 °C water bath and wash cells twice with PBS + 0.5%–1% BSA before proceeding to cell encapsulation.

Validation of protocol

The current protocol has been routinely performed as a standard protocol on human, mouse, and rat tissues and primary cells at the Spatial Technologies Unit, Beth Israel Deaconess Medical Center, Harvard Medical School, being considered robust and reproducible. The whole procedure was repeated once after the completion of this manuscript by a person without previous knowledge using the current version to ensure details are correct, comprehensible, and executable.
Additional verification has been conducted in various studies, as documented, and systematically tested and optimized in Jiménez-Gracia et al. [23] (Figures 1–7) and in Aney et al. [24] (Figure 1).
Authors have validated the following sample types with successful single-cell RNA sequencing: human PBMC, mouse pancreas, mouse and rat liver, mouse lung, mouse colon, human colon biopsies, and human prostate.

General notes and troubleshooting

General notes

Tip 1: Only use freshly prepared DSP
Make single-use aliquots (20–50 µL) for 2–5 fixations. Do not re-freeze leftovers. Keeping the stock fixative away from water is key because it neutralizes the NHS-esters quickly. Currently, this can be a limitation in clinical settings, since clinical staff requires bandwidth, proper instrument setup, and training for the fixation process.

Tip 2: Avoid using Tris, glycine, or any other amine-active buffer to prepare the fixative
Avoid buffer components with primary amines such as Tris and glycine buffers, as they compete with proteins in the sample. DSP reacts with non-protonated aliphatic amine groups, including the N (amine) terminus of polypeptides and the ε-amino group of lysine (K) side chain.

Tip 3: Never bypass the stock preparation step by directly dissolving DSP solid powder in PBS
DSP is hydrophobic and needs to be dissolved in DMSO before being added to the aqueous reaction mixture. Besides the 50 mg (22586) size, DSP is also available in 10 × 1 mg (A35393) or 1 g (22585) packaging. Although a smaller size unit is more expensive per sample, consider using a smaller size matching the experimental scale to avoid exposure to moisture absorption. Lot-to-lot differences in DSP may exist; testing one aliquot on the optimization sample from the batch can be a QC option.

Tip 4: DSP crystal formation is normal but needs to be controlled; not all crystals are from DSP
DSP does not possess a charged group; it is lipophilic and membrane-permeable, making it suitable for intracellular and intramembrane crosslinking. However, it is water-insoluble and can form crystal precipitates when preparing

the working solution. These crystals are removed by filtering in step B3. However, usage of BSA and even a high percentage of FBS are common to promote cell survival and boost viability. Still, calcium oxalate crystals can often be found in commercial fetal bovine serum (FBS) [25], being mistakenly perceived as DSP crystals. Using BSA is not concerning but a high percentage of BSA will deplete reactive DSP and lead to cloudy samples.

Tip 5: Calculate viability before fixation and re-count before storage
Viability is no longer a good measure for single-cell sample quality because DSP permeabilizes the cell membrane. Therefore, it is advisable to measure viability immediately before the fixation. Some cells will inevitably be lost during wash steps, so recounting cells before freezing and storage is necessary. Because fixed cells are not alive, no additive beyond DMSO or glycerol is required, such as FBS, ascorbic acid, or cell culture media components.

Tip 6: Pay attention to other potentially interfering components
The list here is not exhaustive and only includes common examples. With the incorporation of new components into the FixNCut workflow, caution is advised. The central disulfide bridge in DSP provides a reducible link that can be cleaved by reducing agents such as DTT (dithiothreitol) or β-mercaptoethanol. In single-cell assays, DTT is commonly used to inhibit RNase activity, inactivate reverse transcription inhibitors, and dissolve gel beads by breaking their disulfide bonds, thus releasing oligonucleotides essential for mRNA capture. However, DTT also de-crosslinks DSP-induced fixation. DTT, often used with sodium dodecyl sulfate (SDS), can rapidly break down cells. Therefore, even fixed cells should not be left in the reaction mix for more than a few min before chip loading. In contrast, EDTA in cell sorting buffer does not interfere with the crosslinking process. Once cells are fixed, EDTA cannot alter intracellular protein structure through chelating divalent cations or prevent cell clumping by inhibiting calcium-dependent adhesion, making its use in FACS buffer non-essential.

Final remark: As a reactive compound, DSP esters can cause eye and skin irritation and may be harmful if swallowed or inhaled. Therefore, it should be handled with care in a controlled laboratory environment. Always follow safety guidelines when handling chemicals, including wearing appropriate personal protective equipment and working in a well-ventilated area. The waste generated in this protocol is inactive but contains trace contents of DMSO and should be disposed of according to local regulations.

Acknowledgments

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (IMI 2 JU) under grant agreement No 831434 (3TR; Taxonomy, Targets, Treatment, and Remission). The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. Also, this project has received funding from the European Union's H2020 research and innovation program under grant agreement No. 848028 (DoCTIS; Decision On Optimal Combinatorial Therapies In Imids Using Systems Approaches), and the Commonwealth Standard Grant Agreement 4-F26M8TZ. L.J.-G. has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Universities. Australian Prostate Cancer BioResource (APCB) collection in Adelaide is supported by funding from the South Australian Immunogenomics Cancer Institute and the South Australian Health and Medical Research Institute. S.W., A.A.A. and I.S.V. acknowledge support from Spatial Technologies Unit of Precision RNA Medicine Core (RRID:SCR_024905), as well as the National Institutes of Health under award number P01AI179405 and U54HL165440. The authors further thank Dr. Tatsuyuki Sato and Dr. Joji Fujisaki for their valuable feedback and discussion. The graphics were created with BioRender.com.

Competing interests

H.H. is a co-founder and shareholder of Omniscope, a scientific advisory board member of MiRXES and Nanostring, and a consultant to Moderna and Singularity. L.G.M is an advisor and shareholder of Omniscope, and advisor for ArgenTAG and BioScryb. Omniscope has filed a patent related to the application of the FixNCut protocol. All other

authors declare no competing interests.

Received: June 01, 2024; Accepted: July 28, 2024; Available online: August 13, 2024; Published: September 05, 2024

References

1. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. and Enard, W. (2018). Quantitative single-cell transcriptomics. *Briefings Funct Genomics.* 17(4): 220–232. https://doi.org/10.1093/bfgp/ely009
2. Lafzi, A., Moutinho, C., Picelli, S. and Heyn, H. (2018). Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat Protoc.* 13(12): 2742–2757. https://doi.org/10.1038/s41596-018-0073-y
3. Kazer, S. W., Aicher, T. P., Muema, D. M., Carroll, S. L., Ordovas-Montanes, J., Miao, V. N., Tu, A. A., Ziegler, C. G. K., Nyquist, S. K., Wong, E. B., et al. (2020). Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nat Med.* 26(4): 511–518. https://doi.org/10.1038/s41591-020-0799-2
4. Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C. L., Lindberg, E. L., Kanda, M., Polanski, K., Heinig, M., Lee, M., et al. (2020). Cells of the adult human heart. *Nature.* 588(7838): 466–472. https://doi.org/10.1038/s41586-020-2797-4
5. Slyper, M., Porter, C. B. M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., et al. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat Med.* 26(5): 792–802. https://doi.org/10.1038/s41591-020-0844-1
6. Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., Lister, R., et al. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* 21(1): 1–25. https://doi.org/10.1186/s13059-020-02048-6
7. de Ruiter, K., van Staveren, S., Hilvering, B., Knol, E., Vrisekoop, N., Koenderman, L. and Yazdanbakhsh, M. (2018). A field-applicable method for flow cytometric analysis of granulocyte activation: Cryopreservation of fixed granulocytes. *Cytometry Part A* 93(5): 540–547. https://doi.org/10.1002/cyto.a.23354
8. Kotsakis, A., Harasymczuk, M., Schilling, B., Georgoulias, V., Argiris, A. and Whiteside, T. L. (2012). Myeloid-derived suppressor cell measurements in fresh and cryopreserved blood samples. *J Immunol Methods.* 381: 14–22. https://doi.org/10.1016/j.jim.2012.04.004
9. Horie, M., Castaldi, A., Sunohara, M., Wang, H., Ji, Y., Liu, Y., Li, F., Wilkinson, T. A., Hung, L., Shen, H., et al. (2020). Integrated Single-Cell RNA-Sequencing Analysis of Aquaporin 5-Expressing Mouse Lung Epithelial Cells Identifies GPRC5A as a Novel Validated Type I Cell Surface Marker. *Cells.* 9(11): 2460. https://doi.org/10.3390/cells9112460
10. Alles, J., Karaiskos, N., Praktiknjo, S. D., Grosswendt, S., Wahle, P., Ruffault, P. L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., et al. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* 15(1): 1–4. https://doi.org/10.1186/s12915-017-0383-5
11. Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (1979).* 357(6352): 661–667. https://doi.org/10.1126/science.aam8940
12. Chen, J., Cheung, F., Shi, R., Zhou, H., Wenrui, W., Consortium, C., Candia, J., Kotliarov, Y., Stagliano, K. R., Tsang, J. S., et al. (2018). PBMC Fixation and Processing for Chromium Single-Cell RNA Sequencing. *J Transl Med.* 16(1): 1–11. https://doi.org/10.1101/315267
13. García-Castro, H., Kenny, N. J., Iglesias, M., Álvarez-Campos, P., Mason, V., Elek, A., Schönauer, A., Sleight, V. A., Neiro, J., Aboobaker, A., et al. (2021). ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol.* 22(1): 1–34. https://doi.org/10.1186/s13059-021-02302-5
14. Sánchez-Carbonell, M., Jiménez Peinado, P., Bayer-Kaufmann, C., Hennings, J. C., Hofmann, Y., Schmidt, S., Witte, O. W. and Urbach, A. (2023). Effect of methanol fixation on single-cell RNA sequencing of the murine dentate gyrus. *Front Mol Neurosci.* 16: e1223798. https://doi.org/10.3389/fnmol.2023.1223798

15. Wang, X., Yu, L. and Wu, A. R. (2021). The effect of methanol fixation on single-cell RNA sequencing data. *BMC Genomics.* 22(1): 420. https://doi.org/10.1186/s12864-021-07744-6
16. Lomant, A. and Fairbanks, G. (1976). Chemical probes of extended biological structures: Synthesis and properties of the cleavable protein cross-linking reagent [35S]dithiobis(succinimidyl propionate). *J Mol Biol.* 104(1): 243–261. https://doi.org/10.1016/0022-2836(76)90011-5
17. Akaki, K., Mino, T. and Takeuchi, O. (2022). DSP-crosslinking and Immunoprecipitation to Isolate Weak Protein Complex. *Bio Protoc.* 12(15): e4478. https://doi.org/10.21769/bioprotoc.4478
18. Attar, M., Sharma, E., Li, S., Bryer, C., Cubitt, L., Broxholme, J., Lockstone, H., Kinchen, J., Simmons, A., Piazza, P., et al. (2018). A practical solution for preserving single cells for RNA sequencing. *Sci Rep.* 8(1): 2151. https://doi.org/10.1038/s41598-018-20372-7
19. Subramanian Parimalam, S., Oguchi, Y., Abdelmoez, M. N., Tsuchida, A., Ozaki, Y., Yokokawa, R., Kotera, H. and Shintaku, H. (2018). Electrical Lysis and RNA Extraction from Single Cells Fixed by Dithiobis(succinimidyl propionate). *Anal Chem.* 90(21): 12512–12518. https://doi.org/10.1021/acs.analchem.8b02338
20. Reyes, M., Billman, K., Hacohen, N. and Blainey, P. C. (2019). Simultaneous Profiling of Gene Expression and Chromatin Accessibility in Single Cells. *Adv Biosyst.* 3(11): e201900065. https://doi.org/10.1002/adbi.201900065
21. Nesterenko, P. A., McLaughlin, J., Cheng, D., Bangayan, N. J., Burton Sojo, G., Seet, C. S., Qin, Y., Mao, Z., Obusan, M. B., Phillips, J. W., et al. (2021). Droplet-based mRNA sequencing of fixed and permeabilized cells by CLInt-seq allows for antigen-specific TCR cloning. *Proc Natl Acad Sci USA.* 118(3): e2021190118. https://doi.org/10.1073/pnas.2021190118
22. Mutisheva, I., Robatel, S., Bäriswyl, L. and Schenk, M. (2022). An Innovative Approach to Tissue Processing and Cell Sorting of Fixed Cells for Subsequent Single-Cell RNA Sequencing. *Int J Mol Sci.* 23(18): 10233. https://doi.org/10.3390/ijms231810233
23. Jiménez-Gracia, L., Marchese, D., Nieto, J. C., Caratù, G., Melón-Ardanaz, E., Gudiño, V., Roth, S., Wise, K., Ryan, N. K., Jensen, K. B., et al. (2024). FixNCut: single-cell genomics through reversible tissue fixation and dissociation. *Genome Biol.* 25(1): 81. https://doi.org/10.1186/s13059-024-03219-5
24. Aney, K. J., Jeong, W. J., Vallejo, A. F., Burdziak, C., Chen, E., Wang, A., Koak, P., Wise, K., Jensen, K., Pe'er, D., et al. (2024). Novel Approach for Pancreas Transcriptomics Reveals the Cellular Landscape in Homeostasis and Acute Pancreatitis. *Gastroenterology.* 166(6): 1100–1113. https://doi.org/10.1053/j.gastro.2024.01.043
25. Pedraza, C. E., Chien, Y. and McKee, M. D. (2007). Calcium oxalate crystals in fetal bovine serum: Implications for cell culture, phagocytosis and biomineralization studies in vitro. *J Cell Biochem.* 103(5): 1379–1393. https://doi.org/10.1002/jcb.21515

Appendix B. Metastatic recurrence in colorectal cancer arises from residual EMP1+ cells[5]

Adrià Cañellas-Socias, Carme Cortina, Xavier Hernando-Momblona, Sergio Palomo-Ponce, Eoghan J. Mulholland, Gemma Turon, Lidia Mateo, Sefora Conti, Olfa Roman, Marta Sevillano, Felipe Slebe, Diana Stork, Adrià Caballé-Mestres, Antonio Berenguer-Llergo, Adrián Álvarez-Varela, Nicola Fenderico, Laura Novellasdemunt, **Laura Jiménez-Gracia**, Tamara Sipka, Lidia Bardia, Patricia Lorden, Julien Colombelli, Holger Heyn, Xavier Trepat, Sabine Tejpar, Daniele V.F. Tauriello, Elena Sancho, Simon Leedham, Camille Stephan-Otto Attolini, Eduard Batlle. **Metastatic recurrence in colorectal cancer arises from residual *EMP1+* cells**, *Nature*. 611, 603–613 (2022).

DOI: https://doi.org/10.1038/s41586-022-05402-9

Impact Factor 2022: 64.8 (Q1)

Raw data: Accessible via the Sequence Read Archive (SRA), NCBI Gene Expression Omnibus (GEO) and ArrayExpress databases. Further details, including specific accession numbers, are provided in the *Data availability* section of the article.

Processed data: Accessible via the Synapse database. Accession numbers are provided in the *Data availability* section of the article.

Code: Available upon corresponding author's request.

[5] In this paper, I collaborated by analyzing the scRNA-seq data together with the Bioinformatics Facility from Institute for Research in Biomedicine (IRB).

Article

Metastatic recurrence in colorectal cancer arises from residual EMP1^{+} cells

https://doi.org/10.1038/s41586-022-05402-9

Received: 13 December 2021

Accepted: 30 September 2022

Published online: 9 November 2022

Check for updates

Adrià Cañellas-Socias[1,2], Carme Cortina[1,2], Xavier Hernando-Momblona[1,2], Sergio Palomo-Ponce[1,2], Eoghan J. Mulholland[3], Gemma Turon[1], Lidia Mateo[1], Sefora Conti[4], Olga Roman[1], Marta Sevillano[1,2], Felipe Slebe[1], Diana Stork[1], Adrià Caballé-Mestres[1], Antonio Berenguer-Llergo[1], Adrián Álvarez-Varela[1,2], Nicola Fenderico[1], Laura Novellasdemunt[1], Laura Jiménez-Gracia[5], Tamara Sipka[1], Lidia Bardia[1], Patricia Lorden[5], Julien Colombelli[1], Holger Heyn[5,6], Xavier Trepat[4,7,8,9], Sabine Tejpar[10], Elena Sancho[1,2], Daniele V. F. Tauriello[1,11], Simon Leedham[3,12,13], Camille Stephan-Otto Attolini[1] & Eduard Batlle[1,2,9]✉

Around 30–40% of patients with colorectal cancer (CRC) undergoing curative resection of the primary tumour will develop metastases in the subsequent years[1]. Therapies to prevent disease relapse remain an unmet medical need. Here we uncover the identity and features of the residual tumour cells responsible for CRC relapse. An analysis of single-cell transcriptomes of samples from patients with CRC revealed that the majority of genes associated with a poor prognosis are expressed by a unique tumour cell population that we named high-relapse cells (HRCs). We established a human-like mouse model of microsatellite-stable CRC that undergoes metastatic relapse after surgical resection of the primary tumour. Residual HRCs occult in mouse livers after primary CRC surgery gave rise to multiple cell types over time, including LGR5^{+} stem-like tumour cells[2–4], and caused overt metastatic disease. Using *Emp1* (encoding epithelial membrane protein 1) as a marker gene for HRCs, we tracked and selectively eliminated this cell population. Genetic ablation of EMP1^{high} cells prevented metastatic recurrence and mice remained disease-free after surgery. We also found that HRC-rich micrometastases were infiltrated with T cells, yet became progressively immune-excluded during outgrowth. Treatment with neoadjuvant immunotherapy eliminated residual metastatic cells and prevented mice from relapsing after surgery. Together, our findings reveal the cell-state dynamics of residual disease in CRC and anticipate that therapies targeting HRCs may help to avoid metastatic relapse.

Surgical resection of primary CRC effectively cures most patients diagnosed with locoregional disease[1]. However, around 5% of patients with AJCC stage I, 15% of stage II and 40% of stage III CRC will develop metastases in the following years[1]. We and others have previously shown that the vast majority of genes that predict high risk of disease relapse in CRC are expressed by cells of the tumour microenvironment (TME), particularly by cancer-associated fibroblasts (CAFs)[5–7]. To further investigate this finding, we mapped the poor-prognosis CRC transcriptome at the single-cell level. Using a large pooled transcriptomic cohort of primary CRC samples (Supplementary Table 1), we identified 2,530 genes that predicted disease relapse (Fig. 1a). Subsequently, the expression of this poor-prognosis gene set was analysed in two independent single-cell RNA-sequencing (scRNA-seq) CRC datasets that included both tumour epithelial and microenvironment cells[8] (Fig. 1b–d and Extended Data Fig. 1a–c). Supporting our previous findings, CAFs, endothelial cells and, to a lesser extent, myeloid cells expressed the highest levels of genes associated with poor prognosis (Fig. 1b and Extended Data Fig. 1a). However, detailed analysis of the recurrence gene set in specific cell populations purified from primary CRC samples from patients[6] revealed that 99 out of the 2,530 genes were upregulated in epithelial tumour cells compared with TME cells (Fig. 1a and Supplementary Table 2). Indeed, these 99 recurrence-associated genes (hereafter, epithelial-specific high-risk gene set (EpiHR)), showed epithelial tumour cell-restricted expression patterns in the Samsung Medical Center (SMC) (Fig. 1d) and Katholieke Universiteit Leuven (KUL) (Extended Data Fig. 1c)

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain. [3]Gastrointestinal Stem Cell Biology Lab, Wellcome Centre Human Genetics, University of Oxford, Oxford, UK. [4]Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [5]CNAG-CRG, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [6]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [7]Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain. [8]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain. [9]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [10]Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium. [11]Department of Cell Biology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands. [12]Translational Gastroenterology Unit, John Radcliffe Hospital, University of Oxford, Oxford, UK. [13]Oxford National Institute for Health Research Biomedical Research Centre, Oxford, UK. ✉e-mail: eduard.batlle@irbbarcelona.org

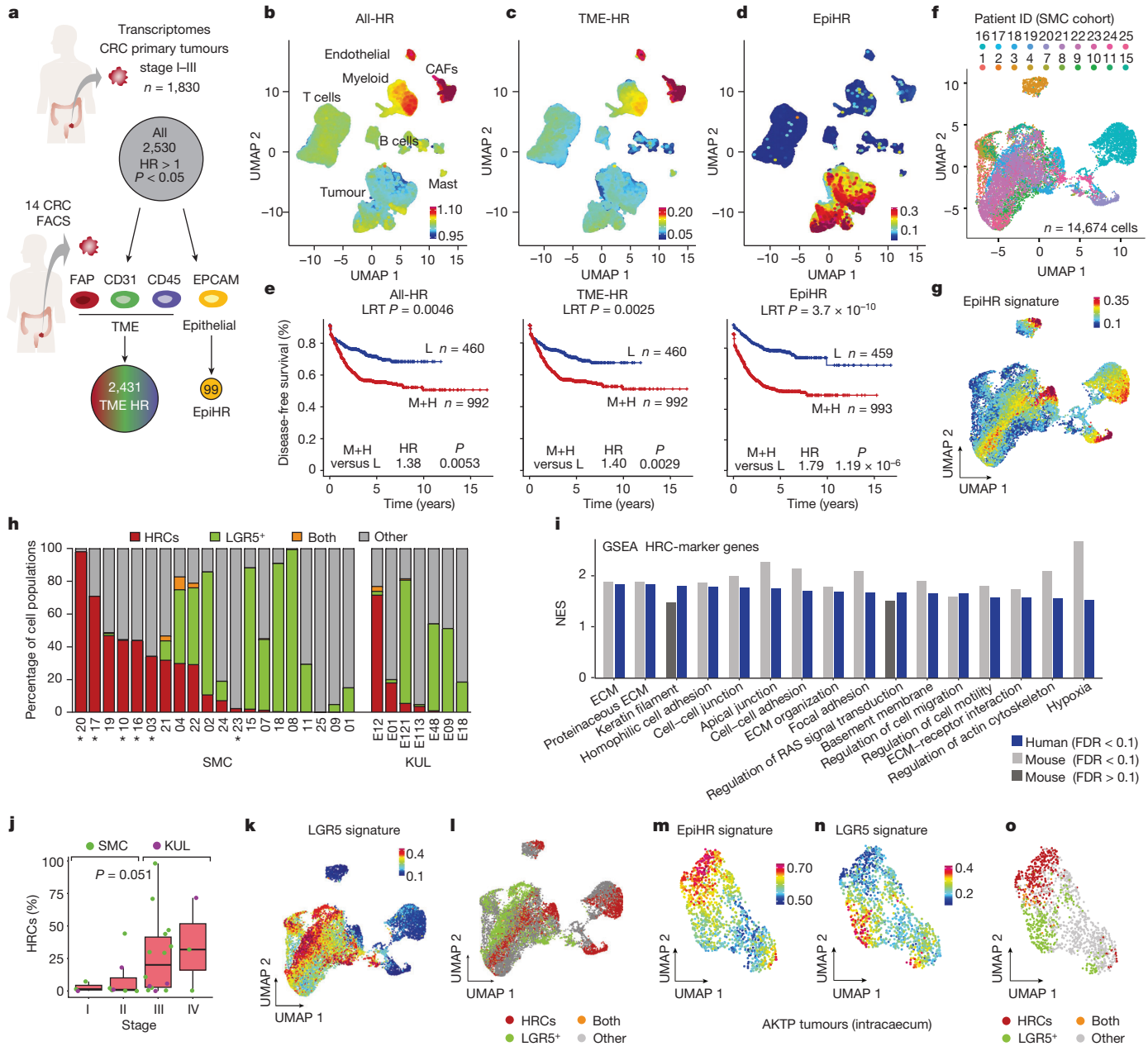


Fig. 1 | Identification of epithelial CRC cells associated with poor prognosis. a, Identification of TME-HR and EpiHR signatures in a metacohort of 7 pooled human stage I–III CRC datasets (Supplementary Table 1). **b**–**d**, UMAP layout of whole tumours (stroma + epithelium cells) belonging to the SMC dataset, coloured by all high-risk (All-HR) relapse (**b**), TME-HR (**c**) and EpiHR (**d**) genes. **e**, Kaplan–Meier survival curves showing relapse-free survival. Statistical analysis was performed using a two-sided Wald test. High (H), medium (M) and low (L) expression levels of All-HR, TME-HR and EpiHR gene signatures; LRT, likelihood-ratio test. **f**,**g**, UMAP layout of 14,674 CRC tumour cells (SMC cohort) coloured by patient ID (**f**) and by the expression of the EpiHR signature (**g**). **h**, The subpopulation composition of each patient. Patient ID is detailed. Patients with low WNT-signature scores are indicated by an asterisk. **i**, Selected Hallmark, GOSLIM and KEGG gene signatures enriched in HRCs compared with the rest of tumour cells in human and mouse CRC samples. NES, normalized enrichment score. ECM, extracellular matrix. **j**, The proportion of HRCs in each clinical stage. The box plots show the median (centre line) first and third quartiles (box limits), and a maximum of 1.5× the interquartile range (whiskers). From left to right, $n = 3$, $n = 7$, $n = 14$ and $n = 3$ patients. Statistical analysis was performed using two-sided Kruskal–Wallis tests. **k**, UMAP analysis of tumour cells coloured by expression of the *LGR5* signature. **l**, UMAP analysis of tumour cells labelled according to their classification as HRCs, LGR5$^+$, double-positive or other cells. **m**–**o**, Primary tumours were generated in the caecum of C57BL/6J mice by injecting syngeneic AKTP MTOs. UMAP plots depicting eGFP$^+$ tumour cells dissociated from primary tumours coloured by expression levels of EpiHR signature (**m**), *LGR5* signature (**n**) and their classification as HRCs, LGR5$^+$, double-positive or other cells (**o**). Colour scales in **b**–**d**,**g**,**k**,**m**–**o** indicate expression levels of the indicated gene signatures.

scRNA-seq cohorts. EpiHR expression levels predicted recurrence with an accuracy equivalent to the subset of TME poor-prognosis genes (Fig. 1e). In a multivariate analysis including the two signatures and clinical variables, the TME high risk (TME-HR) and EpiHR gene sets were independent prognostic factors (EpiHR: hazard ratio (HR) (+1 s.d.) = 2.26, $P = 1.2 \times 10^{-7}$; TME-HR: HR (+1 s.d.) = 1.74, $P = 8 \times 10^{-4}$). The EpiHR signature was significantly associated with right-side colon cancer and to AJCC stages III–IV (Extended Data Fig. 1d). Moreover, it stratified CRC patients within each consensus molecular subtype[9] into high and low risk of relapse (Extended Data

Fig. 1e). Thus, the EpiHR gene set encodes determinants of disease relapse with epithelial tumour-cell-specific expression.

Identification of EpiHR$^+$ tumour cells

Representation of tumour epithelial CRC cells using uniform manifold approximation and projections (UMAP) analysis showed that 18 out of 27 CRCs contained cells labelled with the EpiHR gene set in proportions ranging from 1.4% to 98.1% (Fig. 1f–h and Extended Data Fig. 1f,g). We named this tumour cell population high-relapse cells (HRCs). An analysis of the EpiHR signature revealed a large number of highly correlated genes exhibiting overlapping expression in HRCs (gene clusters 1 and 3 in Extended Data Fig. 1h,i and Supplementary Table 2). The EpiHR contains only epithelial-specific genes, reflecting only a fraction of the HRC transcriptome. We further identified a core gene expression program upregulated in HRCs of most tumour samples (coreHRC; Extended Data Fig. 2a,b and Supplementary Table 2). HRCs belonging to different CRCs displayed a shared enrichment pattern of annotated gene sets, implying that they co-opt a similar phenotype and have equivalent functions (Extended Data Fig. 2c,d). Differential expression analysis of HRCs versus non-HRCs revealed prominent enrichment in genes related to hypoxia, cell-to-cell adhesion, extracellular matrix, actin cytoskeleton and regulation of cell migration (Fig. 1i). Among them (Supplementary Table 2), the collagen-sensing receptor DDR1, the integrins α2, α3 and β4, and the protease PLAUR have been repeatedly associated with tumour cell invasion, extravasation and metastasis in multiple cancer types. Incidentally, we also discovered that the signature of basal-like pancreatic cancer cells[10] marked both human and mouse HRCs, suggesting that they are akin to the most aggressive subtype of pancreatic cancer (Extended Data Fig. 2e and Supplementary Table 3). Further reinforcing our observations, most scRNA-seq samples containing abundant EpiHR$^+$ cells were stage III and IV CRCs (Fig. 1j; $P = 0.051$ stage I + II versus stage III + IV).

Widespread evidence has demonstrated that CRC growth is driven by a subset of LGR5$^+$ stem-cell-like tumour cells[2–4]. However, our analyses revealed that HRCs represent a distinct cell population, shown by their mutually exclusive distribution in UMAP analyses (Fig. 1k–l and Extended Data Fig. 2f,g). Only one tumour sample exhibited a significant number of HRCs co-expressing the LGR5 signature (7.79% sample SMC04; Fig. 1h), whereas five others included a minimal fraction (<3%) of cells marked by both *LGR5* and the HRC gene programs. The expression patterns of several WNT/intestinal stem cell marker genes confirmed that HRCs were not LGR5$^+$ stem-like tumour cells (Extended Data Fig. 2h–o). Indeed, a subset of CRCs exhibited marginal WNT target gene expression levels (Extended Data Fig. 2p,q) yet contained HRCs (Fig. 1h (asterisk)). Furthermore, HRCs represented a subset of cells in both epithelial intrinsic consensus molecular CRC subtypes iCMS2 and iCMS3[11] (Extended Data Fig. 2r).

We previously described that mice bearing mutations in *Apc*, *Kras*, *Tgfbr2* and *Trp53* (AKTP mice) in LGR5$^+$ intestinal stem cells (ISCs) develop human-like metastatic CRCs[12]. We analysed CRCs generated by implantation of AKTP mouse tumour organoids (MTOs) into the caecum of c57BL/6 mice using scRNA-seq (Fig. 1m–o). Mirroring the observations in human tumour samples, we found that mouse CRCs contained abundant HRCs and that this population did not express the LGR5$^+$ ISC-like expression program (Fig. 1m–o and Extended Data Fig. 2s). HRCs were enriched in similar gene categories in both species, including the pancreatic basal-like signature, implying functional equivalence (Fig. 1i and Extended Data Fig. 2s,t). Bulk RNA-seq analysis of our MTO biobank[12] showed elevated EpiHR and coreHRC gene signatures expression in AKTP MTOs derived from metastases compared with those established from primary CRCs (Extended Data Fig. 2u).

Dynamics of metastatic cell states

We developed a new mouse model of metastatic relapse that enabled us to investigate the contribution of HRCs to metastatic recurrence. In brief, we innovated classical needle-based orthotopic injections by relocating them to the tip of the caecum, which enabled complete surgical excision of singular invasive CRCs (Fig. 2a, Extended Data Fig. 3a–c and Supplementary Video 1). Dual eGFP–luciferase-labelled AKTP MTOs grew rapidly in the caecum of c57BL/6 mice, colonized the adjacent mucosa and generated invasive cancers (Extended Data Fig. 3a,b). Bioluminescence imaging (BLI) revealed that mice remained free of primary disease after surgical resection yet, over the following days, they relapsed in the form of liver metastases (Fig. 2b,c and Extended Data Fig. 3d,e). Occasionally, we also observed metastases in the mesenteric lymph nodes, lungs, peritoneum and diaphragm (Extended Data Fig. 3f–h). Primary tumour resection shortly after implantation cured all mice, whereas surgery at later points resulted in increased proportions of mice developing metastatic recurrences (Fig. 2b). In early primary tumour resection experiments (days 11–15), faint bioluminescence could be detected ex vivo in some livers immediately after surgery, implying the presence of residual disseminated tumour cells (DTCs). Light-sheet 3D fluorescence imaging revealed 3-to-10-cell micrometastases at the time of resection (Fig. 2d). On the basis of these observations, we established that 4–5 weeks after implantation is the optimal timepoint to enable a complete surgical resection of the primary CRC. We also developed a CRC relapse model based on the implantation of AKP MTOs. These triple-mutant CRCs exhibited delayed kinetics of metastatic recurrence after surgery (Extended Data Fig. 3i).

We next sought to profile tumour cells along the process of relapse. Isolation of residual tumour cells from large organs has been a major hurdle in cancer research[13]. We devised a tissue dissociation strategy that enriched for residual tumour cells from whole-liver samples (Extended Data Fig. 3j). In brief, during tissue preparation for fluorescence-activated single cell sorting (FACS), the vast majority of luciferase$^+$ tumour cells were retained in 100 µm filters after mild enzymatic digestions, whereas most parenchymal liver cells flowed through in these conditions (Extended Data Fig. 3k,l). By redigesting cells retained in the filter, we obtained a 400-fold eGFP–luciferase$^+$ enriched metastatic tumour cell preparation (Extended Data Fig. 3m). This step enabled the purification of residual tumour cells from individual livers exhibiting absent or very low ex vivo bioluminescence (Extended Data Fig. 3n). Using this approach, we profiled by Smart-seq2 scRNA-seq 900 eGFP$^+$ tumour cells derived from livers collected at different time points after implantation of MTOs as well as from their corresponding primary CRCs (Fig. 2e). We confirmed the presence of micrometastases, small metastases or macrometastases on the basis of bioluminescence measurements in the resected livers (Extended Data Fig. 3n).

A UMAP representation showed that tumour cells from primary tumours and metastases overlapped to a large extent (Fig. 2e,f and Extended Data Fig. 4). Hierarchical clustering analysis identified six cell clusters (Fig. 2g and Extended Data Fig. 4b,c). Cluster 0 included cells that expressed elevated levels of proliferation and biosynthesis-encoding genes (Fig. 2g and Extended Data Fig. 4b,c). LGR5$^+$ ISC-like tumour cells occupied two different clusters depending on high (cluster 1) and low (cluster 2) expression levels of the Ki-67 proliferative signature[14] (Fig. 2g and Extended Data Fig. 4c). Tumour cells of cluster 3 upregulated the differentiation marker *Krt20*. Clusters 4 and 5 were largely enriched in HRCs. Some cells in cluster 4 expressed *Krt20* suggesting that HRCs can also undergo differentiation (Fig. 2g and Extended Data Fig. 4b,c). Quantification of cell types revealed a dynamic distribution of cell populations across the metastatic relapse process (Fig. 2h). Primary CRCs and macrometastases exhibited similar distribution of cell populations, including proliferative cells, LGR5$^+$ ISC-like cells, KRT20$^+$ differentiated tumour cells and HRCs, most of which were also KRT20$^+$ (Fig. 2h). By contrast, micrometastases were largely enriched in undifferentiated

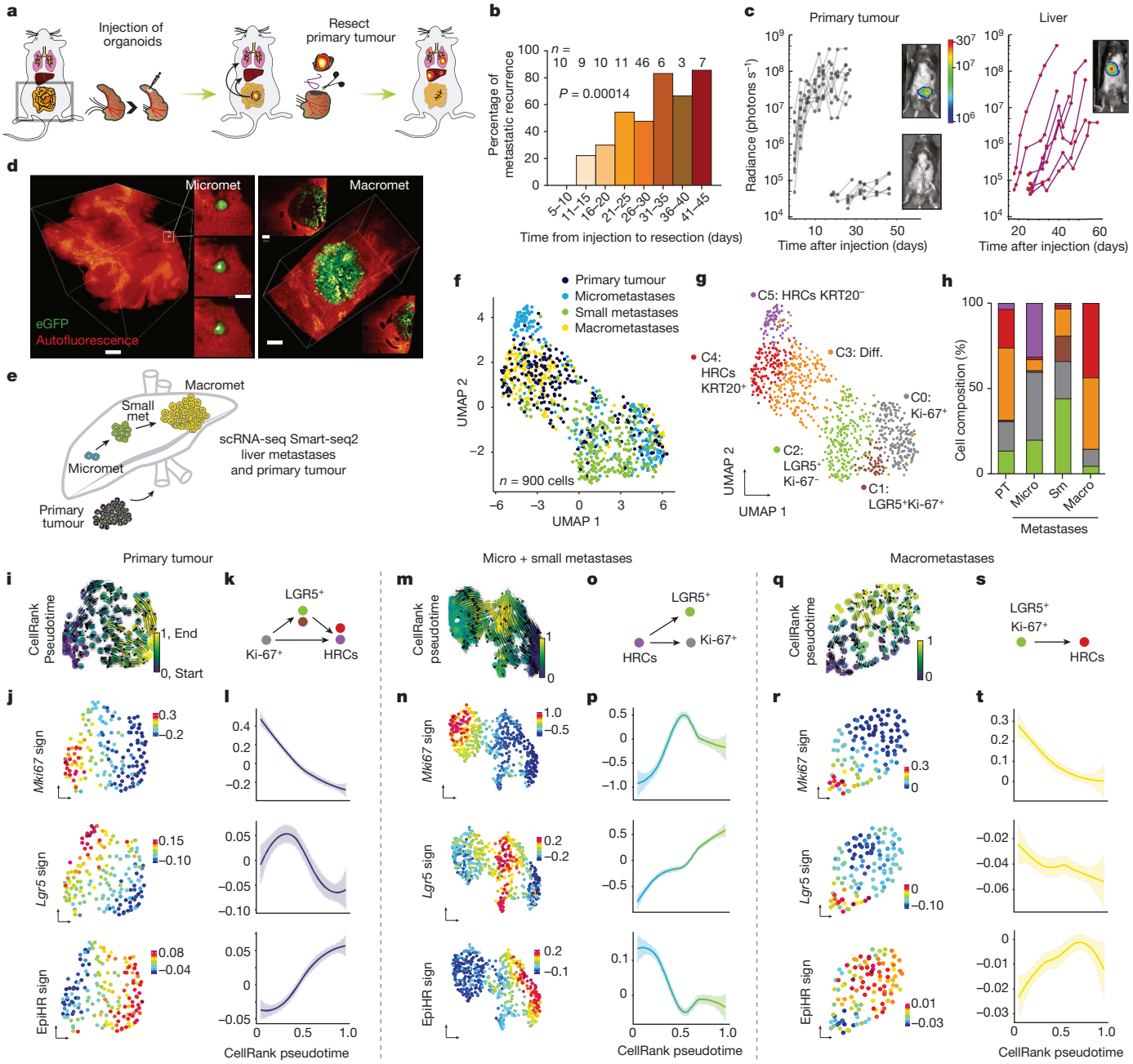


Fig. 2 | Spatiotemporal dynamics of CRC metastases resolved by scRNA-seq. a, Schematic of the mouse model of CRC metastatic relapse developed in this study. **b**, The percentage of metastatic recurrence depending on time to primary tumour resection. The number of mice is shown above the plot. The *P* value was calculated using a generalized linear model. **c**, Intravital BLI quantification (photons s^{-1}) of a representative experiment. The grey points and lines represent BLI in the lower thorax of individual mice and the purple points represent BLI in the liver. Representative images of BLI in the same mouse before surgery, after surgery and after liver metastasis formation are shown. Colour scale indicates intensity of radiance. **d**, Representative images of whole livers containing GFP-expressing tumour cells obtained using light-sheet microscopy. Scale bars, 300 µm (left, main), 50 µm (left and right insets) and 100 µm (right, main). Macromet, macrometastases; micromet, micrometastases. **e**, Illustration of the longitudinal single-cell RNA-expression analysis of tumour cells along the metastatic cascade. Met, metastases. **f**,**g**, UMAP layout of 900 tumour cells isolated from 7 different mice coloured by metastatic stage (**f**) and Seurat clusters (**g**). Diff., differentiated. **h**, Seurat cluster composition by sample stage. PT, primary tumour; Sm, small. **i**,**m**,**q**, Vector fields representing RNA velocity projected onto UMAPs of primary tumours (**i**), micro + small metastases (**m**) and macrometastases (**q**), coloured by the pseudotime estimated for each cell with scVelo. **j**,**n**,**r**, UMAPs with cells separated in primary tumours (**j**), micro + small metastases (**n**) and macrometastases (**r**) and coloured by gene expression of *Mki67*, *Lgr5* and EpiHR gene signatures (sign). **k**,**o**,**s**, Schematics showing distinct hierarchical behaviour during the different stages of metastasis progression in primary tumours (**k**), micro + small metastases (**o**) and large metastases (**s**). **l**,**p**,**t**, Smoothed *Mki67*, *Lgr5* and EpiHR gene signature expression trends fitted with generalized additive models as a function of pseudotime in primary tumours (**l**), micro + small metastases (**p**) and large metastases (**t**). Colour scales in **i**,**m**,**q** indicate start to end of trajectory in pseudotime. Colour scales in **j**,**n**,**r** indicate, expression levels of the indicated gene signatures.

(KRT20$^-$) HRCs and also contained abundant proliferative cells (Fig. 2h). Small metastases were mainly formed by LGR5$^+$ ISC-like cells in both Ki-67$^+$ and Ki-67$^-$ states and contained fewer HRCs than micrometastases (Fig. 2h and Extended Data Fig. 4d).

The trajectory-inference algorithm CellRank[15] predicted diverse hierarchical organizations during metastatic progression. In primary CRCs, proliferative LGR5$^-$ tumour cells gave rise to LGR5$^+$ cells and HRCs (Fig. 2i–l). In macrometastases, the apex of the hierarchy was occupied by proliferative LGR5$^+$ ISC-like cells, which generated HRCs over time (Fig. 2q–t). By contrast, the algorithm prognosticated that the cells that initiate the cellular hierarchy of micrometastases corresponded to undifferentiated HRCs in cluster 5 (Fig. 2m–p and Extended Data Fig. 4e). This cell population gave rise to LGR5$^+$ ISC-like and proliferative tumour cell progeny (Fig. 2m–p), which are abundant in small metastases (Fig. 2h and Extended Data Fig. 4d).

We made equivalent observations in AKP CRCs (Extended Data Fig. 4f–p). Abundant HRCs characterized liver AKP lesions profiled at early time points (day 35 after implantation), whereas micrometastases collected later (day 70) contained a higher proportion of LGR5$^+$ cells (Extended Data Fig. 4f–k). Micrometastatic LGR5$^+$ cells in the AKP model expressed high levels of the signature of latent MEX3A$^+$ cells (Extended Data Fig. 4l–n), which is in agreement with our previous study on the specification of this population in triple-mutant CRCs[16]. CellRank also predicted that HRCs are the origin of micrometastasis in AKP CRCs (Extended Data Fig. 4o,p).

Characterization of HRC features

A comparison of human and mouse scRNA-seq datasets revealed that a large subset of core genes was expressed consistently by HRCs of the two species (Extended Data Fig. 5a). Among them, we focused on EMP1 because it was expressed at high levels in HRCs and exhibited a great degree of overlap with the expression of the EpiHR gene set in both human (Extended Data Fig. 5b,c) and AKTP mouse CRCs (Fig. 3a and Extended Data Fig. 5d). Echoing our results with the EpiHR signature, CellRank predicted that the cell origin of metastatic relapse in AKTP (Extended Data Fig. 5e–g) and AKP (Extended Data Fig. 5h,i) tumours expressed elevated *Emp1* levels.

Thus, we leveraged *Emp1* to track HRCs during disease relapse. To this end, we knocked-in an inducible caspase-9-tdTomato (iCT) cassette[2] into the *Emp1* locus of AKTP MTOs using CRISPR–Cas9 (Fig. 3b). Inspection of knockin MTOs revealed high tdTomato (TOM) expression in a subset of tumour cells (Fig. 3b). We next inoculated *Emp1-iCT* AKTP MTOs into the caecum of c57BL/6 mice. TOM expression in dissociated epithelial tumour cells measured by flow cytometry revealed heterogeneity in *Emp1* expression (Extended Data Fig. 5j). TOM^{high} cells purified by FACS showed substantial upregulation of *Emp1* expression (Extended Data Fig. 5k). By contrast, TOM^{low} cells were characterized by the expression of ISC-specific genes such as *Lgr5* and *Smoc2* (Extended Data Fig. 5k). Gene expression profiling confirmed elevated levels of EpiHR and coreHRC signatures in EMP1–TOM^{high} cells, whereas the WNT/LGR5$^+$ ISC program was downregulated in these cells (Fig. 3c,d and Extended Data Fig. 5l).

Inspection of tissue sections evidenced that cancer cells invading the muscular layer were strongly labelled by the fluorescent reporter (Fig. 3e). In particular, isolated tumour buds and larger clusters in contact with the stroma at the edges on invasion fronts exhibited the highest TOM expression (Fig. 3e (inset) and Extended Data Fig. 6a–d). EMP1–TOM^{high} cell clusters were often found in proximity to peripheral blood vessels in mouse primary CRCs, suggesting a connection with haematogenous or lymphatic dissemination (Extended Data Fig. 6e). CRCs generated by implantation of AKP MTOs in the caecum also exhibited EMP1–TOM^{high} invasion fronts and tumour buds that overexpressed HRC-marker genes (Extended Data Fig. 6f–h).

We also examined livers of mice bearing AKTP tumours at various time points after orthotopic MTO implantation. Micrometastatic lesions trapped within portal veins and liver sinusoids were populated entirely by EMP1–TOM^{high} cells (Fig. 3f,g and Extended Data Fig. 6i,j). Fitting the scRNA-seq analyses, we identified two EMP1–TOM$^+$ subsets; one expressed KRT20 and was located mainly in the tumour cores, whereas the other was positioned at invasion fronts, lacked KRT20 and expressed the highest levels of TOM reporter (Extended Data Fig. 6k). Liver EMP1–TOM^{high} micrometastases were also KRT20 negative (Extended Data Fig. 6l). Notably, EMP1–TOM^{high} tumour buds and micrometastases were labelled with EPCAM and E-cadherin, implying that they retained an epithelial organization (Extended Data Figs. 6 and 7a). Analysis of epithelial-to-mesenchymal transition (EMT) master transcription factors showed equivalent expression levels in EMP1–TOM^{high} and EMP1–TOM^{low} cells (Extended Data Fig. 7b). scRNA-seq analysis also supported the lack of canonical EMT markers in human and mouse HRCs (Extended Data Fig. 7c,d). However, we noticed that the coreHRC signature genes *Lama3*, *Lamc2*, *Itga2* and *Plaur* belong to a recently characterized partial EMT gene expression module[17]. HRCs in mouse primary CRCs and micrometastases upregulated this partial EMT signature (Extended Data Fig. 4e). Notably, LAMC2 has previously been identified as a specific marker for tumour budding in multiple tumour types, including CRC[18], and we confirmed that its expression is correlated with *EMP1* mRNA in samples from patients with CRC (Extended Data Fig. 7e–g). KRT17—a widely used marker of the basal pancreatic cancer subtype[19]—also marked EMP1^{high} invasion fronts and tumour buds (Extended Data Fig. 7f,g). Gene set enrichment analysis (GSEA) also revealed homophilic cell adhesion and apical junctions as central features of HRCs (Fig. 1i). Fittingly, *Emp1* encodes a component of the tight junctions[20,21] and there were multiple other constituents of adherent junction, tight junction and desmosome complexes upregulated in the core HRC program, such as *Pcdh1*, *Dsc2*, *Clnd4* and *Jup* (Extended Data Fig. 7g and Supplementary Table 2). The latter encodes plakoglobin, which mediates the circulation of tumour cells as clusters and confers enhanced metastatic ability to breast cancer cells[22].

To further investigate the relationship between *Emp1* and *Lgr5* expression, we engineered AKTP MTOs bearing both *Emp1-iCT* and *Lgr5-eGFP* knockin reporter cassettes (Extended Data Fig. 8a–f). Confocal imaging of dual-labelled MTOs showed a mutually exclusive expression pattern of the two reporters (Extended Data Fig. 8a) and quantitative PCR with reverse transcription (RT–qPCR) analysis confirmed the upregulation of *Emp1* and *Lgr5* in sorted TOM$^+$ and eGFP$^+$ cells, respectively (Extended Data Fig. 8b). Primary CRCs generated from inoculation of dual-labelled MTOs into the caecum also exhibited a mutually exclusive expression pattern of EMP1–TOM and LGR5–eGFP reporters (Fig. 3h and Extended Data Fig. 8c,d). EMP1–TOM^{high} cells were substantially enriched at tumour buds, which, by contrast, contained few LGR5–eGFP$^+$ cells (Fig. 3h,i). RNA fluorescence in situ hybridization (FISH) analysis in human samples of patients with CRC also showed that *EMP1* expression was elevated at tumour invasion fronts, whereas *LGR5* marked the tumour cores in most cases (examples in Extended Data Fig. 8g–p).

Finally, we analysed the livers of mice bearing primary CRCs. LGR5–eGFP fluorescence was absent in DTCs and micrometastases, but was progressively gained during metastatic outgrowth, in a marked antithetic pattern to EMP1–TOM expression (Fig. 3j,k and Extended Data Fig. 8e,f). Together with the CellRank bioinformatic predictions, these observations suggest that HRCs are endowed with the ability to migrate and disseminate to foreign organs where they initiate metastatic outgrowth and subsequently give rise to non-HRC populations.

Determinants of the HRC population

The transcription factor YAP opposes the activity of the WNT pathway[23,24] and a YAP-driven gene program has been associated with tumour cell plasticity[25,26], regeneration[27] and metastasis formation in CRC[28]. Moreover, YAP promotes the conversion of LGR5$^+$ CRCs cells to a fetal intestine-like progenitor state during chemotherapy[16,25,29,30]. Extended

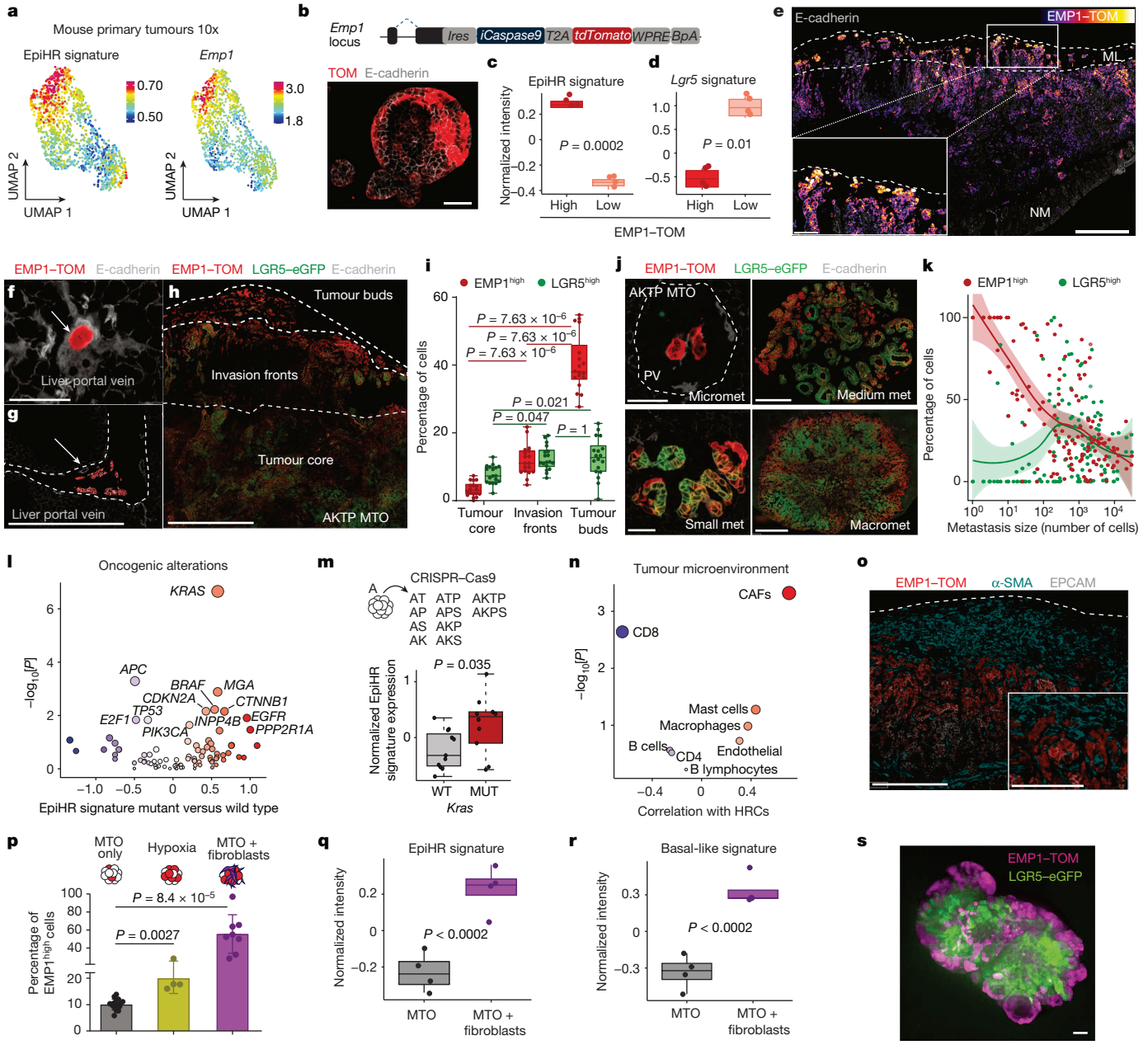


Fig. 3 | EMP1 marks cells enriched in invasion fronts and micrometastases. **a**, UMAP showing EpiHR and *Emp1* expression in mouse primary tumours. Colour scale indicates expression levels. **b**, TOM and E-cadherin immunostaining in *Emp1-iCT* AKTP MTOs. Scale bar, 50 µm. **c**,**d**, Expression (normalized intensity) of EpiHR (**c**) and *Lgr5* (**d**) signatures in EMP1^{high} and EMP1^{low} cells 4 weeks after implantation. $n = 4$ mice per condition. ROAST-GSA adjusted P values are shown. **e**, Representative images of EMP1-TOM^{+} cells in primary CRCs. The dashed lines encompass tumour buds and invasion fronts. ML, muscle layer; NM, normal mucosa. Scale bars, 500 µm (main image), 50 µm (inset). **f**–**h**, Representative images of EMP1–TOM^{+} and LGR5–eGFP^{+} cells in primary CRCs (**h**) and liver metastases (**f**,**g**). Scale bars, 500 µm (**g**,**h**), 50 µm (**f**). The dashed lines label the portal vein (**g**), and encompass the tumour buds, invasion fronts and tumour core (**h**). **i**, The percentage of LGR5–eGFP^{high} and EMP1–TOM^{high} cells in different CRC regions. The whiskers show the maximum and minimum values. $n = 855{,}330$ cells from 18 mice. Statistical analysis was performed using two-sided Wilcoxon signed-rank tests. **j**, Examples of liver metastases of increasing size. Scale bars, 50 µm (micrometastases and small metastases), 250 µm (medium metastases) and 500 µm (macrometastases). **k**, The percentage of EMP1–TOM^{high} and LGR5–eGFP^{high} cells per metastasis size. $n = 318{,}276$ cells from 137 liver metastases from 17 different mice. LOESS model with a 95% confidence interval. **l**, EpiHR levels versus driver mutations. The difference of median values (x axis) versus P values (Wilcoxon rank-sum test) (y axis) are shown. **m**, Normalized intensity of EpiHR signature expression in CTOs of different genotypes. $n = 6$ (wild-type (WT) *Kras*) and $n = 5$ (*Kras*^{G12D}) CTOs; 2 technical replicates per genotype. The P value was calculated using a two-sided t-test. **n**, The association between HRC frequency and TME cell populations in patients from the SMC cohort. Pearson correlation coefficients and P values are shown. **o**, Representative patterns of TOM, α-SMA and EPCAM in primary CRCs. The dashed lines delimit the caecum. Scale bars, 250 µm (main) and 125 µm (inset). **p**, The percentage of EMP1–TOM^{high} cells in the indicated conditions. Data are mean ± s.d. $n = 17$ (control), $n = 4$ (hypoxia) and $n = 8$ (fibroblast co-culture). Statistical analysis was performed using two-sided Wilcoxon rank-sum tests. **q**,**r**, EpiHR expression (normalized intensity) (**q**) or basal-like (**r**) signatures. $n = 4$ biological replicates. ROAST-GSA adjusted P values are shown. **s**, Representative confocal images of AKTP MTOs co-cultured with colon fibroblasts. Scale bar, 50 µm. For the box plots in **c**, **d**, **i**, **m**, **q** and **r**, the median (centre line) and first and third quartiles (box limits) are shown; the whiskers in **m**, **q** and **r** show a maximum of 1.5× the interquartile range.

Data Fig. 9a,b shows the expression of the top 50 upregulated fetal intestine progenitor genes from ref. [31], and the broadly used YAP_22 target gene signature[32]. As a reference, we also show the expression of the top 50 HRC core genes (Extended Data Fig. 9c). Only 3 out of 22 YAP core signature genes were upregulated in HRCs of patients with CRC, whereas most bona fide YAP-target genes, including *CTGF* and *CYR61*, were not (Extended Data Fig. 9a). HRCs upregulated only a few canonical markers of the regenerative fetal-like intestinal population, but the majority of them were either not expressed or were expressed at low levels by HRCs (Extended Data Fig. 9b). We also found little overlap between coreHRC and YAP_22 or fetal intestinal progenitor signatures (Extended Data Fig. 9d). EMP1–TOM$^+$ cells isolated from mouse primary CRCs were not enriched in YAP-target genes (Extended Data Fig. 9e,f). There are therefore only marginal similarities between YAP$^+$/fetal intestinal progenitors and the HRC state. To obtain further insights into this question, we knocked down YAP levels using short hairpin RNAs in MTOs or blocked YAP-driven transcription by overexpressing an inducible dominant-negative TEAD transcription factor that inhibits both YAP and TAZ activity[33] (Extended Data Fig. 9g–m). These genetic manipulations decreased the levels of the canonical YAP-target genes *Ctgf* and *Cyr61*, but did not affect the expression of core HRC genes *Emp1* or *Lamc2* in MTOs, or the abundance of EMP1–TOM^{high} cells (Extended Data Fig. 9g–m). As a control for these experiments, we used chemotherapy (folfiri), which upregulated the YAP_22 signature, including *Ctgf* and *Cyr61*, and the intestinal fetal progenitor program, but did not induce the HRC state (Extended Data Fig. 9n,o).

Association analyses in the TGCA COAD cohort revealed a strong correlation between activating *KRAS* mutations and EpiHR expression levels (Fig. 3l). Using a series of mouse organoids with compound mutations in main CRC driver genes engineered by CRISPR–Cas9 (CRISPR-derived tumour organoids (CTOs)), we confirmed that genotypes containing *Kras*^{G12D} mutations exhibited upregulation of coreHRC and EpiHR gene signatures (Fig. 3m and Extended Data Fig. 10a,b). We next searched for associations between HRCs and TME composition using the scRNA-seq dataset of patients with CRC. This analysis revealed a direct correlation between CAF and HRC abundance (Fig. 3n). Consistent with this finding, α-SMA$^+$ CAFs surrounded EMP1–TOM^{high} cells in primary AKTP CRCs (Fig. 3o). Co-culture of AKTP MTOs with CAFs augmented sixfold EMP1–TOM$^+$ cell numbers (Fig. 3p), but *YAP* knockdown did not block this effect (Extended Data Fig. 10c). Expression profiling of MTO cells isolated from co-cultures demonstrated upregulation of EpiHR, coreHRC and basal pancreatic cancer signature levels (Fig. 3q,r and Extended Data Fig. 10d). Furthermore, an arrangement of tumour cell types reminiscent of the in vivo organization was evident—LGR5–eGFP cells were positioned at the organoid centre, whereas EMP1–TOM$^+$ cells relocated to the boundaries in contact with the fibroblast population (Fig. 3s and Extended Data Fig. 10e). Fibroblasts surrounding MTOs appeared to be activated, as shown by the expression of α-SMA (Extended Data Fig. 10f). We also noticed that some organoids gained expression of the pancreatic cancer basal cell marker KRT17 (Extended Data Fig. 10g).

Relapse by residual EMP1$^+$ cells

We next leveraged the inducible caspase-9 cassette inserted in the *Emp1* locus to perform cell-ablation experiments in intact tumours[2,34] (Fig. 4a,b). Inoculation of mice with AP20187 (DIM) dimerized the chimeric caspase-9 expressed under the *Emp1* locus and specifically killed cells expressing the highest levels of EMP1–TOM reporter (Fig. 4c–e). This effect was reversible, shown by a slow but progressive recovery of the EMP1–TOM^{high} cell population at invasion fronts after cessation of treatment with DIM (Extended Data Fig. 10h–j). DIM was administered only during primary tumour growth, but treatment ceased the day before primary CRC resection (Fig. 4b). Macrometastases were not yet present when DIM treatment finished; thus, this experimental setting aimed to ablate EMP1^{high} cells in the primary tumour and possibly in incipient metastatic lesions. Notably, although DIM treatment did not affect primary tumour growth (Fig. 4f), the majority of mice showed no signs of liver and lung metastatic recurrence and were disease-free at the experimental end points (Fig. 4g,h and Extended Data Fig. 10k). We made equivalent observations in AKTP CRC growing in nude mice, therefore ruling out that the genetic cell ablation strategy impaired metastatic progression through activation of the adaptive immune system (Extended Data Fig. 10l–n). When EMP1^{high} cell ablation started 1 week after primary CRC resection, we observed no changes in metastatic progression and all of the mice suffered metastatic relapse (Fig. 4i,j). EMP1^{high} cell ablation did not halt metastasis formation when MTOs were directly inoculated in the liver through the spleen (Fig. 4k,l). Reinforcing this observation, we found that intrasplenic inoculation of EMP1^{high} cells isolated from MTOs generated more and larger liver metastases compared with either LGR5^{high} cells or EMP1^{low}LGR5^{low} cells, but the difference was not substantial, suggesting that all of these tumour cell populations had metastasis-initiating ability in this assay (Extended Data Fig. 10o–r). Moreover, metastases produced by these three cell populations contained the other tumour cell types owing to extensive cell plasticity (Extended Data Fig. 10s). Overall, these data demonstrate that EMP1^{high} HRCs drive metastatic relapse after primary CRC resection, yet they are dispensable after metastatic seeding is completed. This model is further supported by the observation that HRC ablation just after surgery and during metastatic outgrowth decreased the number of small metastases significantly, but did not modify the frequency of large metastases (Extended Data Fig. 10t–w).

Despite the slow-growing metastatic lesions generated by the AKP CRCs, HRC ablation before primary tumour surgery also prevented metastatic recurrence in these triple-mutant models (Extended Data Fig. 11a–c). Moreover, we demonstrated that HRCs mediated recurrence in CRCs with mutations in *Smad4* (AKPS) (Extended Data Fig. 11d–f). To extend our observations beyond the colon-to-liver metastasis axis, we implanted AKTP MTOs into the rectum, which generated invasive rectal cancers that, as in humans, metastasized preferentially to the lungs (Extended Data Fig. 11g–k). Inspection of lung metastases revealed that, akin to liver metastases, micrometastases were mostly EMP1^{high}, whereas larger metastases decreased the percentage of EMP1^{high} cells (Extended Data Fig. 11i). Although we could not surgically resect mouse rectal tumours, ablation of EMP1^{high} cells caused a 20-fold reduction in lung metastasis burden in this model (Extended Data Fig. 11k).

Using a diphtheria toxin receptor (DTR)-based ablation strategy in CRC models, it was previously shown that LGR5$^+$ CRC cells are necessary for liver metastasis formation[3]. To assess the role of LGR5$^+$ tumour cells in our relapse models, we knocked-in a DTR cassette into the *Lgr5* locus of AKTP MTOs (Extended Data Fig. 11l,m). Inoculation of this MTO line into the caecum further validated our previous observations that invasion fronts, tumour buds and micrometastases seldom contain LGR5$^+$ cells (Extended Data Fig. 11n–q). Importantly, treatment with diphtheria toxin (DT) before surgical removal of the primary CRC effectively eliminated LGR5$^+$ cells (Fig. 4m–p), yet it did not prevent disease relapse and mice developed overt liver metastatic disease (Fig. 4q–s). Thus, LGR5$^+$ cells are dispensable for dissemination and metastatic colonization. We validated these results using an independent AKTP MTO line engineered with an iCT cassette knocked-in into the *Lgr5* locus (Extended Data Fig. 11r–t). Again, effective ablation of LGR5$^+$ cells in the iCT model (Extended Data Fig. 11u–w) neither altered CRC tumour growth (Extended Data Fig. 11x) nor prevented metastatic recurrence (Extended Data Fig. 11y). By contrast, ablation of LGR5$^+$ cells after direct inoculation of MTOs into the liver through the portal vein halted metastasis formation (Fig. 4t,u), further supporting a requirement for LGR5$^+$ cells during metastatic outgrowth[3].

Neoadjuvant immunotherapy avoids relapse

We previously showed that mouse primary AKTP CRCs are models of microsatellite-stable/mismatch-repair-proficient CRCs characterized by a relatively low mutational burden, abundant stroma cells and

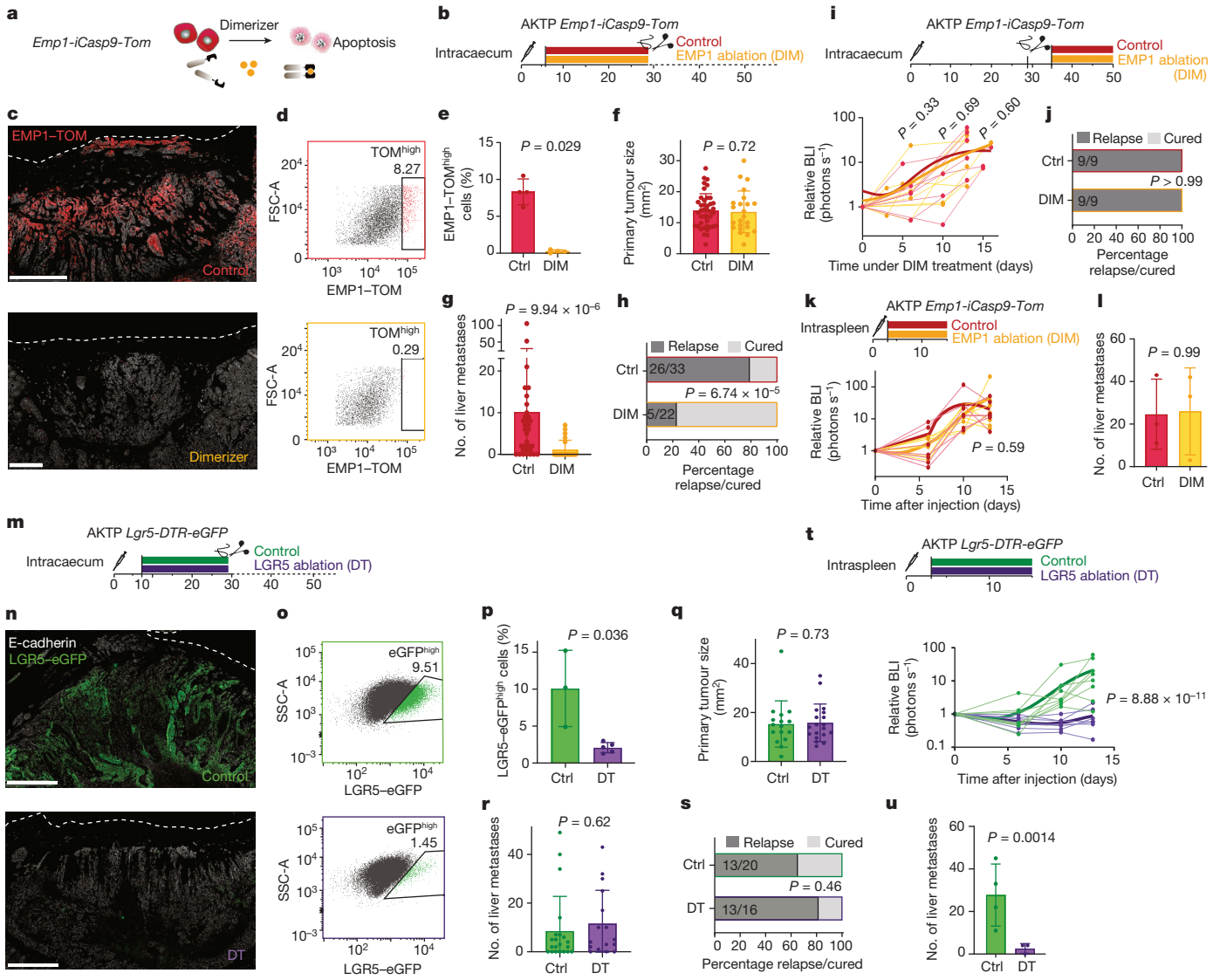


Fig. 4 | EMP1^{high} cells are the origin of metastatic relapse. a, EMP1–iCasp9-mediated cell ablation. **b**, Inducible EMP1$^+$ cell ablation and surgery. **c**, EMP1–TOM$^+$ cells in primary tumours after ablation. The lines mark the caecum borders. Scale bars, 500 µm (top) and 250 µm (bottom). **d**, Representative flow cytometry plot of TOM fluorescence in EPCAM$^+$ cells. **e**, The percentage of EMP1–TOM^{high} tumour cells at the time of surgery. Data are mean ± s.d. Every dot represents a mouse. $n = 4$ in both groups. *P* values were calculated using two-sided Wilcoxon rank-sum tests. Ctrl, control. **f**, Primary tumour area after resection. Data are mean ± s.d. Each dot represents a mouse. $n = 33$ (control) and $n = 22$ (DIM) mice. Statistical analysis was performed using a linear model after Box–Cox transformation. **g**, Liver metastases at the experimental end points. Data are mean ± s.d. Each dot represents a mouse. *n* values are as described in **f**. **h**, The percentage of mice that relapse with liver metastases. *P* values were calculated using two-sided Fisher's tests. **i**, Late ablation of EMP1$^+$ cells and metastatic growth on the basis of BLI monitoring. $n = 9$ mice per group. **j**, The percentage of mice that relapse with liver metastases. *P* values were calculated using two-sided Fisher's tests. **k**, Metastatic growth on the basis of BLI monitoring after ablation of EMP1$^+$ cells 3 days after intrasplenic inoculation of *Emp1-iCasp9-Tom* organoids. $n = 3$ mice per each group. **l**, The number of liver metastases in **k**. Data are mean ± s.d. $n = 3$ mice per group. **m**, The experimental timeline for **n**–**s**. **n**, Representative stainings of LGR5–DTR–eGFP$^+$ cells in primary CRCs. Dashed lines outline the serosa. Scale bars, 100 µm. **o**, Representative flow cytometry plot of LGR5–eGFP fluorescence. **p**, The percentage of LGR5–eGFP^{high} tumour cells. Data are mean ± s.d. $n = 3$ (control) and $n = 5$ (DT) mice. The *P* value was calculated using a generalized linear model. **q**, Primary tumour area measured after resection. Data are mean ± s.d. $n = 20$ mice (control) and $n = 16$ (DT). *P* values were calculated using two-sided Wilcoxon rank-sum tests. **r**, Liver metastases at the experimental end points. Data are mean ± s.d. Each dot represents a mouse. *n* values are as described in **r**. **s**, The percentage of mice that relapse with liver metastases. *P* values were calculated using two-sided Fisher's tests. **t**, Experimental timeline and liver metastasis growth monitored by BLI after intrasplenic inoculation of MTOs. **u**, The number of liver metastases in **t**. Data are mean ± s.d. For **t** and **u**, $n = 7$ (control) and $n = 8$ (DT treated) mice. The points and lines of BLI measurements in **i**, **k** and **t** represent individual mice, the trend lines (bold) show a LOESS model and *P* values were calculated using a mixed-effects linear model. For **g**, **r**, **l** and **u**, *P* values were calculated using a generalized linear model.

T cell exclusion[12]. Transplantation of AKTP MTOs in the caecum also gave rise to T-cell-excluded CRCs (Extended Data Fig. 12a). CAFs surrounded *Emp1*-expressing invasion fronts and tumour buds, but T cells did not reach these structures and remained at the tumour periphery (Extended Data Fig. 12b). By contrast, liver micrometastases generated by these primary CRCs exhibited high CD3$^+$ cell density (Fig. 5a). The

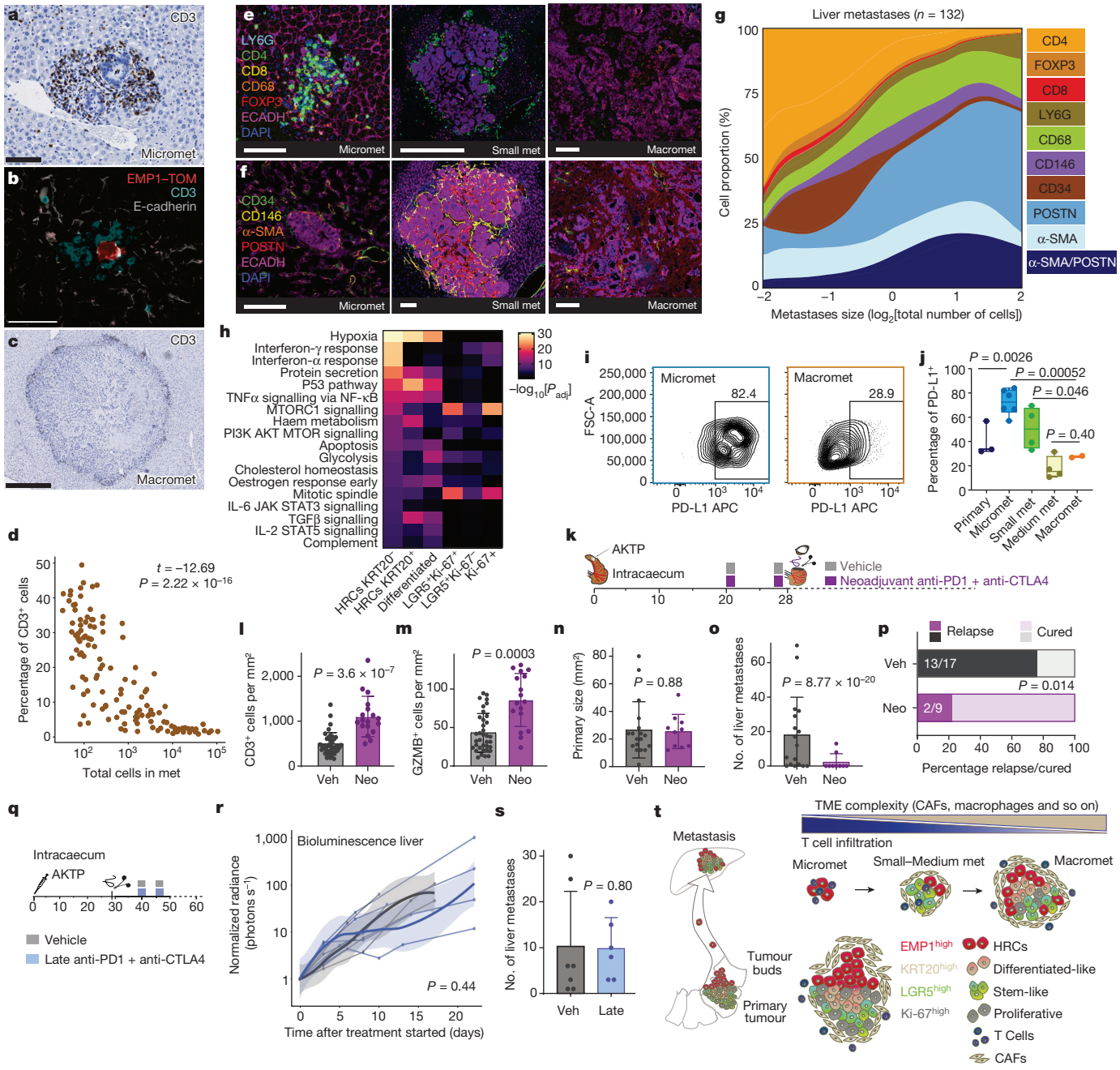


Fig. 5 | Neoadjuvant immunotherapy prevents metastatic relapse in CRC. **a**, Representative images of CD3 immunohistochemistry on a liver micrometastasis. Scale bars, 100 µm. **b**, tdTomato, CD3 and E-cadherin immunofluorescence showing Emp1-tdTomato$^+$ liver micrometastatic cells surrounded by CD3$^+$ cells. E-cadherin fluorescence is used for cell delimitation. Scale bar, 50 µm. **c**, Representative images of CD3 immunohistochemistry on a liver macrometastasis. Scale bar, 500 µm. **d**, The percentage of CD3$^+$ T cells (of total TME cells) versus metastasis size. $n = 12$ mice, 133 metastases. Statistical analysis was performed using a linear model with mouse as a fixed effect. **e**,**f**, Representative examples of multiplex immunofluorescence of immune (**e**) and stromal (**f**) markers in metastases. Scale bars, 100 µm (micro and macro), 500 µm (small, **e**) and 100 µm (small, **f**). **g**, Proportional stacked area graph showing TME cell types in metastases of different sizes; $n = 132$ metastases of 20 mice. **h**, Heat map showing Hallmark GSEA in the cell populations described in Fig. 2g. **i**, Representative flow cytometry contour plot showing PD-L1 expression in tumour cells of micro- and macrometastases. **j**, The percentage of PD-L1$^+$ tumour cells determined using flow cytometry. The median (centre line), first and third quartiles (box limits) and the smallest and largest values (whiskers) are shown. $n = 3$ (primary tumour), $n = 6$ (micrometastases), $n = 4$ (small and medium metastases) and $n = 2$ (macrometastases). Statistical analysis was performed using Wilcoxon rank-sum test. **k**, Neoadjuvant immunotherapy treatment. **l**,**m**, CD3$^+$ (**l**) and GZMB$^+$ (**m**) cell densities in primary CRCs. Data are mean ± s.d. $n = 19$ (control) and $n = 10$ (neoadjuvant (neo)). Statistical analysis was performed using a mixed-effects linear model. Veh, vehicle. **n**, Primary CRC area after resection. Data are mean ± s.d. Each dot represents a mouse. $n = 18$ (control) and $n = 10$ (treated). Statistical analysis was performed using a linear model. **o**, Liver metastases at the experimental end points. Data are mean ± s.d. Each dot represents a mouse. $n = 17$ (control) and $n = 9$ (treated). The P value was calculated using a generalized linear model. **p**, The percentage of mice that developed liver metastases at the experimental end points. n values are as described in **o**. The P value was calculated using a generalized linear model. **q**, Late immunotherapy. **r**, Liver metastasis growth measured by normalized BLI. The points and lines represent individual mice, and trend lines show a LOESS model. $n = 4$ (control) and $n = 5$ (late immunotherapy). Statistical analysis was performed using a linear model. **s**, Liver metastases generated by AKTP primary tumours. Data are mean ± s.d. Dots represent individual mice. $n = 7$ (control) and $n = 6$ (late immunotherapy). The P value was calculated using a generalized linear model. **t**, The proposed model for metastatic dissemination of CRC and TME co-evolution.

lack of T cell exclusion was evident and we could visualize T cell–HRC interactions in these early lesions (Fig. 5b). However, metastatic outgrowth was accompanied by a gradual decrease in $CD3^+$ cells and the relocalization of T cells to the periphery (Fig. 5c,d). Multiplex immunohistochemistry analysis showed that most infiltrating T cells were $CD4^+FOXP3^-$ (examples are shown in Fig. 5e and quantification is shown in Fig. 5g and Extended Data Fig. 12c). T cell exclusion coincided with progressive CAF (α-SMA^+ and/or $POSTN^+$) and macrophage ($CD68^+$) recruitment to the metastatic TME (Fig. 5f,g and Extended Data Fig. 12c).

scRNA-seq revealed elevated expression of interferon-alpha- and interferon-gamma-target genes in undifferentiated ($KRT20^-$) HRCs (Fig. 5h and Supplementary Table 4), >95% of which belong to micrometastatic lesions, suggestive of an ongoing inflammatory response as HRCs reach the liver. Indeed, undifferentiated HRCs upregulate multiple interferon-response genes compared with the other tumour cell populations in primary CRCs and large metastases (Extended Data Fig. 12d). It was also apparent that HRCs in micrometastases expressed elevated *Cd274* (PD-L1) and *Ido1* levels, two negative regulators of the immune response (Extended Data Fig. 12d). Flow cytometry analysis confirmed that there was elevated PD-L1 expression on the surface of micrometastatic CRC cells and a progressive decline over subsequent outgrowth (Fig. 5i,j). On the basis of these findings, we hypothesized that, at the onset of organ colonization, the lack of a mature TME exposes HRCs to the adaptive immune system, yet disseminated cells bypass immune attack through the expression of immunomodulatory molecules. To test the susceptibility of micrometastatic disease to immunotherapy, we treated mice with anti-PD1 antibodies in combination with anti-CTLA4 antibodies in the neoadjuvant setting, that is, before the primary tumour was extirpated by surgery (Fig. 5k–p). This approach increased $CD8^+$ cytotoxic T cell numbers in the primary CRC (Fig. 5l,m) but we did not observe an alteration in the tumour growth rate (Extended Data Fig. 12e) or curative effects on the primary disease at the experimental end points (Fig. 5n). Notably, however, most of these mice did not develop metastasis after surgical removal of the primary CRC and remained disease free (Fig. 5o,p). We obtained similar results with neoadjuvant anti-PD1 monotherapy (Extended Data Fig. 12f–i). By contrast, the same regime applied 2 weeks after surgery did not stop metastatic outgrowth (Fig. 5q–s), which is consistent with the failure of checkpoint immunotherapy observed in patients with metastatic microsatellite-stable CRC[35]. Thus, during a temporal window after metastatic colonization, immunotherapy is effective in eliminating the residual disease and preventing subsequent metastatic relapse.

Discussion

HRCs represent a defined cell state in a large proportion of patient samples, implying a common origin and mechanism of metastatic recurrence across CRC genotypes and molecular subtypes. A pioneering study revealed that $LGR5^+$ cancer stem cells are dispensable for primary CRC growth yet necessary for metastasis formation in experimental models[3]. Subsequently, it was proposed that metastases are initiated by disseminated differentiated tumour cells that, through plasticity, produce $LGR5^+$ cancer stem cells upon reaching the liver[36]. Another study showed that the adhesion molecule L1CAM in $LGR5^+$ and $LGR5^-$ cells is important to regenerate metastasis[37]. We unequivocally show that HRCs are neither differentiated nor stem-like but, rather, co-opt a distinct state that enables migration and colonization of foreign organs. A defining feature of the HRC state is the upregulation of genes encoding cell-to-cell adhesion molecules. We therefore speculate that HRCs may extravasate as cell clusters and colonize foreign organs as oligocellular structures as previously shown for other cancer types[22,38]. Reacquisition of the $LGR5^+$ stem cell and proliferation programs occurs at a later phase and is necessary for metastatic outgrowth (Fig. 5t). Although ablation of HRCs in primary tumours prevents the vast majority of metastatic relapses, it is formally possible that other tumour cell populations, including $LGR5^+$ cells, may generate metastases if they reach foreign organs, as suggested by experiments of direct inoculation of tumour cells into the blood stream performed herein and elsewhere[3,28,36,37].

Our findings also reveal that residual metastatic cells lodged in foreign organs lack a mature TME and are susceptible to the attack of the adaptive immune system after immunotherapy treatment. This window of vulnerability could be exploited to prevent metastatic relapse. Our results support current efforts to use neoadjuvant immunotherapy in patients with early-stage CRC[39]. This and other therapeutic strategies with the ability to eliminate HRCs may prevent disease relapse if applied before metastatic disease is overt.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05402-9.

1. Amin, M.B. et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA* **67**, 93–99 (2017).
2. Shimokawa, M. et al. Visualization and targeting of $LGR5^+$ human colon cancer stem cells. *Nature* **545**, 187–192 (2017).
3. de Sousa e Melo, F. et al. A distinct role for $Lgr5^+$ stem cells in primary and metastatic colon cancer. *Nature* **543**, 676–680 (2017).
4. Cortina, C. et al. A genome editing approach to study cancer stem cells in human tumors. *EMBO Mol. Med.* **9**, 869–879 (2017).
5. Calon, A. et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer Cell* **22**, 571–584 (2012).
6. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
7. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
8. Lee, H.-O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
9. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
10. Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137 (2021).
11. Joanito, I. et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat. Genet.* **54**, 963–975 (2022).
12. Tauriello, D. V. F. et al. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
13. Massagué, J. & Obenauf, A. C. Metastatic colonization by circulating tumour cells. *Nature* **529**, 298–306 (2016).
14. Barriga, F. M. et al. Mex3a marks a slowly dividing subpopulation of $Lgr5^+$ intestinal stem cells. *Cell Stem Cell* **20**, 801–816 (2017).
15. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
16. Álvarez-Varela, A. et al. Mex3a marks drug-tolerant persister colorectal cancer cells that mediate relapse after chemotherapy. *Nat. Cancer* **3**, 1052–1070 (2022).
17. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat. Commun.* **12**, 2592 (2021).
18. Grigore, A. D., Jolly, M. K., Jia, D., Farach-Carson, M. C. & Levine, H. Tumor budding: the name is EMT. Partial EMT. *J. Clin. Med.* **5**, 51 (2016).
19. Roa-Peña, L. et al. Keratin 17 identifies the most lethal molecular subtype of pancreatic cancer. *Sci. Rep.* **9**, 11239 (2019).
20. Durgan, J. et al. SOS1 and Ras regulate epithelial tight junction formation in the human airway through EMP1. *EMBO Rep.* **16**, 87–96 (2015).
21. Bangsow, T. et al. The epithelial membrane protein 1 is a novel tight junction protein of the blood-brain barrier. *J. Cereb. Blood Flow Metab.* **28**, 1249–1260 (2008).
22. Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110–1122 (2014).
23. Barry, E. R. et al. Restriction of intestinal stem cell expansion and the regenerative response by YAP. *Nature* **493**, 106–110 (2013).
24. Cheung, P. et al. Regenerative reprogramming of the intestinal stem cell state via hippo signaling suppresses metastatic colorectal cancer. *Cell Stem Cell* **27**, 590–604 (2020).
25. Vasquez, E. G. et al. Dynamic and adaptive cancer stem cell population admixture in colorectal neoplasia. *Cell Stem Cell* **29**, 1213–1228 (2022).
26. Han, T. et al. Lineage reversion drives WNT independence in intestinal cancer. *Cancer Discov.* **10**, 1590–1609 (2020).
27. Lupo, B. et al. Colorectal cancer residual disease at maximal response to EGFR blockade displays a druggable Paneth cell-like phenotype. *Sci. Transl. Med.* **12**, eaax8313 (2020).
28. Heinz, M. C. et al. Liver colonization by colorectal cancer metastases requires YAP-controlled plasticity at the micrometastatic stage. *Cancer Res.* **82**, 1953–1968 (2022).

29. Solé, L. et al. p53 wild-type colorectal cancer cells that express a fetal gene signature are associated with metastasis and poor prognosis. *Nat. Commun.* **13**, 2866 (2022).
30. Ohta, Y. et al. Cell-matrix interface regulates dormancy in human colon cancer stem cells. *Nature* **680**, 784–794 (2022).
31. Mustata, R. C. et al. Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep.* **5**, 421–432 (2013).
32. Wang, Y. et al. Comprehensive molecular characterization of the hippo signaling pathway in cancer. *Cell Rep.* **25**, 1304–1317 (2018).
33. Yuan, Y. et al. YAP1/TAZ-TEAD transcriptional networks maintain skin homeostasis by regulating cell proliferation and limiting KLF4 activity. *Nat. Commun.* **11**, 1472 (2020).
34. Morral, C. et al. Zonation of ribosomal DNA transcription defines a stem cell hierarchy in colorectal cancer. *Cell Stem Cell* **26**, 845–861 (2020).
35. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
36. Fumagalli, A. et al. Plasticity of Lgr5-negative cancer cells drives metastasis in colorectal cancer. *Cell Stem Cell* **26**, 569–578 (2020).
37. Ganesh, K. et al. L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat. Cancer* **1**, 28–45 (2020).
38. Padmanaban, V. et al. E-cadherin is required for metastasis in multiple models of breast cancer. *Nature* **573**, 439–444 (2019).
39. Chalabi, M. et al. Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. *Nat. Med.* **26**, 566–576 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

MTOs and integration of cassettes using CRISPR

We previously described[12] the establishment of MTOs from primary tumours arising in genetically engineered mouse models with compound genetic alterations (*Apc*, *Kras*, *Trp53*, *Tgfbr2*). MTOs were cultured as described previously[12] and checked bimonthly for mycoplasma contamination. Small guide RNAs (sgRNA; 20 bp) were designed to cut 9–11 bp after the STOP codons using the http://crispr.mit.edu web tool and were cloned into the pX330-IRFP hSp-enhanced-Cas9 plasmid[4]. The sgRNA sequence for *Emp1* was AAATAAGCCGAATACGCTCA; and for *Lgr5* the sequence was GTCTCTAGTGACTATGAGAG. Then, 1 kb 5′ and 3′ homology arms were sequentially cloned into pShuttle vectors containing the inducible Caspase9-tdTomato cassette[34], a gift from T. Sato. The Ires-DTR-T2A-eGFP-WPRE-BGHpA sequence was cloned between the *Lgr5* homology arms flanking the gene stop codon. *eGFP* was cloned after the *Ires* sequence to generate the Lgr5-Ires-eGFP-WPRE-BGHpA donor. CRISPR–Cas9 knockin editing was performed as described previously[4]. AKTP-MTO-93 *Emp1-iCT* was generated from a single tdTomato$^+$ cell (clone 14). MTO-93 *Emp1-iCT Lgr5-eGFP* was first generated from a single tdTomato$^+$ cell (clone 49), then nucleofected with the *Lgr5-eGFP* construct and established from a pool of sorted eGFP$^+$ cells. MTO-93 *Lgr5-iCT* was generated by a single tdTomato$^+$ cell (clone 2). MTO-93 *Lgr5-DTR-eGFP* was generated by sorting a pool of eGFP$^+$ cells. Triple-mutant AKP MTO-54[12] *Emp1-iCT* was generated by sorting a pool of tdTomato$^+$ cells. CRISPR–Cas9 gene editing was subsequently used to introduce a *Smad4* mutation in AKP *Emp1-iCT* to generate AKPS MTOs as previously described[40]. Correct integration of the knockin cassettes was checked by PCR and sequencing of the genomic regions. Correct expression of the reporter cassette was tested by RT–qPCR or by sorting cells from tumours as described in the text. A list of genotyping primers is provided in Supplementary Table 5.

Generation of CTOs

To study the effect of gene driver mutations on the HRC phenotype, we sequentially introduced mutations into *Apc*, *Kras*, *Trp53*, *Tgfbr2* and/or *Smad4* in normal mucosa colonic organoids using CRISPR–Cas9 as described elsewhere[16,40,41]. Compound mutations were introduced into organoids derived from one c57BL/6J mouse, which enabled us to assess the effect of individual mutations on gene expression without other confounding variables. We named these organoids CTOs to distinguish them from MTOs, that were established from genetically engineered mouse models.

Lentivirus production and MTO infection

For bioluminescence tracking, MTOs were infected with a lentivirus encoding an eGFP–firefly luciferase fusion reporter construct under the control of the *Pgk1* promoter. For YAP inhibition experiments, AKTP organoids were infected with shControl (SHC002) or shYAP1 (TRCN0000095864, TRCN0000095865, TRCN0000095866, TRCN0000095867) Mission Sigma-Aldrich lentiviral constructs. We also used a pInducer20 eGFP-TEADi vector, a gift from R. Iglesias-Bartolome (Addgene plasmid, 140145), that blocks the activity of both YAP and TAZ[33].

Animal experimentation approval and maintenance

Experiments with mice were approved by the Animal Care and Use Committee of Barcelona Science Park under protocol CEEA-PCB-14-000053. Mice were maintained in a specific-pathogen-free (SPF) facility under a 12 h–12 h light–dark cycle, under controlled temperature and humidity (18–23 °C and 40–60%, respectively) and given ad libitum access to standard diet and water. All of the mice were closely monitored by the authors, facility technicians and by an external veterinary scientist responsible for animal welfare. We monitored primary tumour and metastasis growth using intravital bioluminescence at least once a week.

Inoculation of MTOs into the caecum and rectum

For all injections, c57BL/6J mice were purchased from Janvier Labs at six weeks of age and injected at 7 to 9 weeks of age. Sex always matched the origin of the tumour. Intracaecum injections were used to generate primary tumours. Organoids were collected and incubated for 30 min with cold HBSS to break down BME (Bio-Techne, 3533-010-02) without disrupting their structure. Cells were then counted and resuspended in 70% BME in HBSS for injection (0.1×10^6 cells in 10 µl per mouse). Full organoids were injected with a 30G syringe into the submucosal wall of the distal caecum while looking through a binocular lens. We introduced a significant modification to previous protocols[42] by moving the injection site to the apex of the caecum, which enabled posterior surgical resection. For liver colonization studies, we used intrasplenic injections of organoids as described before[12]. Rectal injections were performed according to a previously described procedure[43]. Primary CRC and metastasis growth were monitored using bioluminescence. Experiments were terminated when tumour bioluminescence measurements reached saturation or if the animals presented signs of suffering or poor health status according to protocols approved by the animal experimentation committee of the Barcelona Science Park.

Primary tumour resection

Mice were anesthetized with isoflurane and placed into dorsal recumbency. The abdomen was shaved and sterilized with povidone-iodine surgical solution. A small midline incision—slightly to the left—was performed to open the skin and peritoneum and expose the abdominal area. We placed a sterile surgical drape on top of the abdomen with a circular hole above the incision and sprawled the caecum over the drape using cotton swabs and saline to keep it hydrated. After confirming the presence of a primary tumour, Kelly forceps were used to first knot the surgical suture into the caecum wall, in between the ileocecal junction and the primary tumour. This provided a grip for subsequent caecum ligation. After ligation, the apical caecum containing the primary tumour was excised and any remaining caecal tissue was trimmed. After resection and organ fixation, we measured primary tumour size using callipers. A video of the surgery, which usually lasted 5–10 min, is provided (Supplementary video 1). The fitness of the mice was monitored weekly throughout the experiment. Mice were euthanized four weeks after resection and metastases were scored macroscopically.

Pharmacological treatments

For iCaspase-9-inducible ablation experiments, animals were treated with dimerizer (AP20187, MedChem Express, HY-13992) through intraperitoneal injection at 2.5 mg per kg three times per week (*Emp1-iCasp9*) or 5 mg per kg every day (*Lgr5-iCasp9*). For DTR-inducible ablation, mice were treated with 16.7 µg per kg of DT (Sigma-Aldrich, 322326) three times per week. For immunotherapy experiments, 2 shots of 250 µg of anti-mouse CTLA-4 (C2444, Leinco) and/or anti-mouse PD1 (P372, Leinco) were administered through intraperitoneal injection 5 days apart.

Tumour dissociation for flow cytometry

Primary tumours and microdissected liver metastases were chopped with razor blades. Subsequent enzymatic digestion was performed with 200 U ml^{-1} collagenase IV (Sigma Aldrich, C5138) in HBSS (Lenovo) for 30 min at 37 °C, in a shaking water bath or a gentleMACS Dissociator (Miltenyi Biotec). Digested tissue fragments were then filtered through 100 µm and 40 µm meshes, washed and treated for 5 min with ammonium chloride. Single-cell preparations were first blocked with anti-CD16/32 (93, eBioscience) and then stained with APC anti-EPCAM (BioLegend, 118214) and BV605 anti-CD45 (BioLegend, 103155) antibodies at 1:200 concentration. In experiments measuring PD-L1 expression in tumour cells, cells were stained with FITC anti-EPCAM (Santa Cruz,

G8.8, 53532) and APC anti-PD-L1 (BD, 564715, MIH5) antibodies at 1:200 concentration. Finally, cells were resuspended in HBSS with 0.5% FBS and DAPI (Sigma Aldrich, D9542).

Collection of mouse primary CRC and metastasis samples

Metastatic seeding in the CRC relapse models is not synchronized; as it occurs in patients, tumour cells are shed from the primary tumour continuously during the weeks until the day of resection. As a result, metastases expand for different periods in the liver and lungs, and exhibit different sizes at the experimental time points depending on when HRCs colonized the organ. We therefore decided to classify according to size instead to time. Four different CRC samples were collected—primary, micrometastases, small metastases and macrometastases. Micrometastases and small metastases were collected from liver of mice 29 to 31 days after primary tumour implantations (that is, at the time of primary CRC resection). Micrometastasis samples were DTCs collected from livers with absent or residual bioluminescence ex vivo in which metastases were not visible. For small metastases, metastatic nodules were visible but small in size (<1.5 mm). Macrometastases samples were derived from metastatic nodules larger than 4 mm in livers from mice 64 to 66 days after primary tumour implantation. Primary tumour samples were paired with micro, small or macrometastasis samples and were collected from both time points (29 to 31 and 64 to 66 days after primary tumour implantation).

Isolation of liver residual DTCs

For isolation of low numbers of tumour cells from mice without visible metastases, whole livers were thoroughly minced with razor blades. After an initial 30 min digestion with 200 U ml^{-1} collagenase IV (Sigma-Aldrich, C5138), samples were filtered through 100 µm meshes. Although most cells flowed through, a small fraction of the sample—highly enriched for tumour cells—was retained in the filter (Extended Data Fig. 3k,l). The filters were then laid into a six-well plate and covered with HBSS containing 200 U ml^{-1} collagenase IV, 200 µg ml^{-1} dispase II (Sigma-Aldrich, D4693), 40 µg ml^{-1} DNase I (Sigma-Aldrich, 10104159001) and 13 µM rock inhibitor (MedChem Express, HY-10583). After redigestion in a water bath for 30 min al 37 °C, most cells now seeped through the filters. The protocol continued with washing, ammonium chloride treatment and antibody staining as described above. $DAPI^{-}eGFP^{+}CD45^{-}$ cells were gated to sort tumour cells.

Histology and tissue staining

Standard haematoxylin and eosin and antibody staining were performed on 4 µm tissue sections using standard procedures, as described previously[12]. Details are provided in Supplementary Table 6. The following secondary antibodies were used: donkey anti-goat conjugated to Alexa 488/568/647 (Life Technologies A11055, A11057, A21447), donkey anti-rabbit conjugated to Alexa 488/568/647 (Life Technologies A21206, A10042, A31573) and donkey anti-mouse conjugated to Alexa 488/568/647 (Life Technologies A-21202, A10037, A31571) at room temperature. Digital scanned bright-field and fluorescent images were acquired with a NanoZoomer-2.0 HT C9600 scanner (Hamamatsu Photonics). All of the images were visualized with the NDP.view 2 U123888-01 software (Hamamatsu Photonics) with a gamma correction set at 1.8 in the image control panel.

Co-culture of MTOs and colon fibroblasts

To address the role of fibroblasts in inducing HRCs, we seeded 5,000 AKTP *Emp1-iCT* MTO cells with or without 100,000 mouse colon fibroblasts in BME (Bio-Techne, 3533-010-02) and supplemented them with MTO cell culture medium[12] without noggin and galunisertib. Co-cultures were analysed 7 days after by flow cytometry and confocal microscopy. For Fig. 3s and Extended Data Fig. 10e, MTOs and mouse colon fibroblasts were also co-cultured in hanging drops as previously described[44].

In situ hybridization analysis of samples of patients with CRC

Paraffin-embedded tissue sections (thickness, 2–3 µm) of human CRC primary tumours were air-dried and further dried at 60 °C over-night before any staining. To compare *EMP1* and *LGR5* expression, sections were hybridized with RNAscope Probe Hs-LGR5 (311021, Bio-Techne, R&D Systems) in the C1 channel, a custom-made RNAscope Probe Hs-EMP1 (Bio-Techne, R&D Systems, 895051-C2) in the C2 channel and an Alexa568-conjugated antibody against E-cadherin. To compare the expression of *EMP1* with other HRC-marker genes, sections were hybridized using the RNAscope Probe Hs-EMP1 mRNA in channel 1 (Bio-Techne, R&D Systems, 895051-C1) and stained with the relevant antibodies. FISH probes were detected using the RNAscope Multiplex Fluorescent Detection Reagents Kit v2 (Bio-Techne R&D Systems, 323110).

Multiplex immunofluorescence

Multiplex immunofluorescence staining was performed on 4-µm-thick formalin-fixed paraffin-embedded sections using the OPAL protocol (Akoya Biosciences) on the Leica BOND RXm autostainer (Leica Microsystems). Six consecutive staining cycles were performed using the following primary antibody–Opal fluorophore pairings. Stroma panel: CD34 (1:3,000, ab81289; Abcam) with Opal 520; CD146 (1:500, ab75769; Abcam) with Opal 570; α-SMA (1:1,000, ab5694; Abcam) with Opal 620; periostin (1:1,000; Abcam, ab227049) with Opal 690; and E-E-cadherin (1:500; Cell Signaling, 3195) with Opal 650. Immune panel: (1) LY6G (1:300; BD Pharmingen, 551459) with Opal 540; (2) CD4 (1:500; Abcam, ab183685) with Opal 520; (3) CD8 (1:800; Cell Signaling, 98941) with Opal 570; (4) CD68 (1:1,200, Abcam, ab125212) with Opal 620; (5) FOXP3 (1:400; Cell Signaling, 126553) with Opal 650; and (6) E-cadherin (1:500; Cell Signaling, 3195) with Opal 690. Primary antibodies were incubated for 60 min and detected using the BOND Polymer Refine Detection System (Leica Biosystems, DS9800) according to the manufacturer's instructions, substituting DAB for the Opal fluorophores, with a 10 min incubation time and without adding haematoxylin. Whole-slide scans and multispectral images (MSI) were obtained on the Akoya Biosciences Vectra Polaris system. Batch analysis of the MSIs from each case was performed using the inForm (v.2.4.8) software provided. Finally, batch-analysed MSIs were fused in HALO (Indica Labs) to produce a spectrally unmixed reconstructed whole-tissue image, ready for analysis.

Statistical analyses multiplex immunofluorescence

The Halo Image Analysis Platform software was used for quantification of cell phenotypes within metastases and primary tumours. Two matrices of counts (distinguishing immune and stromal panels) with the total number of cells per metastasis/primary (in rows) assigned to each cell population (in columns) were obtained. Multiple positives were present in both immune and stromal panels. These cases observed in the immune panel represented only 0.7% of the total assignations and were removed from the analysis. For the stromal panel, α-SMA and POSTN double positives were retained and labelled in a distinct category α-SMA/POSTN. The rest of the multiple positives, comprising 7% of the total cells, were amalgamated and labelled as stromal others. For all metastases, information regarding metastasis size (log_2-transformed total number of cells), metastatic burden (defining micro, small and macrometastases) and organ site were considered for subsequent analysis.

For the proportional stacked area graph, only the percentages of CD4, FOXP3, CD8, LY6G and CD68 for the immune panel, and the percentages of CD146, CD34, POSTN, α-SMA and α-SMA/POSTN for the stromal panel were considered. The cell composition of all measured metastases (in percentages) was averaged out using the R function aggregate, taking as grouping elements both metastasis size and metastatic burden. Linear mixed-effects models were fitted independently for every cell population using the CLR-transformed values as the response variable,

the metastatic burden and the metastasis size as fixed effects, and the tumour ID as a random effect to consider the dependence between metastases for the same tumour.

Quantification of tdTomato and eGFP fluorescence intensity

For the quantification of percentages of EMP1–tdTomato high and LGR5–eGFP^{high} in tumour sections, we used the Halo Image Analysis Platform. In brief, the epithelial tumour area was classified apart from the stroma, background and necrosis using a random-forest algorithm. Single cells were detected and tdTomato/eGFP fluorescence intensity was measured for every cell. In primary tumours, TOM^{high} and eGFP^{high} tumour cells were defined as the cells in the 90th percentile of each sample. In liver metastases, TOM^{high} and eGFP^{high} tumours cells were defined as the cells in the 90th percentile for all metastases measured. For *Emp1-iCT* liver metastases (Extended Data Fig. 6j) and *Lgr5-DTR-eGFP* liver metastases (Extended Data Fig. 11q), tumour cell fluorescence intensity was analysed using ImageJ with a custom-made macro. Tumour cells were first detected and isolated using E-cadherin to create a mask. TdTomato or eGFP intensity was calculated for every pixel inside the masked area. We then plotted the percentage of fluorescence high and low pixels as a function of the area of the metastases (measured in pixels).

Gene expression analysis using RT–qPCR

RT–qPCR and microarrays were used to compare subpopulations of EMP1^{high} and EMP1^{low}, and LGR5^{high} and LGR5^{low} cells dissociated from MTO-93 organoids grown in vitro or in vivo. Subpopulations were defined in flow cytometry as the top/bottom 10% in fluorescence expression. RNA from 2,000 cells was extracted and retrotranscribed to cDNA as described previously[45]. To analyse changes in gene expression, RT–qPCR was performed using 5 ng of cDNA per each qPCR well. qPCRs were performed using the TaqMan Universal PCR Master Mix (Applied Biosystems, 4369016) or the PowerUp SYBR Green Master Mix (Applied Biosystems, 100029284) in triplicates, according to the manufacturer's instructions. Gene expression levels were normalized using the housekeeping genes *PPIA* or *B2M*. The following TaqMan assays were used: *TdTomato-BGHPA* (custom-made probe; F, GGGCATGGCACCGGCAGCACC; R, CCTACTTGTACAGCTCGTCCATGCC), Mm*Ppia* (Mm002342430_g1), Mm*B2m* (Mm00437762_m1), *eGFP* (Mr04097229mr), Mm*Emp1* (Mm00515678_m1), Mm*Lgr5* (Mm0043889_m1), Mm*Smoc2* (Mm00491553_m1), Mm*LamC2* (Mm00500494), *Ctgf* (Mm01192033_g1). The following Sybr primers were used: MmYap (F, ACCCTCGTTTTGCCATGAAC; R, TGTGCTGGGATTGATATTCCGTA), Mm*Cyr61* (F, AGGTCTGCGCTAAACAACTCA; R, ATATTCACAGGGTCTGCCTTCT).

Western blotting

Cells were collected in Cell Recovery Solution (Corning, 354257), pelleted and lysed in RIPA buffer (Tris-HCl pH 7.4, 150 mM NaCl, 0.5% Sodium deoxycolate, 0.1% SDS, 1 mM EDTA, 10 mM NaF, 1 mM PMSF, 1% Triton X-100) supplemented with protease and phosphatase inhibitors (Sigma-Aldrich). Primary antibodies were incubated overnight at 4 °C at the following dilutions: YAP/TAZ (Cell Signalling, CS93622, 1:1,000) and vinculin (Sigma Aldrich, 9264, 1:20,000). Anti-Rabbit IgG HRP (NA934V) and anti-mouse IgG HRP (NXA931V) secondary antibodies were diluted 1:5,000 and incubated for 1 h at room temperature with the PVDF membranes. Membranes were visualized using Hyper Processor (Amersham Pharmacia Biotech).

RNA-seq and analysis

We used RNA-seq to profile CTOs with different genotypes and to profile the effects of chemotherapy treatment on MTOs. For the first dataset, 5,000 cells from CTOs with multiple genotypes (AT, AP, AS, AK, ATP, APS, AKP, AKS, AKPT, AKPS) were seeded in BME (Bio-Techne, 3533-010-02) and cultured in full stem cell medium[4]. Organoids were collected 7–10 days after seeding. For chemotherapy experiments, MTO54 organoids were dissociated with TryplE express (Thermo Fisher Scientific, 12604039) and single cells were resuspended in BME in six-well plates in complete stem cell medium. Then, 2 days later, the medium was removed and fresh medium containing folfiri (5FU at 50 μg μl^{-1} (Sigma-Aldrich, F6627) plus SN38 at 100 nM (MedChem, HY-13704)) or control (MTO stem cell medium[12]) was added for 2 days.

RNA was extracted and sequenced as we previously described[16]. RNA-seq reads from datasets (CTOs or chemotherapy treatment) were aligned with STAR (v.2.5.2)[46] using the default parameters to the *Mus musculus* reference genome built with annotations version GENCODE_mmusculus_vM25. SAM files were converted to BAM and sorted using Sambamba (v.0.7.1)[47]. Count matrices were generated using the R (v.4.0.5) package Rsubread (v.2.4.3)[48] with the GENCODE_mmusculus_vM25 custom annotation. Data from parental versus metastatic MTOs were processed as described previously[12]. Gene-wise differential expression in the chemotherapy dataset between controls and folfiri treatment was performed using the R package DESeq2 (v.1.30.1)[49]. Normalized values for plots were obtained using the rlog function of the same package. Signature scores were defined as the scaled mean of all genes in the signature after scaling the expression matrix. Comparison of signature scores between conditions was assessed using *t*-tests.

Microarray expression analyses

Microarrays (GeneChip Mapping 250K Nsp Assay Kit, Affymetrix) were used to compare EMP1^{high} and EMP1^{low} cells from MTO-93 *Emp1-iCT* primary tumours grown for 4 weeks; and to compare MTO-93 *Emp1-iCT* MTOs grown in vitro alone or in combination with mouse colon fibroblasts for 7 days. Samples were processed with oligo (v.1.46.0)[50] (fitProbeLevelModel: background = TRUE, normalize = TRUE, target = "core", method = "plm"). Raw cel files were normalized using the RMA method[51] (with the default parameters). Probe sets were annotated with Clariom_S_Mouse_HT-na36-mm10-transcript Affymetrix databases. Standard quality controls were considered to identify abnormal samples[52]. No samples were excluded due to quality issues.

Differential expression analysis of EMP1^{high} versus EMP1^{low} data was performed using a linear model with empirical shrinkage (limma R package)[53], taking into account the paired data setting. Differential expression analysis of co-culture data was performed using the same regression method, but considering as adjustment variable the Eklund metric pm.iqr[54] to reduce the influence of technical variation. Benjamini–Hochberg false-discovery rate (FDR) correction was used for multiple-comparison correction. Gene set analysis was used to explore the enrichment in custom gene sets. limma's rotation-based approach for enrichment[55] was considered to represent the null distribution. The maxmean enrichment statistic proposed in ref. [56], under restandardization, was used for competitive testing. Gene signature (Supplementary Table 7) *z*-scores[57] were used to measure pathway activity. For doing so, normalized expression was adjusted for biological replicate, centred and scaled gene-wise according to the mean and the s.d. computed across samples. Gene signature *z*-score was summarized by taking the average of its constituent genes. Furthermore, a global signature was computed using all genes and used for a priori centring of signature scores. This strategy has been proved to be useful to avoid systematic biases. Only the expression of the most variable probe sets per gene were considered for gene set analyses.

10x mouse scRNA-seq analysis

CellRanger[58] (v.4.0.0) was used to align reads to a custom refdata-gex-mm10-2020-A transcriptome including the *eGFP* and luciferase genes. Gene expression was analysed using Seurat (v.4.0.3)[59–62]. A total of 1,330 cells with <20% mitochondrial content and >3,000 detected genes were considered. Ribosomal reads (17% of the total) were removed. Mitochondrial content was regressed out during SCT normalization[63].

SCT-transformed counts were smoothed using MAGIC (v.2.0.3)[64]. Gene signature expression was summarized by taking the average MAGIC score of its constituent genes. HRCs and LGR5$^+$ cell populations were defined by having a score above the 75th percentile. FindAllMarkers was used to identify differentially expressed genes from raw counts. The HRC cell population was compared to the rest to identify mouse HRC markers. Testing was limited to genes detected in >10% of cells and showing >0.25 log-transformed fold difference. Functional enrichment analysis was performed using the GSEApreranked[65] method ranking genes by the $\log_2$-transformed average fold change. The significance threshold was set at 5% Benjamini–Hochberg FDR.

To select for marker genes to track HRCs in mouse tumours, we computed the correlation scores for all genes with the EpiHR signature in the SMC human dataset and the mouse 10x dataset. We used MAGIC score for genes and signature scores as defined previously.

Mouse Smart-seq2 scRNA-seq analysis

Smart-seq2 reads were aligned to the UCSC_GRCm38.mm10 genome with zUMIs[66] and analysed using Seurat[59–62] (v.4.0.3). Four technical batches of AKTP at different stages were merged into a single object and two technical batches of AKP micrometastases were merged into another object. A total of 1,057 cells with <20% mitochondrial content and >20,000 UMIs were considered for AKTP, whereas 414 cells with <20% mitochondrial content and >100,000 UMIs were isolated from AKP micrometastases. Ribosomal reads were removed. Mitochondrial content was regressed out during SCT normalization. SCT-transformed counts were further imputed and smoothed using MAGIC (v.2.0.3)[64]. Gene signature expression (Supplementary Table 7) was summarized by taking the average MAGIC score of its constituent genes. Non-epithelial cells were removed and normalized again. To improve the integration of the four AKTP batches, the IntegrateData function was used with precomputed anchors based on 3,000 features. The integrated dataset was SCT normalized, smoothed with MAGIC (v.2.0.3)[64] and clustered with the FindClusters Seurat function (resolution = 1.2). FindMarkers was used to identify differentially expressed genes from raw counts. CellRank[15] was used to uncover the cell-state dynamics of CRC metastases from RNA velocity estimates[67,68]. Gene expression or signature expression was represented as a function of latent time with R[69]. Moreover, the integration, normalization, imputation and trajectory analysis were performed independently for the subset of cells collected from primary tumours, incipient metastasis and macrometastases.

Creation of CRC transcriptomic meta-cohort

Public CRC transcriptomic datasets were downloaded from the Gene Expression Omnibus (GEO)[70] and NCI GDC commons[71], preprocessed and homogenized into a unique meta-cohort, including 1,830 samples from TCGA[72], GSE38832 (ref. [73]), GSE44076 (ref. [74]), GSE33113 (ref. [75]), GSE14333 (ref. [76]), GSE39582 (ref. [77]) and GSE37892 (ref. [78]) (Supplementary Table 1). The last four datasets include disease-free survival information with a median follow-up of 3.7 years and clinical information across all datasets is gender, age, stage and location of primary tumour (Supplementary Table 1). When not available, MSI status was imputed using the transcriptomic signature reported in ref. [79] through density-based nonparametric clustering[80,81]. Signature scores were computed as the scaled mean of the genes in the signature after scaling the expression matrix.

Microarray data were processed separately using RMA[52]. Information about sample processing and hybridization was retrieved from the CEL files. For TCGA, the Legacy version was used for expression with clinical annotation from October 2016. Genes were annotated using the Ensembl Biomart database (GRCh37)[82,83]. Duplicated samples across platforms were removed from the GA dataset, as well as samples from locations other than the colon or rectum. RSEM[84] expression values were already $\log_2$-transformed and quantile-normalized. Samples TCGA-A6-2679-01A and TCGA-AA-A004-01A were excluded owing to an anomalous expression distribution among all samples.

Each microarray series was corrected by Eklund metrics[54], centre and scanning date using a mixed-effect linear model[85]. Age, gender, stage, site and MSI were also included in these models. TCGA sets were corrected for occurrences of combinations of centre and plate identifier (random effect). For microarrays, probe sets were summarized at the gene level using the first principal component from probe sets in that gene. This component was then centred and scaled to the weighted mean of the means and s.d. values of the probe sets. The sign of the component was changed to the sign of the probe set contributing the most. All of the datasets were merged after gene-wise standardization to the GSE39582 series according to the distribution of gender, age, MSI and stage using undersampling: a subsample of the same number of patients and the same distribution according to these clinical variables was selected from the GSE39582 series for each dataset. Expression values were truncated to the maximum and minimum values observed in the reference dataset.

Gene screening for association with relapse

Each gene in the metacohort was evaluated for linear association of its expression with recurrence using a frailty Cox proportional hazards model[86,87], including dataset and technical variables. Statistical significance was assessed using a Wald test. HR values and their corresponding 95% confidence intervals were computed as a measure of association. The 2,530 genes with HR > 1 and $P < 0.05$ were included in the all_HR signature.

We used the GSE39397 dataset[6], which includes expression profiles of epithelial cancer cells (EPCAM$^+$), CAFs (FAP$^+$), leukocytes (CD45$^+$) and endothelial cells (CD31$^+$) isolated by FACs from dissociated primary CRCs ($n = 14$), to classify genes according to their expressions in these populations. The EpiHR signature contains genes from the allHR signature, which are upregulated (fold change > 1; $P < 0{,}05$) in EPCAM$^+$ cells compared with the three TME populations (FAP$^+$, CD31$^+$ and CD45$^+$). AllHR genes that did not pass this cut-off comprised the TME-HR signature.

Signature scores were computed as the scaled mean of the genes in the signature after scaling the expression matrix. The association with recurrence of the EpiHR signature in the whole dataset as well as the subclasses of consensus molecular subtype samples was assessed as described above. The likelihood ratio test comparing the EpiHR and AllHR P value was computed using the drop1 R function. Kaplan–Meier plots were generated using the survfit and plot functions.

Association between clinical variables and the EpiHR signature in the CRC metacohort was assessed by fitting a linear model for each variable of interest independently. Technical factors (dataset and centre) were included as covariates.

Association between oncogenic alterations and the EpiHR score

Annotations of oncogenic alterations for the TCGA samples were downloaded from ref. [88]. A Wilcoxon test comparing the expression of mutated versus wild-type samples was performed independently for every alteration. The difference in expression median values was used as a measure of the impact of each mutation in the gene expression.

Patient 10x single-cell analysis

Count matrices were downloaded from ArrayExpress (E-MTAB-8107 for samples EXT001, EXT002, EXT003, EXT009, EXT010, EXT011, EXT012, EXT013, EXT014, EXT018, EXT019, EXT020, EXT021, EXT022, EXT023, EXT024, EXT025, EXT026, EXT027 and EXT028) and GEO (GSE132465 for all Samsung Medical Center tumour (SMC-T) samples)[8]. The remaining EXT samples were processed as referred in E-MTAB-8107 and deposited in ArrayExpress under accession number E-MTAB-9934. Cells with mitochondrial content higher than 20%, less than 1,000 counts, more than 6,000 or less than 200 genes were discarded. Ribosomal genes

were also removed from the matrix to avoid technical biases during normalization. Samples with less than 500 cells that were not classified as core tumour were discarded from further analyses. The Korea (SMC samples) and Leuven (EXT samples) cohorts were processed independently according to the recommendations of the R package Seurat v.3[62]: samples were combined and normalized using the SCTransform function regressing mitochondrial percentage, with the method "glmGamPoi", min.cells = 1 and return.only.genes=FALSE to keep the maximum number of genes. Dimensionality reduction and visualization were performed using RunPCA and RunUMAP, with 26 principal components. Expression was imputed and smoothened using the MAGIC algorithm[64]. The expression of the *EPCAM* gene was used to define the connected components corresponding to epithelial cells.

The epithelial component of each cohort was processed as follows: cells with less than 1,000/3,000 (SMC/KUL) counts and less than 200/1,250 genes detected were removed from the dataset. No further normalization was applied. RunPCA, RunUMAP, FindNeighbors and FindClusters were applied, with 5/7 principal components and resolution of 0.7. Expression was imputed and smoothened using the MAGIC algorithm. Signature scores were computed as the mean value of the MAGIC expression per cell for all genes in the signature. The EpiHR and *Lgr5* cell populations were defined as cells with the corresponding signature expression above the 75th percentile. Population markers were found using FindMarkers. Functional enrichment was computed through the GSEA implementation in ref. [89] with genes ordered by fold change. Samples in the SMC and KUL datasets were annotated according to their iCMS class[11].

To identify the core gene expression program upregulated in HRCs, we computed the correlation scores for all genes with the EpiHR signature in the SMC and KUL human dataset. The resulting list was ranked by the average correlation across all samples (Supplementary Table 2). The coreHRC signature was defined as the top 100 genes in the SMC dataset.

Association of HRCs and TME populations

We classified clusters according to the expression of known markers of microenvironment components[8]. For each population we estimated the association between its percentage per sample and the percentage of HRCs. Spearman correlation coefficient and *P* value were used to assess said association.

Clustering of EpiHR genes according to expression correlation

We computed the Pearson correlation coefficient for all pairwise combinations of genes in the EpiHR signature. We then applied hierarchical clustering (hclus in R with method 'complete') and defined 6 clusters using the cutree function. After visual inspection, we decided to merge the three clusters with higher correlation, resulting in cluster 1 (Extended Data Fig. 1h).

Ethics oversight

Samples of primary CRC from patients used for IF and ISH analysis were obtained from the Hospital Clinic de Barcelona-IDIBAPS Biobank (B.0000575), which is integrated in the Spanish National Biobanks Network. Samples were donated by patients under informed consent and they were processed according to standard operating procedures with the appropriate approval of the Ethics and Scientific Committee of the Hospital Clinic de Barcelona (register code: HCB/2020/1478) and according to the guidelines of the European Network of Research Ethics Committees, following European, national and local laws.

Statistics

No statistical test was used to determine sample size upfront. Instead, sample size was determined empirically according to previous knowledge of the variation in the experimental set-up[4,12,34]. A minimum of four mice were quantified in each experiment and for each condition. For the majority of in vitro experiments, we used $n > 3$ according to previous experience with similar experiments. Additional information is provided in the Reporting summary. Automated blind quantifications and blind data analysis were performed whenever possible. The sample size typically results in a standard error <25% of the mean. No data from in vitro or in vivo experiments were excluded, except for the CRC relapse model, in which a small fraction of mice, typically 1 in every 10 in each experiment, were not included in the follow up due to invasion into the ileocaecal junction, which impeded successful surgery. Occasionally, mice bearing CRCs were euthanized 1–2 days after surgery owing to significant weight loss or unhealthy aspect, according to protocols approved by the animal experimental committees. Data distributions were assumed to be normal but this was not formally tested. Transformations were applied whenever needed. For percentage data, normality was assumed for values far from 0 or 100. Statistical analyses were performed using the R software package (v.4.0.5) and GraphPad Prism (v.7.03).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data relevant to this study are available from the corresponding author on reasonable request. Expression arrays and RNA-seq data are available at the Gene Expression Omnibus (GEO). Accession numbers for gene expression sequencing experiments reported in this paper are GEO: GSE190055 (arrays, $EMP1^{high}$ versus $EMP1^{low}$ AKTP tumour cells), GSE208139 (arrays, MTOs co-cultured with fibroblasts), GSE207974 (RNA-seq chemotherapy) and GSE207668 (RNA-seq CTOs). Count matrices for single-cell RNA-seq experiments were deposited at ArrayExpress under accession number E-MTAB-11284 (10x AKTP primary tumours), E-MTAB-11302 (Smart-seq metastatic progression) and E-MTAB-11981 (Smart-seq AKP micrometastases). Additional metadata and processed data files, including UMAP embeddings and gene signature scores, are available at Synapse (syn35000645). Source data are provided with this paper.

40. Matano, M. et al. Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids. *Nat. Med.* **21**, 256–262 (2015).
41. Drost, J. et al. Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**, 43–47 (2015).
42. Céspedes, M. V. et al. Orthotopic microinjection of human colon cancer cells in nude mice induces tumor foci in all clinically relevant metastatic sites. *Am. J. Pathol.* **170**, 1077–1085 (2007).
43. Chen, Y.-C. et al. Gut fecal microbiota transplant in a mouse model of orthotopic rectal cancer. *Front. Oncol.* **10**, 568012 (2020).
44. Conti, S. et al. CAFs and cancer cells co-migration in 3D spheroid invasion assay. *Methods Mol. Biol.* **2179**, 243–256 (2020).
45. Gonzalez-Roca, E. et al. Accurate expression profiling of very small cell populations. *PLoS ONE* **5**, e14418 (2010).
46. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
48. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
51. Bolstad, B. M. et al. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman, R. et al.) (Springer, 2005).
52. Fridlyand, J. Microarray Data Analysis. in *Selected Works in Probability and Statistics* (ed Dudoit, S.) https://doi.org/10.1007/978-1-4614-1347-9_15 (Springer, 2012).
53. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
54. Eklund, A. C. & Szallasi, Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.* **9**, R26 (2008).
55. Wu, D. et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).
56. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).

57. Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, e1000217 (2008).
58. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
59. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
60. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
61. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
62. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
63. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
64. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
65. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
66. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
67. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
68. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
69. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
70. Barrett, T. & Edgar, R. [19] Gene Expression Omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.* **411**, 352–369 (2006).
71. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
72. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
73. Tripathi, M. K. et al. Nuclear factor of activated T-cell activity is associated with metastatic capacity in colon cancer. *Cancer Res.* **74**, 6947–6957 (2014).
74. Sanz-Pamplona, R. et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol. Cancer* **13**, 46 (2014).
75. Kemper, K. et al. Mutations in the Ras-Raf axis underlie the prognostic value of CD133 in colorectal cancer. *Clin. Cancer Res.* **18**, 3132–3141 (2012).
76. Jorissen, R. N. et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. *Clin. Cancer Res.* **15**, 7642–7651 (2009).
77. Marisa, L. et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
78. Laibe, S. et al. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. *OMICS* **16**, 560–565 (2012).
79. Jorissen, R. N. et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin. Cancer Res.* **14**, 8061–8069 (2008).
80. Azzalini, A. & Menardi, G. Clustering via nonparametric density estimation: the R package pdfcluster. *J. Stat. Softw.* **57**, 1–26 (2014).
81. Azzalini, A. & Torelli, N. Clustering via nonparametric density estimation. *Stat. Comput.* **17**, 71–80 (2007).
82. Smedley, D. et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**, W589–W598 (2015).
83. Drost, H. G. & Paszkowski, J. Biomartr: genomic data retrieval with R. *Bioinformatics* **33**, 1216–1217 (2017).
84. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
85. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
86. Therneau, T. M., Grambsch, P. M. & Pankratz, V. S. Penalized survival models and frailty. *J. Comput. Graph. Stat.* **12**, 156–175 (2003).
87. Therneau, T. coxme: mixed effects Cox models. R package version 2.2-3 www.cran.R-project.org/package=coxme.Oikos (2012).
88. Sanchez-Vega, F. et al. Oncogenic signaling pathways in the Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
89. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

Acknowledgements We thank all of the members of the laboratory for their support and discussions; the staff at the IRB Barcelona core facilities for biostatistics, histopathology, functional genomics and advanced digital microscopy, as well as the flow cytometry, animal facilities of the UB/PCB, and the CRG genomic unit for assistance; the staff at the HCB-IDIBAPS Biobank for sample and data procurement. Sample collection of this work was also supported by the Xarxa de Bancs de Tumours de Catalunya (XBTC) sponsored by Pla Director d'Oncologia (PDO). A.C.-S., G.T. and L.J.-G. have held FPU fellowships from Spanish Ministry of Universities and A.A.-V. has held a La Caixa predoctoral fellowship. S.C. holds a FPI fellowship from the Spanish Ministry of Economy and Competitiveness (MINECO). H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII) and the Agencia Estatal de Investigación (AEI) and FEDER (SAF2017-89109-P). This work has been supported by ERC advanced grants 884623 (residualCRC to E.B.) and 883739 (Epifold to X.T.); LCF/PR/HR19/52160018 from La Caixa foundation; PID2020-119917RB-I00 from the Spanish MICINN; and CRUK Accelerator Award C7932/A26825 (ACRCelerate), in collaboration with AECC (grant GEACC19006BAT_2021). E.B., X.T. and H.H. are supported by the Fundació La Marató de TV3 (201903-30-31-32). H.H. received support for the project PID2020-115439GB-I00- funded by MCINN/AEI/10.13039/501100011033. S.L. was supported by a Wellcome Trust Senior Clinical Research Fellowship (206314/Z/17/Z)), E.J.M. is funded by the Lee Placito Medical Research Fund (University of Oxford). IRB Barcelona and IBEC are recipients of a Severo Ochoa Award of Excellence from MINECO. Single-cell profiling of CRC samples was supported by the Belgian Federation against Cancer grant nos 2018-127 and 2016-133 and by a grant from Fondation Roi-Baudouin. S.T. is supported by a Fundamental Clinical Researchers KU Leuven grant and Foundation against Cancer grant for this work. The Magdalena Socias Moyà fund supports metastasis research at IRB Barcelona.

Author contributions E.B. and A.C.-S. conceived the study, coordinated experiments and wrote the manuscript. A.C.-S. designed and performed key experiments including the profiling of residual disease, the analyses of tumour buds and micrometastases, genetic ablation studies and immunotherapy experiments in the CRC relapse model. A.C.-S., C.C., G.T., F.S. and D.S. generated and characterized MTO knock-in lines. A.C.-S., X.H.-M., S.P.-P. and G.T. developed the CRC relapse model. A.C.-S., X.H.-M. and S.P.-P. performed all mouse work. L.M. and C.S.-O.A. analysed scRNA-seq data. C.S.-O.A. and A.B.-L. analysed human CRC transcriptomic datasets. A.C.-M., L.M. and C.S.-O.A. performed statistical analyses. C.C. and T.S. performed IF and imaged organoids in vitro. A.C.-S., C.C. and O.R. generated and characterized YAP-KD models. A.C.-S., C.C., O.R., S.C. and X.T. performed in vitro co-cultures. A.C.-S. and A.A.-V. quantified immunofluorescence stainings. A.C.-S. and N.F. developed the method to purify residual tumour cells in whole livers. M.S. performed IF and IHC. E.J.M. and S.L. performed multiplex IF. L.N. performed chemotherapy experiments. L.B. and J.C. performed 3D light-sheet imaging. L.J.-G., C.C., P.L. and H.H. provided support with scRNA-seq experiments. S.T. generated scRNA-seq data from CRC patient samples. D.V.F.T. and D.S. generated MTO and CTO biobanks. E.S. provided strategic support and helped with figures and manuscript writing. E.B. supervised the study.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05402-9.

Correspondence and requests for materials should be addressed to Eduard Batlle.

Peer review information *Nature* thanks Hugo Snippert, Itay Tirosh, Nicola Valeri and the other, anonymous, reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

Appendix C. Mex3a marks drug-tolerant persistent colorectal cancer cells that mediate relapse after chemotherapy[6]

Adrián Álvarez-Varela, Laura Novellasdemunt, Francisco M. Barriga, Xavier Hernando-Momblona, Adrià Cañellas-Socias, Sara Cano-Crespo, Marta Sevillano, Carme Cortina, Diana Stork, Clara Morral, Gemma Turon, Felipe Slebe, **Laura Jiménez-Gracia**, Ginevra Caratù, Peter Jung, Giorgio Stassi, Holger Heyn, Daniele V.F. Tauriello, Lidia Mateo, Sabine Tejpar, Elena Sancho, Camille Stephan-Otto Attolini, Eduard Batlle†. ***Mex3a* marks drug-tolerant persistent colorectal cancer cells that mediate relapse after chemotherapy**, *Nature Cancer.* 3, 1052–1070 (2022).

DOI: https://doi.org/10.1038/s43018-022-00402-0

Impact Factor 2022: 22.7 (Q1)

Raw data: Accessible via the Sequence Read Archive (SRA), NCBI Gene Expression Omnibus (GEO) and ArrayExpress databases. Further details, including specific accession numbers, are provided in the *Data availability* section of the article.

Processed data: Available upon corresponding author's request.

Code: Available upon corresponding author's request.

[6] In this paper, I collaborated by analyzing the scRNA-seq data together with the Bioinformatics Facility from Institute for Research in Biomedicine (IRB).

ARTICLES
https://doi.org/10.1038/s43018-022-00402-0

nature cancer

Check for updates

Mex3a marks drug-tolerant persister colorectal cancer cells that mediate relapse after chemotherapy

Adrián Álvarez-Varela[1,2], Laura Novellasdemunt[1], Francisco M. Barriga[1,3], Xavier Hernando-Momblona[1,2], Adrià Cañellas-Socias[1,2], Sara Cano-Crespo[1], Marta Sevillano[1,2], Carme Cortina[1,2], Diana Stork[1], Clara Morral[1], Gemma Turon[1], Felipe Slebe[1], Laura Jiménez-Gracia[4], Ginevra Caratù[4], Peter Jung[5,6], Giorgio Stassi[7], Holger Heyn[4,8], Daniele V. F. Tauriello[1,9], Lidia Mateo[1], Sabine Tejpar[10], Elena Sancho[1,2], Camille Stephan-Otto Attolini[1] and Eduard Batlle[1,2,11] ✉

Colorectal cancer (CRC) patient-derived organoids predict responses to chemotherapy. Here we used them to investigate relapse after treatment. Patient-derived organoids expand from highly proliferative LGR5$^+$ tumor cells; however, we discovered that lack of optimal growth conditions specifies a latent LGR5$^+$ cell state. This cell population expressed the gene *MEX3A*, is chemoresistant and regenerated the organoid culture after treatment. In CRC mouse models, Mex3a$^+$ cells contributed marginally to metastatic outgrowth; however, after chemotherapy, Mex3a$^+$ cells produced large cell clones that regenerated the disease. Lineage-tracing analysis showed that persister Mex3a$^+$ cells downregulate the WNT/stem cell gene program immediately after chemotherapy and adopt a transient state reminiscent to that of YAP$^+$ fetal intestinal progenitors. In contrast, Mex3a-deficient cells differentiated toward a goblet cell-like phenotype and were unable to resist chemotherapy. Our findings reveal that adaptation of cancer stem cells to suboptimal niche environments protects them from chemotherapy and identify a candidate cell of origin of relapse after treatment in CRC.

The standard of care for advanced CRC includes treatment with chemotherapeutic drugs that target the cell proliferation machinery[1]. In patients with CRC with overt metastases, chemotherapy initially halts tumor growth but, almost inevitably, disease progresses after some cycles of treatment. Adjuvant chemotherapy is also administered to eliminate minimal residual disease, yet it only diminishes the risk of relapse by 10–25%[2]. Several adaptive mechanisms have been proposed as the culprit of therapy resistance in CRC, including adaptive mutability[3], epigenetic remodeling[4,5] or the acquisition of a diapause-like program[6,7]. Less clear is how the extensive phenotypic heterogeneity and plasticity present in CRC before the treatment influence chemotherapy resistance and which cell states are adopted by drug-tolerant persister cells during treatment and relapse.

CRC patient-derived organoids (PDOs) predict responses to chemotherapy[8–11] and are therefore excellent models to investigate the mechanisms behind the limited benefit of the treatment. The standard organoid culture medium mimics the ideal colonic stem cell niche[12–16]. Supplementation with WNT promotes the expression of the stem cell program, epidermal growth factor (EGF) provides mitogenic signals and inhibitors of transforming growth factor (TGF)-β and bone morphogenetic protein signaling are necessary to counteract cytostasis and differentiation[12–16]. The most common genetic alterations during CRC progression occur in genes that control key signaling pathways that allow colon stem cell expansion. Mutations in Apc, Kras and Smad4 or Tgfbr2 render tumor stem cells independent of the crypt niche and facilitate growth in foreign environments[17]. Supporting this view, we and others showed that colon organoids engineered with four driver mutations (Apc, Kras, P53 and Tgfbr2 or Smad4; AKPT/AKPS) can be cultured in the absence of stem cell factors and, upon inoculation in mice, they generate metastatic outgrowths in the liver and lungs with high efficiency[18–21]; however, the vast majority of human primary CRCs and metastases only carry mutations in two or three driver pathways[12,16,22]. Indeed, the establishment of PDO biobanks showed that a large proportion of CRCs arising in patients exhibit partial dependency on niche signals and require supplementation with complete stem cell medium for expansion[12,16]. These conditions favor continuous expansion of cancer stem cells for all tumor genotypes and enable the establishment of PDOs from most patient samples; however, we hypothesized that in intact primary CRCs and metastasis, cancer stem cells may encounter suboptimal tumor niches that do

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain. [3]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [5] German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]German Cancer Consortium (DKTK), Partner site Munich, Institute of Pathology, Ludwig Maximilian University, Munich, Germany. [7]Department of Surgical Oncological and Stomatological Sciences, University of Palermo, Palermo, Italy. [8]Universitat Pompeu Fabra, Barcelona, Spain. [9]Department of Cell Biology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands. [10]Molecular Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium. [11]ICREA, Barcelona, Spain. ✉e-mail: eduard.batlle@irbbarcelona.org

not support their growth. To explore this possibility, we leverage mouse and human CRC organoids engineered with different driver mutations. Given that tumor organoids recreate the cellular heterogeneity and organization of CRCs[12,15,16], we speculated that they may also contain drug-tolerant persister cells. We discovered that cancer stem cells cultured with suboptimal niche factor medium co-opt a latent Mex3a$^+$ state. These cells exist in a large proportion of human CRCs and behave as a bonafide drug-tolerant persister (DTP) cell population capable of regenerating the tumor after chemotherapy in in vitro and in vivo models.

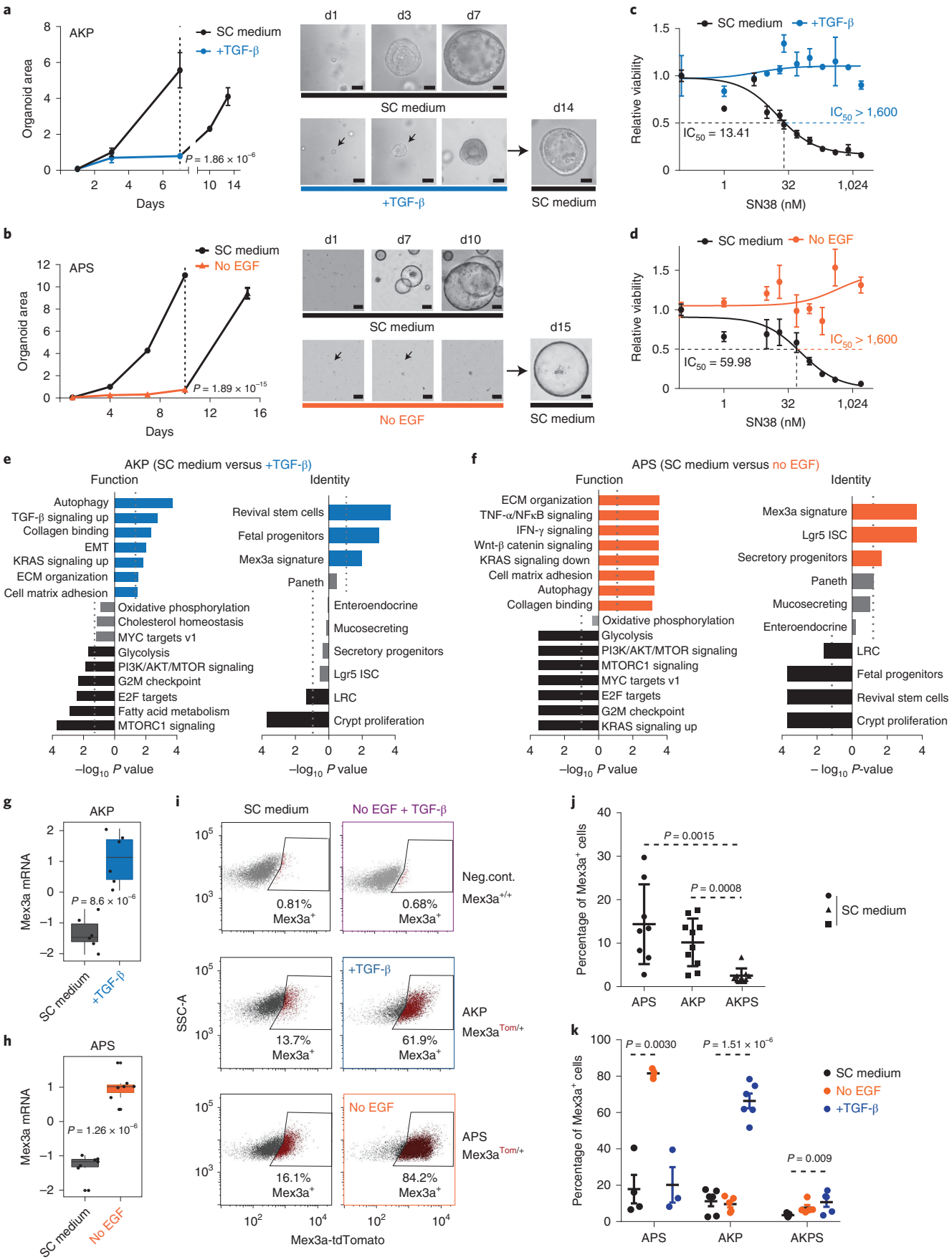
Results

Suboptimal stem cell niche induces a latent Mex3a$^+$ state. We cultured mouse tumor organoids (MTOs) engineered with defined genetic alterations in driver pathways[21] in suboptimal niche factor conditions. Supplementation with recombinant TGF-β arrested the growth of AKP (Apc$^{-/-}$; Kras G12D$^{/+}$; P53$^{-/-}$) MTO cultures (Fig. 1a). Likewise, MTOs bearing wild-type KRAS alleles, APS (Apc$^{-/-}$; P53$^{-/-}$; Smad4$^{-/-}$) or APT (Apc$^{-/-}$; P53$^{-/-}$; Tgfbr2$^{-/-}$), did not expand in the absence of EGF (Fig. 1b and Extended Data Fig. 1a). We, however, found that MTOs remained viable for more than a week under these suboptimal culture conditions. AKP MTOs resumed expansion after removal of TGF-β (Fig. 1a), whereas supplementation of EGF after 10 d restored growth rates of APS (Fig. 1b) or APT organoids (Extended Data Fig. 1a). While performing these experiments, we discovered that the suboptimal niche factor combination (no EGF for APS or APT and TGF-β supplementation for AKP) conferred broad chemotherapy resistance to triple-mutant MTOs. Cell viability assays demonstrated a large increase in half-maximum inhibitory concentration (IC_{50}) values for fluorouracil (5FU), oxaliplatin and SN38 (the active metabolite of irinotecan), the three chemotherapeutic drugs most commonly used to treat patients with CRC, in arrested versus exponentially growing triple-mutant MTOs (Fig. 1c,d and Extended Data Fig. 1b–e). MTOs with single loss-of-function mutations in Apc (A) or with Apc combined with p53 (AP) mutations also acquired chemotherapy resistance under suboptimal niche factor conditions (Extended Data Fig. 1c,d). In contrast, quadruple-mutant AKPS or AKPT MTOs kept proliferating in a no-EGF/+TGF-β medium and remained sensitive to chemotherapy (Extended Data Fig. 1c–e). The DTP state was reversible, shown by experiments where APS MTOs underwent two consecutive cycles of culturing under suboptimal stem cell conditions and then were re-challenged with chemotherapy (Extended Data Fig. 1f).

These observations suggest that CRC cells exposed to a suboptimal niche environment acquire a latent phenotype and gain intrinsic resistance to chemotherapy. To gain insights into this phenomenon, we performed RNA sequencing (RNA-seq) followed by gene set enrichment analysis (GSEA) (Fig. 1e,f, Extended Data Fig. 1g and Supplementary Table 1). TGF-β signaling in AKP MTOs induced the expression of the signature genes of revival stem cells[23] and fetal progenitors[24] (Fig. 1e), in agreement with previous studies[25]. TGF-β also induced upregulation of pathways and genes related to epithelial-to-mesenchymal transition (Fig. 1e). Lack of EGF receptor (EGFR) signaling in APS and APT MTOs increased the levels of a subset of secretory and Paneth cell-specific genes as previously reported for human CRCs[26] (Fig. 1f and Extended Data Fig. 1g). Although growth arrest triggered by lack of the mitogen EGF or by activation of TGF-β signaling involved substantially different mechanisms and downstream gene programs, we identified some shared features. The latent phenotype in both MTO genotypes was characterized by silencing of proliferation gene programs and general downregulation of metabolic pathways, including glycolysis, oxidative phosphorylation and fatty acid metabolism (Fig. 1e,f, Extended Data Fig. 1g and Supplementary Table 1). Extracellular matrix remodeling and sensing genes were also upregulated in both AKP and APS MTOs (Fig. 1e,f). It called our attention that the Mex3a$^+$ cell signature[27] was among the gene sets enriched in common between latent AKP and APS or APT MTO cultures (Fig. 1e,f and Extended Data Fig. 1g). Furthermore, Mex3a messenger RNA was upregulated in all three triple-mutant CRC genotypes (Fig. 1g,h and Extended Data Fig. 1h). Using AKP and APS MTOs engineered with a tdTomato reporter cassette inserted in the endogenous Mex3a locus, we confirmed that MTO cells gained Mex3a expression under suboptimal stem cell niche culture conditions (Fig. 1i). Quadruple-mutant AKPS MTOs contained fewer Mex3a$^+$ cells than triple-mutant MTOs when cultured in either complete stem cell medium (Fig. 1j) or no-EGF/+TGF-β conditions (Fig. 1k).

MEX3A marks latent LGR5$^+$ cells in human CRCs. We next investigated chemoresistance in PDOs from patients with CRC. Overall, 5 out the 11 PDOs analyzed acquired broad resistance to chemotherapy under suboptimal growth factor conditions (Fig. 2a,b and Extended Data Fig. 2a). A prime example was PDO7, which arrested in no-EGF/+TGF-β medium and became resistant to 5FU, SN38 and oxaliplatin (Fig. 2c and Extended Data Fig. 2a). PDO7 bears biallelic inactivating mutations in APC and SMAD4 plus a KRAS G13D allele[28] (Supplementary Table 2), which did not confer complete EGF independence (Fig. 2c), as previously observed in other models[29] and also shown for other KRAS-activating mutations in PDO biobanks[12,30]. Consistent with our findings in MTOs, we observed a fivefold increase in the frequency MEX3A$^+$ cells in PDO7 cultured in suboptimal stem cell medium, measured using a tdTomato reporter cassette integrated into the MEX3A locus (Fig. 2d–f).

Fig. 1 | Identification of Mex3a+ cells in chemoresistant MTOs. a,**b**, Quantification of organoid growth of AKP MTOs in the presence of TGF-β (**a**) and APS MTOs in the absence of EGF (**b**) compared to complete stem cell (SC) medium. Dashed lines indicate the day when TGF-β was removed or EGF was re-added, respectively; $n = 4$ technical replicates (independent tissue culture wells). Linear model on log-transformed values with no adjustment for multiple comparisons. The experiment was conducted three times with equivalent results. Scale bars, 50 µm. d, day. **c**,**d** Viability of AKP and APS MTOs grown upon treatment with increasing concentrations of SN38. The trend line was calculated using a LOESS model; $n = 3$ technical replicates (independent tissue culture wells). The experiment was conducted three times with equivalent results. **e**,**f**, GSEA of AKP (**e**) and APS MTOs (**f**) grown in SC medium versus either +TGF-β1 or no-EGF medium, respectively. Profiling was performed at day 7. Gene sets are ordered based on their $-\log_{10} P$ value and the dashed line indicates $P = 0.05$, calculated with rotation gene set testing for linear models. Signatures are detailed in Supplementary Table 6. EMT, epithelial-to-mesenchymal transition; ECM, extracellular matrix; LRC, label-retaining cells. **g**,**h**, Mex3a mRNA expression levels in AKP MTOs (**g**) treated with TGF-β versus control SC medium ($n = 6$ independent experiments); and in APS MTOs (**h**) grown in the absence of EGF versus control SC medium ($n = 8$ independent experiments). Boxes represent the first and third quartiles. The line indicates adjusted values corrected by experimental batch. Whiskers indicate maximum and minimum values. A linear model with no adjustment for multiple comparisons was used to calculate statistics value. **i**, Representative flow cytometry profiles of dissociated Mex3a$^{Tom/+}$ AKP and APS organoids. Profile of control MTOs (Mex3a$^{+/+}$) were used to establish the gates (top). **j**, Percentages of Mex3a$^+$ cells in three different genotypes in SC medium calculated by flow cytometry; $n = 8$ (APS), $n = 10$ (AKP) and $n = 9$ (AKPS) independent experiments. Unpaired two-tailed Wilcoxon test. **k**, Percentages of Mex3a$^+$ cells in the absence of EGF or in the presence of TGF-β1 in APS, AKP and AKPS organoids calculated by flow cytometry; $n = 3$ (APS), $n = 6$(AKP) and $n = 4$ (AKPS) independent experiments. Paired two-tailed Wilcoxon test. Data in **a**–**d**, **j** and **k** are presented as mean ± s.e.m.


a
AKP
SC medium
+TGF-β
Organoid area
Days
P = 1.86 × 10−6
d1
d3
d7
SC medium
d14
+TGF-β
SC medium
b
APS
SC medium
No EGF
Organoid area
Days
P = 1.89 × 10−15
d1
d7
d10
SC medium
d15
No EGF
SC medium
c
SC medium
+TGF-β
Relative viability
IC50 = 13.41
IC50 > 1,600
SN38 (nM)
d
SC medium
No EGF
Relative viability
IC50 = 59.98
IC50 > 1,600
SN38 (nM)
e
AKP (SC medium versus +TGF-β)
Function
Autophagy
TGF-β signaling up
Collagen binding
EMT
KRAS signaling up
ECM organization
Cell matrix adhesion
Oxidative phosphorylation
Cholesterol homeostasis
MYC targets v1
Glycolysis
PI3K/AKT/MTOR signaling
G2M checkpoint
E2F targets
Fatty acid metabolism
MTORC1 signaling
−log10 P value
Identity
Revival stem cells
Fetal progenitors
Mex3a signature
Paneth
Enteroendocrine
Mucosecreting
Secretory progenitors
Lgr5 ISC
LRC
Crypt proliferation
−log10 P value
f
APS (SC medium versus no EGF)
Function
ECM organization
TNF-α/NFκB signaling
IFN-γ signaling
Wnt-β catenin signaling
KRAS signaling down
Cell matrix adhesion
Autophagy
Collagen binding
Oxidative phosphorylation
Glycolysis
PI3K/AKT/MTOR signaling
MTORC1 signaling
MYC targets v1
E2F targets
G2M checkpoint
KRAS signaling up
−log10 P value
Identity
Mex3a signature
Lgr5 ISC
Secretory progenitors
Paneth
Mucosecreting
Enteroendocrine
LRC
Fetal progenitors
Revival stem cells
Crypt proliferation
− log10 P-value
g
AKP
Mex3a mRNA
P = 8.6 × 10−6
SC medium
+TGF-β
h
APS
Mex3a mRNA
P = 1.26 × 10−6
SC medium
No EGF
i
SC medium
No EGF + TGF-β
Neg.cont.
Mex3a+/+
0.81%
Mex3a+
0.68%
Mex3a+
+TGF-β
AKP
Mex3aTom/+
13.7%
Mex3a+
61.9%
Mex3a+
No EGF
APS
Mex3aTom/+
16.1%
Mex3a+
84.2%
Mex3a+
SSC-A
Mex3a-tdTomato
j
Percentage of Mex3a+ cells
P = 0.0015
P = 0.0008
APS
AKP
AKPS
SC medium
k
Percentage of Mex3a+ cells
P = 0.0030
P = 1.51 × 10−6
P = 0.009
APS
AKP
AKPS
SC medium
No EGF
+TGF-β

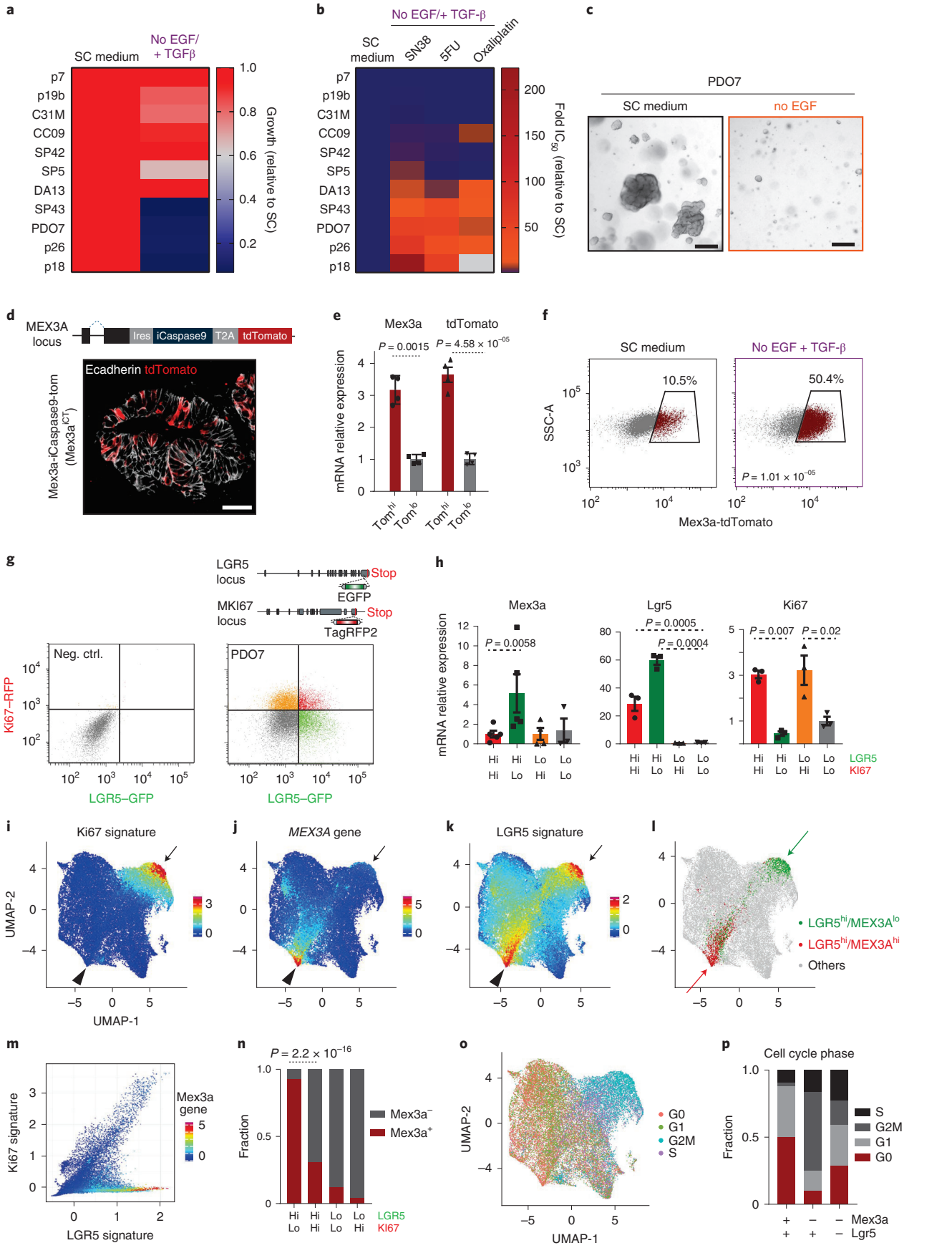
We also engineered PDO7 with Ki67–RFP and LGR5–GFP knock-in cassettes[28,31] and found upregulation of MEX3A mRNA in LGR5–GFP^{hi}/Ki67^{lo} cells purified by FACS from xenografts (Fig. 2g,h).

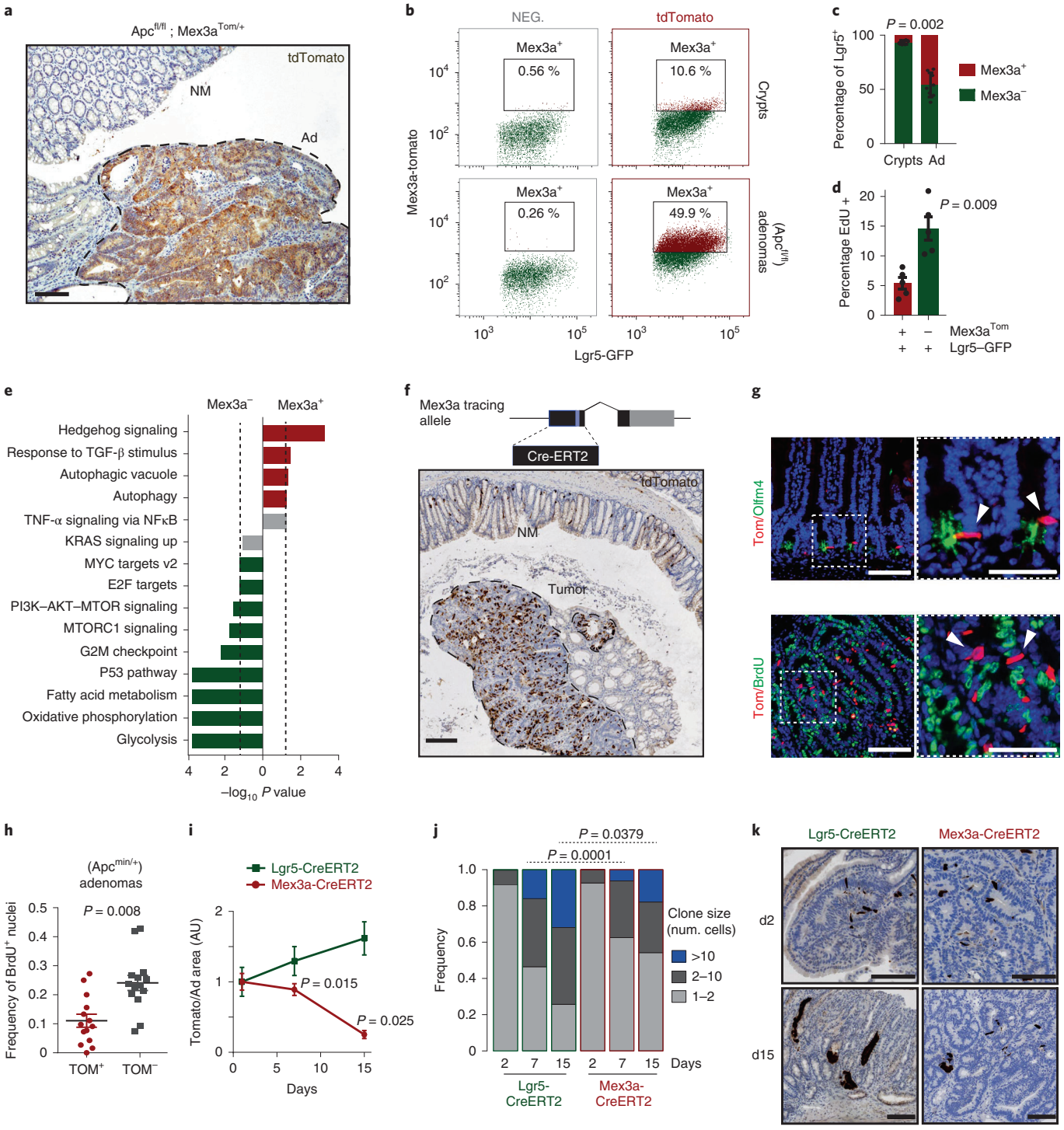
To broaden the relevance of these findings, we studied the relationship between MEX3A, LGR5 and cell proliferation in human CRC using a patient sample cohort analyzed by single-cell RNA-seq (scRNA-seq)[32]. Analysis of pooled epithelial tumor cells corresponding to 37 CRC samples revealed two subsets of LGR5^{hi} cells defined by either high or low expression of the Ki67 proliferation signature (Fig. 2i–l). The LGR5^{hi}/KI67^{lo} cell population was specifically marked by MEX3A mRNA (Fig. 2m,n). Analysis of individual samples in this cohort confirmed that two-thirds contained a LGR5^{hi}/MEX3A^{hi} cell population characterized by low expression of proliferation genes (Extended Data Fig. 2b). According to the expression levels of cell cycle signature genes, 50% of LGR5^{hi}/MEX3A^{hi} cells were predicted to reside in G0 and 38% in G1 phase of the cell cycle (Fig. 1o,p). In contrast, most LGR5^{hi}/MEX3A^{lo} cells were in G2/M and S phases (Fig. 1o,p). Furthermore, a gene expression signature containing the top 100 differentially expressed genes in LGR5^{hi}/MEX3A^{hi} versus LGR5^{hi}/MEX3A^{lo} cells predicted increased risk of disease relapse in a large transcriptomic cohort from patients with CRC ($n=816$ AJCC stage I–III) even after adjusting for clinical variables (age, sex, site, microsatellite instability (MSI) status and stage; hazard ratio (HR) = 1.50, $P=0.007$; Fig. 2l, Extended Data Fig. 2c,d and Supplementary Table 3), whereas LGR5^{hi}/MEX3A^{lo} cells associated with good prognosis (HR = 0.66, $P=0.006$; Extended Data Fig. 2e,f). From all these results, we tentatively conclude that MEX3A marks a subset of latent LGR5^{+} cells that associates with chemoresistance in both mouse and human CRC.

Clonal dynamics of Lgr5^{+} and Mex3a^{+} cells in adenomas. The *Mex3a* gene encodes an RNA-binding protein expressed by a restricted subset of Lgr5^{+} intestinal stem cells (ISCs) in healthy crypts[27]. Lgr5^{hi}/Mex3a^{hi} cells are multipotent ISCs that exhibit a low proliferative rate[27]. To investigate the relationship between Mex3a and Lgr5 during intestinal carcinogenesis, we generated Apc$^{fl/fl}$ mice bearing both Mex3a-tdTomato[27] and Lgr5-EGFP-CreERT2 (ref. [33]) alleles (Fig. 3a–c). We induced recombination of the Apc locus in Lgr5^{+} cells and analyzed Lgr5–GFP^{+} cell populations. Consistent with our previous study[27], only few (~10%) Lgr5^{+} cells in healthy colonic mucosa expressed the Mex3a-tdTomato reporter (Fig. 3b,c and Extended Data Fig. 3a). In colon adenomas, however, around 50% of Lgr5^{+} cells were marked by Mex3a-tdTomato (Fig. 3b,c and Extended Data Fig. 3b). The Lgr5^{+}/Mex3a^{+} cell population exhibited lower proliferation rates than Lgr5^{+}/Mex3a^{-} cells shown by 5-ethynyl-2′-deoxyuridine (EdU) incorporation (Fig. 3d). Gene expression profiling confirmed global downregulation of proliferation and metabolism-related gene sets in Mex3a^{+}/Lgr5^{+} adenoma cells (Fig. 3e). This analysis also revealed increased TGF-β signaling and downregulation of EGFR/KRAS signature genes among the top significantly enriched gene sets, suggesting a role for these pathways in the specification of the Lgr5^{+}/Mex3a^{+} population in adenomas (Fig. 3e and Supplementary Table 4). Indeed, Apc-mutant adenoma-derived organoids cultured in the absence of EGF or supplemented with TGF-β exhibited a large increase in the number of Mex3a-tdTomato^{+}/Lgr5−GFP^{+} cells (Extended Data Fig. 3c,d) and, as we previously showed, resisted chemotherapy (Extended Data Fig. 1d). We also assessed the organoid forming capacity of Mex3a^{+} and Lgr5^{+} populations (Extended Data Fig. 3e–g), a property that correlates with cancer stem cell potential. Mex3a^{+}/Lgr5^{+} tumor cells isolated from adenomas formed organoids more efficiently than Mex3a^{-}/Lgr5^{+} cells (Extended Data Fig. 3e) albeit organoids generated by these two populations exhibited similar growth rates (Extended Data Fig. 3f). Of note, organoids that originated from the two populations contained Mex3a^{+} and Lgr5^{+} cells in similar proportions (Extended Data Fig. 3g). Moreover, Mex3a^{-}/Lgr5^{+} adenoma cells cultured in suboptimal niche factor conditions (no EGF or +TGF-β) acquired a latent Mex3a^{+}/Lgr5^{+} state (Extended Data Fig. 3h,i). Overall, these findings denote that the Mex3a program is not hardwired and demonstrate extensive plasticity between the active Mex3a^{-} and latent Mex3a^{+} cancer stem cell states depending upon optimal and suboptimal niche environments.

We next introduced the Mex3a-CreERT2 allele in APC$^{Min/+}$; Rosa26^{tdTomato} mice. At 24–48 h after induction with tamoxifen, we observed a tdTomato-labeled cell population in both healthy epithelium and adenomas (Fig. 3f,g). In normal crypts, Mex3a-CreERT2-mediated recombination was restricted to 1–2 cells per crypt located preferentially around position +4 (Fig. 3g), which is consistent with our previous study[27]. In contrast, the frequency of recombined cells was largely increased in adenomas (Fig. 3f,g). At early time points after tamoxifen administration (16 h), tdTomato^{+} adenoma cells retained increased levels of Mex3a mRNA (Extended Data Fig. 3j) and showed lower bromodeoxyuridine (BrdU) incorporation than tdTomato^{-} cells (Fig. 3g,h). We next performed clonal analysis in Apc$^{Min/+}$ adenomas using Mex3a-CreERT2 and Lgr5-CreERT2 alleles. In these experiments, we inoculated a low tamoxifen dose so that the progeny of individual cells could be tracked over time. Quantification of tdTomato labeling on tissue sections at the initial time point (2 d) indicated that we marked similar numbers of Mex3a^{+} and Lgr5^{+} cells (1.05 ± 0.01% versus 0.53 ± 0.01% of the adenoma area, respectively). Although the quantification of clone number and size in two dimensions precluded conclusions about the absolute contribution of Mex3a^{+} and Lgr5^{+} cells to tumor growth, a comparative analysis revealed important differences in the dynamics of the two populations (Fig. 3i–k and

Fig. 2 | Characterization of MEX3A^{+} cells in human CRCs. a, Heat map summarizing the effect on organoid growth in medium supplemented with recombinant active TGF-β1 and in absence of EGF compared to full SC medium in a collection of human PDOs. **b**, Heat map summarizing the increase in SN38, 5FU and oxaliplatin IC_{50} values, promoted by absence of EGF or presence of TGF-β1 in the indicated PDOs. **c**, Representative pictures of growth arrest of PDO7 cultured in absence of EGF. Scale bar, 100 µM. **d**, CRISPR–Cas9 targeting strategy to introduce an inducible ablation and reporter cassette into the MEX3A locus in PDO7. Immunofluorescence shows tdTomato and E-cadherin in PDO7-edited xenografts. Scale bar, 50 µM. **e**, RT–qPCR analysis of MEX3A and tdTomato in sorted epithelial (EPCAM^{+}) Mex3a-tdTomato^{hi} versus Mex3a-tdTomato^{lo} cells from PDO7 xenografts; $n=4$ mice, paired two-tailed Student's *t*-test. **f**, Representative flow cytometry profiles of dissociated Mex3a$^{Tom/+}$ PDO7-edited organoids grown in complete SC medium (left) or in medium without EGF and added TGF-β1; $n=3$ independent experiments, Paired two-tailed Wilcoxon test. **g**, Scheme and flow cytometry plot of a dissociated subcutaneous xenograft generated by inoculation of PDO7 targeted with KI67–RFP fusion and LGR5–GFP reporter cassettes. Epithelial cells (EPCAM^{+}) are shown. **h**, Quantification of Mex3a, Lgr5 and Ki67 mRNA by RT–qPCR in the indicated four PDO7 xenograft cell subpopulations; $n=5$ xenograft tumors. Paired two-tailed Student's *t*-test. **i**–**k**, UMAPs of pooled epithelial cancer cells corresponding to 37 samples from patients with CRC. Cells are colored according to levels of KI67 signature (**i**), MEX3A mRNA (**j**) or LGR5^{+} ISC signature (**k**). **l**, Identification of LGR5^{hi}/Mex3a^{hi} and LGR5^{hi}/Mex3a^{lo} population in the pooled dataset from patients with CRC. **m**, Distribution of CRC cells according to KI67 and LGR5 signatures levels in the pooled scRNA-seq CRC patient cohort. Cells are colored by Mex3a mRNA levels. **n**, Frequency of cells expressing Mex3a in the indicated cell subpopulations defined by the expression of Lgr5 and Ki67; $n=37$ tumors, Fisher's exact test. **o**,**p**, Distribution (**o**) and quantification (**p**) of Mex3a^{+}/Lgr5^{+}, Mex3a^{+}/Lgr5^{-} and Mex3a^{-}/Lgr5^{-} in cell cycle phases according to expression of cell cycle gene signatures (Methods). Data in **e** and **h** are presented as mean ± s.e.m.


a
SC medium
No EGF/ + TGFβ
p7
p19b
C31M
CC09
SP42
SP5
DA13
SP43
PDO7
p26
p18
1.0
0.8
0.6
0.4
0.2
Growth (relative to SC)
b
No EGF/+ TGF-β
SC medium
SN38
5FU
Oxaliplatin
200
150
100
50
Fold IC50 (relative to SC)
c
PDO7
SC medium
no EGF
d
MEX3A locus
Ires
iCaspase9
T2A
tdTomato
Mex3a-iCaspase9-tom (Mex3a^iCT)
Ecadherin tdTomato
e
Mex3a
tdTomato
P = 0.0015
P = 4.58 × 10−05
mRNA relative expression
Tom^hi
Tom^lo
f
SC medium
No EGF + TGF-β
10.5%
50.4%
SSC-A
P = 1.01 × 10−05
Mex3a-tdTomato
g
LGR5 locus
Stop
EGFP
MKI67 locus
Stop
TagRFP2
Neg. ctrl.
PDO7
Ki67–RFP
LGR5–GFP
h
Mex3a
Lgr5
Ki67
P = 0.0058
P = 0.0005
P = 0.0004
P = 0.007
P = 0.02
mRNA relative expression
Hi Hi
Hi Lo
Lo Hi
Lo Lo
LGR5
KI67
i
Ki67 signature
UMAP-2
UMAP-1
j
MEX3A gene
k
LGR5 signature
l
LGR5hi/MEX3Alo
LGR5hi/MEX3Ahi
Others
m
Ki67 signature
Mex3a gene
LGR5 signature
n
P = 2.2 × 10−16
Fraction
Mex3a−
Mex3a+
o
G0
G1
G2M
S
UMAP-2
UMAP-1
p
Cell cycle phase
Fraction
S
G2M
G1
G0
Mex3a
Lgr5



Extended Data Fig. 3k). Consistent with previous studies[34,35], Lgr5$^+$ colonic adenoma cells generated large clonal patches (Fig. 3j,k); however, we also observed a substantial fraction of small-size Lgr5$^+$ cell-derived clones even after 15 d (Fig. 3j and Extended Data Fig. 3k). In contrast, Mex3a$^+$ cells contributed substantially lower amount of progeny over time compared to Lgr5$^+$ cells (Fig. 3i), a difference that is also reflected in the frequency of small clones (<10 cells) generated by the two populations (Fig. 3i,j and Extended Data Fig. 3l). Taken together, these data reveal the existence of a slow proliferative Lgr5$^+$ cell population in adenomas characterized by high Mex3a expression levels.

Mex3a impedes secretory differentiation of Lgr5$^+$ tumor cells. Mex3a-null mice die around birth, which precludes analyses of Mex3a function in the adult epithelium[36]. To overcome this limitation, we generated mice bearing Mex3a-floxed alleles and crossed them with a villin-CreERT2 strain that enables efficient recombination in the entire intestinal epithelium (Fig. 4a and Methods). Mex3a deficiency in the adult intestinal epithelium did not cause an overt phenotype neither morphological alterations or obvious changes in the specification of the main intestinal cell lineages implying that the function of Mex3a is dispensable for intestinal homeostasis (Fig. 4a,b and Extended Data Fig. 4). We also obtained mice with

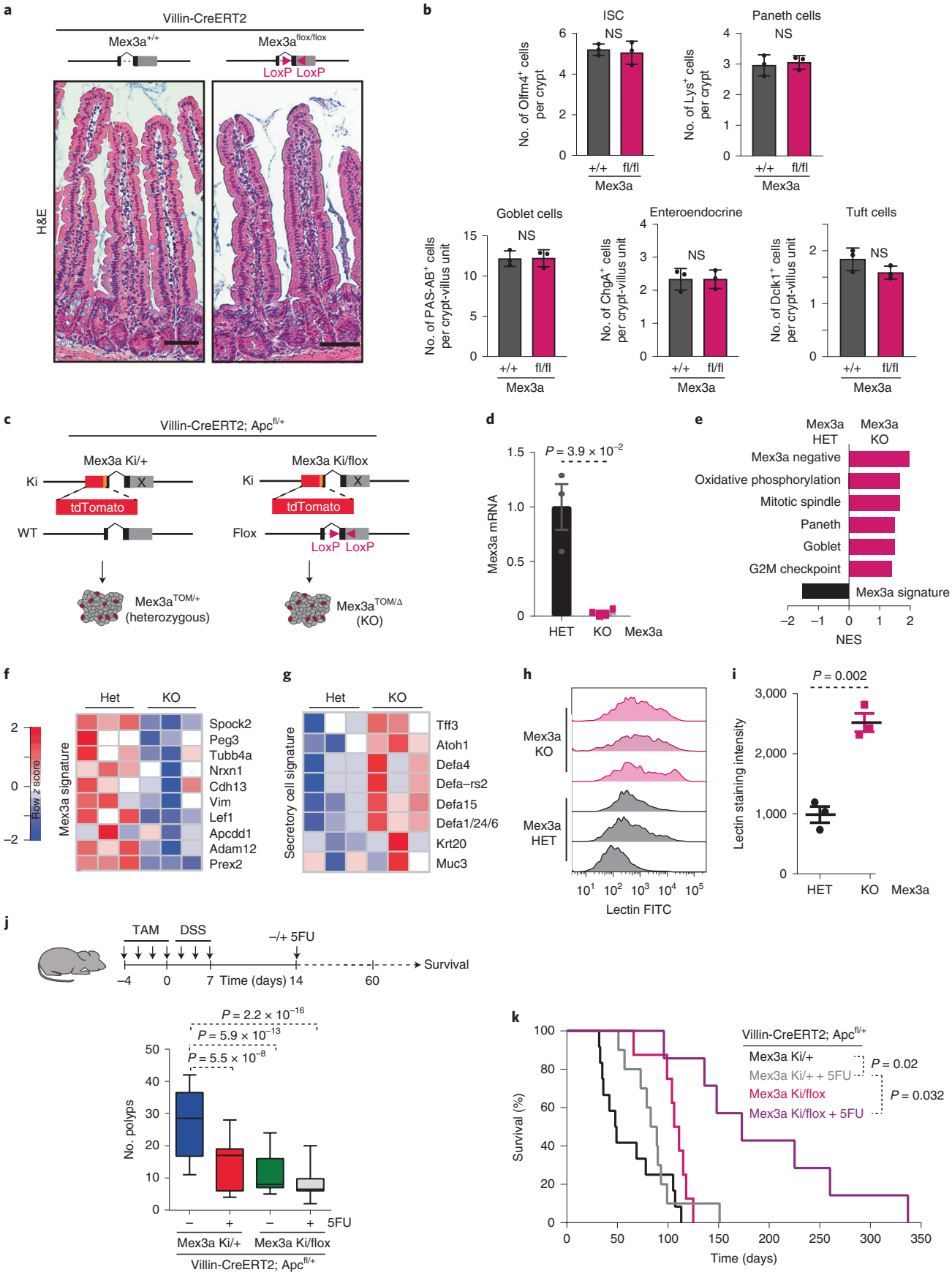
Fig. 3 | Clonal analysis of Mex3a$^+$ cells in mouse Apc-mutant adenomas. a, Representative immunohistochemistry of tdTomato in Lgr5-CreERT2; APC$^{fl/fl}$; Mex3a$^{Tom/+}$ mouse colon. Adenomas are delimited by a dashed line. Ad, adenoma; NM, normal mucosa. Scale bar, 250 µM. **b**, Representative flow cytometry plots of dissociated Mex3a$^{Tom/+}$ crypts and adenomas. Negative controls correspond to Mex3a$^{+/+}$ mice. Numbers indicate frequency of Mex3a$^+$ cells within the Lgr5-GFP$^+$ cell gate. **c**, Quantification of percentage of Mex3a$^+$ cells within Lgr5−GFP$^+$ cells in crypts compared to adenomas; $n=6$ and $n=8$ mice, respectively, two-tailed Wilcoxon test. **d**, Percentage of EdU incorporation in Mex3a$^+$/Lgr5$^+$ and Mex3a$^-$/Lgr5$^+$ adenoma cells; $n=5$ mice, two-tailed Wilcoxon test. **e**, GSEA in Mex3a$^+$ versus Mex3a$^-$ tumor cells sorted from Lgr5-GFP$^+$ adenomas. Most relevant gene sets are shown, dashed line indicates $P=0.05$ calculated with rotation gene set testing for linear models. **f**, Cre-ERT2/WPRE/bGH-polyA knock-in cassette (top). Immunohistochemistry against tdTomato 48 h after tamoxifen injection (bottom). Dashed line marks colonic adenomas. Scale bar, 200 µM. **g**, Immunofluorescence of tdTomato 24 h after tamoxifen administration, co-staining with stem cell marker Olfm4 in normal crypts (top), co-staining with BrdU in adenomas (bottom). Scale bar, 100 µM in low magnification, 50 µM in detail. **h**, Quantification of BrdU incorporation in tissue sections; $n=14$ adenomas from three independent mice; linear model with no adjustment for multiple comparisons. **i**, Area occupied by tdTomato$^+$ cells normalized per total adenoma area quantified from tissue sections. *P* values were calculated with a linear model with no adjustment for multiple comparisons; $n=366$ tumors (10 mice; day 2), $n=372$ tumors (9 mice; day 7) and $n=121$ tumors (4 mice; day 15) for Mex3a-CreERT2 mice. Total of 110 tumors (5 mice, day 2), 95 tumors (5 mice, day 7) and 126 tumors (5 mice; day 15) for Lgr5-CreERT2 tracing. **j**, Size distribution of clones generated by Lgr5-CreERT2 and Mex3a-CreERT2 cassettes in **i**. *P* values were calculated with a linear mixed model with Box-Cox transformation and mouse/tumor as random effects and no adjustment for multiple comparisons. **k**, Representative immunohistochemistry against reporter gene tdTomato at d2 and d15 after tamoxifen administration in Lgr5-CreERT2 and Mex3a-CreERT2. Scale bars, 100 µM. Data in **c**, **d**, **h**, **i** and **j** are presented as mean ± s.e.m.

a Mex3a-floxed allele combined with a Mex3a-tdTomato knock-in allele. In the latter, the tdTomato cassette disrupts the Mex3a open reading frame, which results in a null allele (Fig. 4c). This configuration enabled tracking cells that switched on Mex3a promoter in a background of Mex3a loss of function (Fig. 4c). Adenomas arising in villin-creERT2; Apc$^{fl/+}$; Mex3a$^{fl/TOM}$ mice lacked Mex3a expression (Fig. 4d), downregulated signature genes of Mex3a$^+$ adenoma cells (Fig. 4e,f) and showed upregulation of secretory cell-specific genes (Fig. 4e,g). GSEA confirmed enrichment of signature genes of goblet and Paneth cells in Mex3a$^-$Tom$^{+/KO}$ cells (Fig. 4g). Accordingly, Mex3a-deficient cells isolated from adenoma organoids exhibited increased UEA-1 lectin surface staining, a marker of glycans commonly used to identify secretory cells in the gut (Fig. 4h,i). Therefore, Mex3a not only marks slow proliferative Lgr5$^+$ stem cell-like tumor cells but it is also necessary to sustain their undifferentiated state. Villin-CreERT2; Apc$^{fl/+}$; Mex3a$^{fl/TOM}$ exhibited reduced tumor burden in the gastrointestinal tract compared to control villin-CreERT2; Apc$^{fl/+}$; Mex3a$^{+/TOM}$ mice (Figs. 4j) and 5FU treatment substantially extended their survival (Fig. 4k).

Mex3a$^+$ DTP cells regenerate CRC organoids after therapy. To explore the contribution of Mex3a$^+$ cells to chemotherapy resistance in advanced CRC, we derived Apc$^{Min/+}$; Mex3a-CreERT2; Rosa26-tdTomato colon adenoma organoids and, by means of CRISPR/Cas9 editing, introduced loss-of-function mutations in P53 (P) combined with either oncogenic mutations in Kras (G12D; K) or loss-of-function Smad4 (S) alleles (Fig. 5a and Extended Data Fig. 5a). These APS and AKP MTOs enabled the study of Mex3a$^+$ cells in advanced CRC through lineage-tracing analysis. A tdTomato-labeled population emerged in triple-mutant MTOs after 4-OH-tamoxifen addition (Fig. 5b and Extended Data Fig. 5b,c). After 16 h, the number of tdTomato$^+$ cells in APS and AKP MTOs cultured in complete stem cell medium was around 10% (Fig. 5b and Extended Data Fig. 5b), which is consistent with our previous observations using Mex3a-tdTomato knock-in MTOs (Fig. 1i). AKP and APS MTO cultures switched to suboptimal niche factor conditions after OH-tamoxifen treatment conserved similar proportions of tdTomato-labeled cells implying that both pre-existing Mex3a$^+$ and Mex3a$^-$ cells adapted similarly to these growth-limiting conditions (Extended Data Fig. 5d). We also found that the proportion of the Mex3a tdTomato-labeled progeny in control cultures did not change substantially over time (Fig. 5c,d); however, levels of Mex3a measured by quantitative PCR with reverse transcription (RT–qPCR) declined over time in the tdTomato$^+$ progeny whereas expression of the proliferation marker AurkB increased suggesting that Mex3a$^+$ cells give rise to proliferating Mex3a$^-$ daughter cells (Extended Data Fig. 5e).

Treatment with chemotherapy (FOLFIRI or FOLFOX regimes) caused extensive cell death and organoid shrinkage (Fig. 5c, Extended Data Fig. 5f and Supplementary Videos 1 and 2). After 20 d, most organoids seemed necrotic (arrows in Fig. 5c). Yet, we observed the presence of some MTOs of small size, which according to their morphology under a bright-field microscope remained viable. Invariably, these MTOs contained increased proportions of

Fig. 4 | Analysis of Mex3a deficiency in Apc-mutant mice. a, Schematic representation of the generation of an intestinal specific Mex3a knockout (KO) and representative images of representative intestinal crypts (**a**) and quantification (**b**) of enteroendocrine- (ChgA), Tuft- (Dclk1), Goblet- (Periodic acid-Schiff–Alcian blue (PAS–AB)), intestinal stem- (Olfm4) and Paneth- (lysozyme) cells; $n=3$ independent mice per genotype. NS, not significant. Generalized mixed-effects linear model with Poisson distribution with no adjustment for multiple comparisons. Scale bar, 100 µm. H&E, hematoxylin and eosin. **c**, Genotypes of mice bearing Mex3a-tdTomato knock-in/$^+$ (HET; Mex3a heterozygous) and Mex3a-tdTomato knock-in/floxed alleles (KO; Mex3a knock-out). **d**, RT–qPCR analysis of MEX3A mRNA in Mex3a HET and KO adenomas; $n=3$ adenomas (HET) and 4 (KO), Unpaired two-tailed *t*-test. **e**, Relevant GSEAs obtained after transcriptomic analysis of Tom/$^+$ (HET) and Tom/flox (KO) in vivo adenomas sorted by expression of the tdTomato reporter gene. Selected GSEAs are ranked by their normalized enrichment score (NES). **f**. Expression levels according to microarray data of Mex3a signature genes in tdTomato$^+$ cells isolated by FACS from Mex3a HET and KO adenomas ($n=3$ independent colonic tumors). **g**, Expression of secretory cell marker genes in tdTomato$^+$ cells isolated by FACS from Mex3a HET and KO adenomas ($n=3$ independent colonic tumors). **h**, Flow cytometry histogram of tdTomato$^+$ cells in Tom/$^+$ (HET) and Tom/floxed (KO) for lectin. **i**, Quantification of lectin staining in Tom$^+$ cells in adenoma organoids derived from HET and KO adenomas; $n=3$ independent organoids; unpaired two-tailed Wilcoxon test. **j**, Timeline of treatments (top). Quantification of number of colon adenomas per mice in treated and untreated Mex3a HET or KO animals (bottom); $n=8$ Ki/$^+$ mice, 7 for Ki/$^+$ treated with 5FU, 8 for Ki/flox and 10 for Ki/flox treated with 5FU. Boxes represent the first and third quartiles. The line within the box indicates the median. Whiskers indicate maximum and minimum values. *P* value was calculated with a generalized linear model with Poisson distribution, with no adjustment for multiple comparisons. TAM, tamoxifen. **k**, Kaplan–Meier survival plot of mice with the indicated genotypes and treatments; $n=14$ Ki/$^+$ mice, 10 for Ki/$^+$ treated with 5FU, 8 for Ki/flox and 7 for Ki/flox treated with 5FU. *P* value is calculated using a two-tailed Wilcoxon test. Data in **b**, **d** and **i** are presented as mean ± s.e.m.


a
Villin-CreERT2
Mex3a+/+
Mex3aflox/flox
LoxP LoxP
H&E
b
ISC
Paneth cells
Goblet cells
Enteroendocrine
Tuft cells
NS
No. of Olfm4+ cells per crypt
No. of Lys+ cells per crypt
No. of PAS-AB+ cells per crypt-villus unit
No. of ChgA+ cells per crypt-villus unit
No. of Dclk1+ cells per crypt-villus unit
+/+
fl/fl
Mex3a
c
Villin-CreERT2; Apcfl/+
Mex3a Ki/+
Mex3a Ki/flox
Ki
X
tdTomato
WT
Flox
LoxP LoxP
Mex3aTOM/+ (heterozygous)
Mex3aTOM/Δ (KO)
d
P = 3.9 × 10−2
Mex3a mRNA
HET
KO
Mex3a
e
Mex3a HET
Mex3a KO
Mex3a negative
Oxidative phosphorylation
Mitotic spindle
Paneth
Goblet
G2M checkpoint
Mex3a signature
NES
f
Het
KO
Row z score
Mex3a signature
Spock2
Peg3
Tubb4a
Nrxn1
Cdh13
Vim
Lef1
Apcdd1
Adam12
Prex2
g
Secretory cell signature
Tff3
Atoh1
Defa4
Defa–rs2
Defa15
Defa1/24/6
Krt20
Muc3
h
Mex3a KO
Mex3a HET
Lectin FITC
i
P = 0.002
Lectin staining intensity
j
TAM
DSS
–/+ 5FU
Survival
Time (days)
P = 2.2 × 10−16
P = 5.9 × 10−13
P = 5.5 × 10−8
No. polyps
5FU
Mex3a Ki/+
Mex3a Ki/flox
Villin-CreERT2; Apcfl/+
k
Survival (%)
Time (days)
Villin-CreERT2; Apcfl/+
Mex3a Ki/+
Mex3a Ki/+ + 5FU
Mex3a Ki/flox
Mex3a Ki/flox + 5FU
P = 0.02
P = 0.032

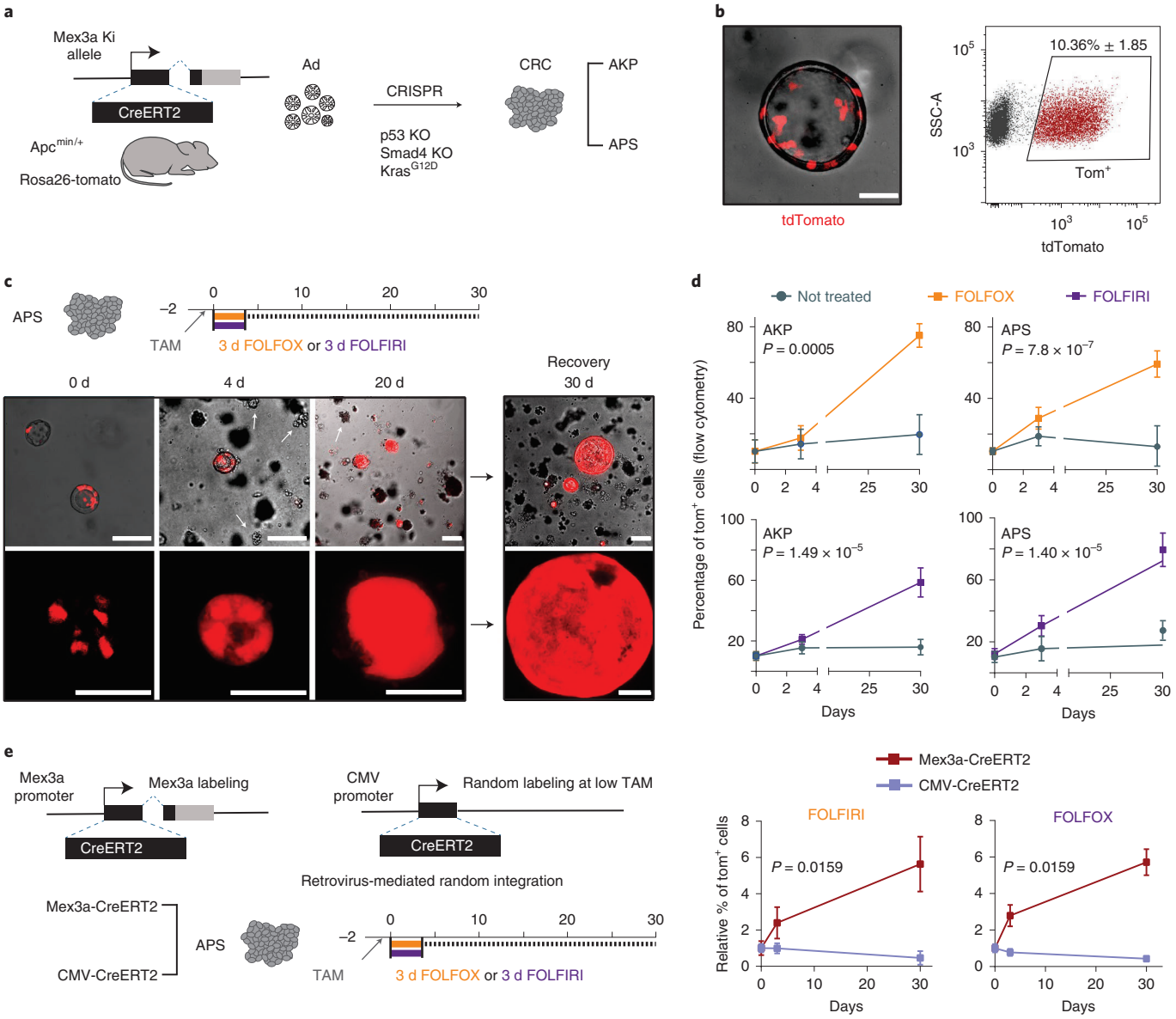


Fig. 5 | Lineage-tracing analysis of Mex3a⁺ cells in AKP and APS MTOs. a, Schematic representation of CRISPR–Cas9-based generation of advanced colorectal cancer organoids from Mex3a-CreERT2 mice. **b**, Representative image and flow cytometry plot of an APS organoid 16 h after addition of tamoxifen. Scale bar, 50 µM. Values indicate mean ± s.e.m. of percentages of tdTomato⁺ cells; $n = 3$ independent experiments. **c**, Recombination in Mex3a⁺ cells present in APS organoids was induced with tamoxifen 16 h before addition of chemotherapy. Organoids were treated during 3 d with FOLFOX or FOLFIRI and then allowed to recover for 30 d. Representative pictures of organoids at different time points are shown. White arrows indicate necrotic organoids. Scale bar, 100 µm for low magnification, 50 µm for detail. **d**, Flow cytometry quantification of tdTomato⁺ cells in MTOs treated with FOLFIRI and FOLFOX. Frequency of populations referred to the number of viable cells (4,6-diamidino-2-phenylindole (DAPI) negative). For APS, $n = 4$ independent experiments for FOLFOX and FOLFIRI, respectively; for AKP, $n = 2$ and $n = 7$; P value was calculated by a mixed-effects linear model, with no adjustment for multiple comparisons. **e**, Scheme of the experimental design. We aimed to trace a random population of tdTomato⁺ cells in APS organoids with a CMV-CreERT2. At low doses of tamoxifen we achieved recombination of equivalent proportion of cells from organoids carrying Mex3a-CreERT2 and CMV-CreERT2 alleles. Recombination was induced with 4-OH TAM for 16 h following FOLFIRI or FOLFOX treatment. The fraction of tdTomato⁺ cells over time was quantified by flow cytometry. A linear model was used to estimate P values, no adjustment was made for multiple comparisons; $n = 4$ independent organoids for random-Cre and $n = 5$ for Mex3a-CreERT. Data in **d** and **e** are presented as mean ± s.e.m.

tdTomato⁺ cells (Fig. 5c,d, Extended Data Fig. 5f and Supplementary Videos 1 and 2). Approximately 20 d after treatment cessation, some organoids resumed growth. In the resulting cultures, the majority of viable cells were tdTomato⁺ (60–80% by flow cytometry; Fig. 5d) implying that they originated from chemotherapy-persister Mex3a⁺ cells. A and AP MTOs also regenerated from Mex3a⁺ cells after chemotherapy treatment (Extended Data Fig. 5g–k). As a control for lineage-tracing experiments, we used a retroviral vector to introduce a CreERT2 cassette downstream of the CMV promoter that drives constitutive expression irrespective of the cell type. We titrated the tamoxifen dose necessary to recombine approximately the same number of cells as with the Mex3a-CreERT2 cassette (5–10%) and tracked the fate of labeled cells after treatment. These experiments revealed that, as opposed to Mex3a⁺ cells, the progeny of a

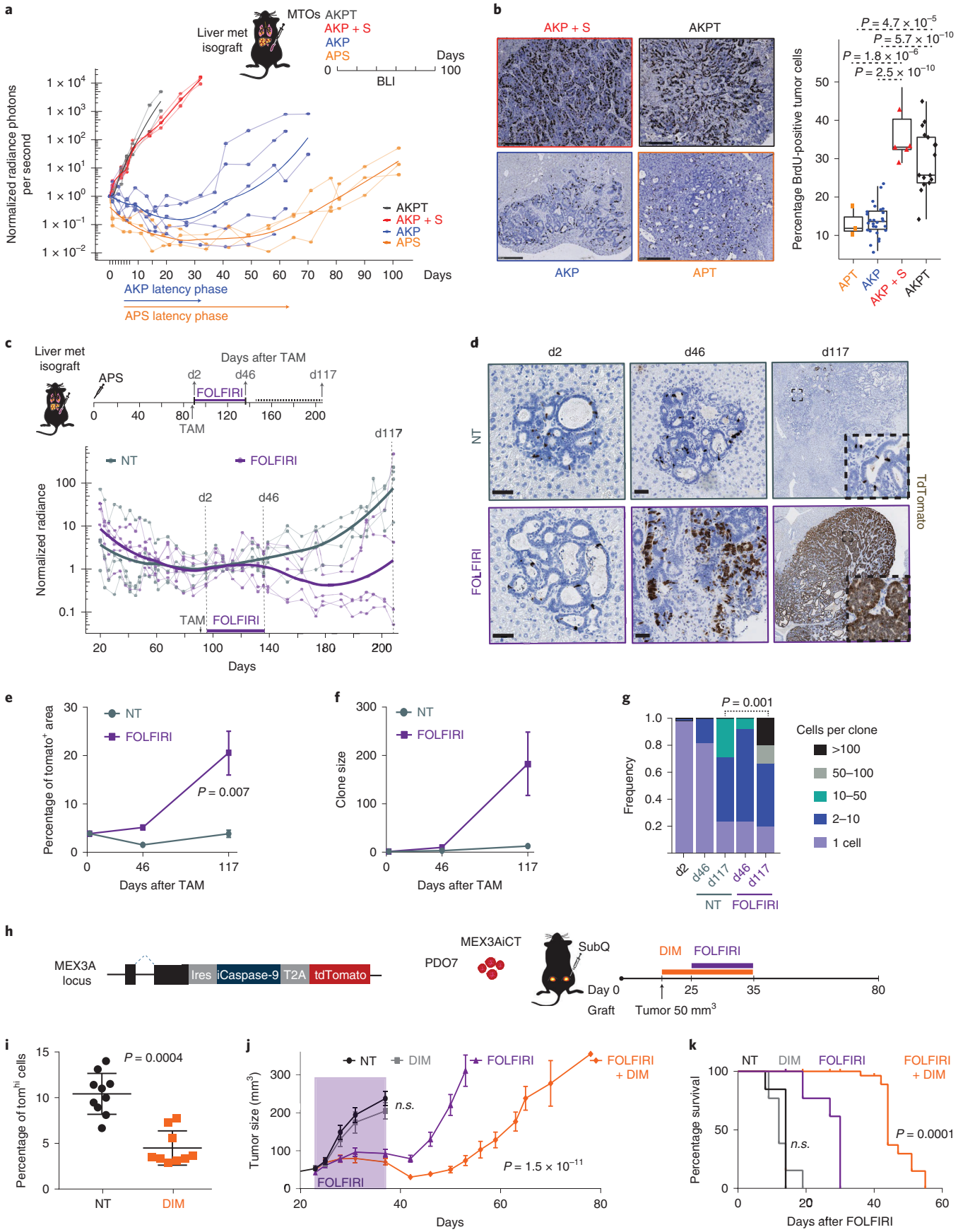

a
Liver met isograft
MTOs AKPT
AKP + S
AKP
APS
Days
0
100
BLI
Normalized radiance photons per second
AKPT
AKP + S
AKP
APS
AKP latency phase
APS latency phase
b
AKP + S
AKPT
AKP
APT
Percentage BrdU-positive tumor cells
P = 4.7 × 10−5
P = 5.7 × 10−10
P = 1.8 × 10−6
P = 2.5 × 10−10
c
Liver met isograft
APS
Days after TAM
d2
d46
d117
FOLFIRI
TAM
NT
FOLFIRI
Normalized radiance
Days
d
d2
d46
d117
NT
FOLFIRI
TdTomato
e
Percentage of tomato+ area
NT
FOLFIRI
P = 0.007
Days after TAM
f
Clone size
NT
FOLFIRI
Days after TAM
g
P = 0.001
Frequency
Cells per clone
>100
50–100
10–50
2–10
1 cell
d2
d46
d117
d46
d117
NT
FOLFIRI
h
MEX3A locus
Ires
iCaspase-9
T2A
tdTomato
MEX3AiCT
PDO7
SubQ
DIM
FOLFIRI
Day 0
Graft
Tumor 50 mm3
i
Percentage of tomhi cells
P = 0.0004
NT
DIM
j
NT
DIM
FOLFIRI
FOLFIRI + DIM
Tumor size (mm3)
n.s.
FOLFIRI
P = 1.5 × 10−11
Days
k
NT
DIM
FOLFIRI
FOLFIRI + DIM
Percentage survival
n.s.
P = 0.0001
Days after FOLFIRI

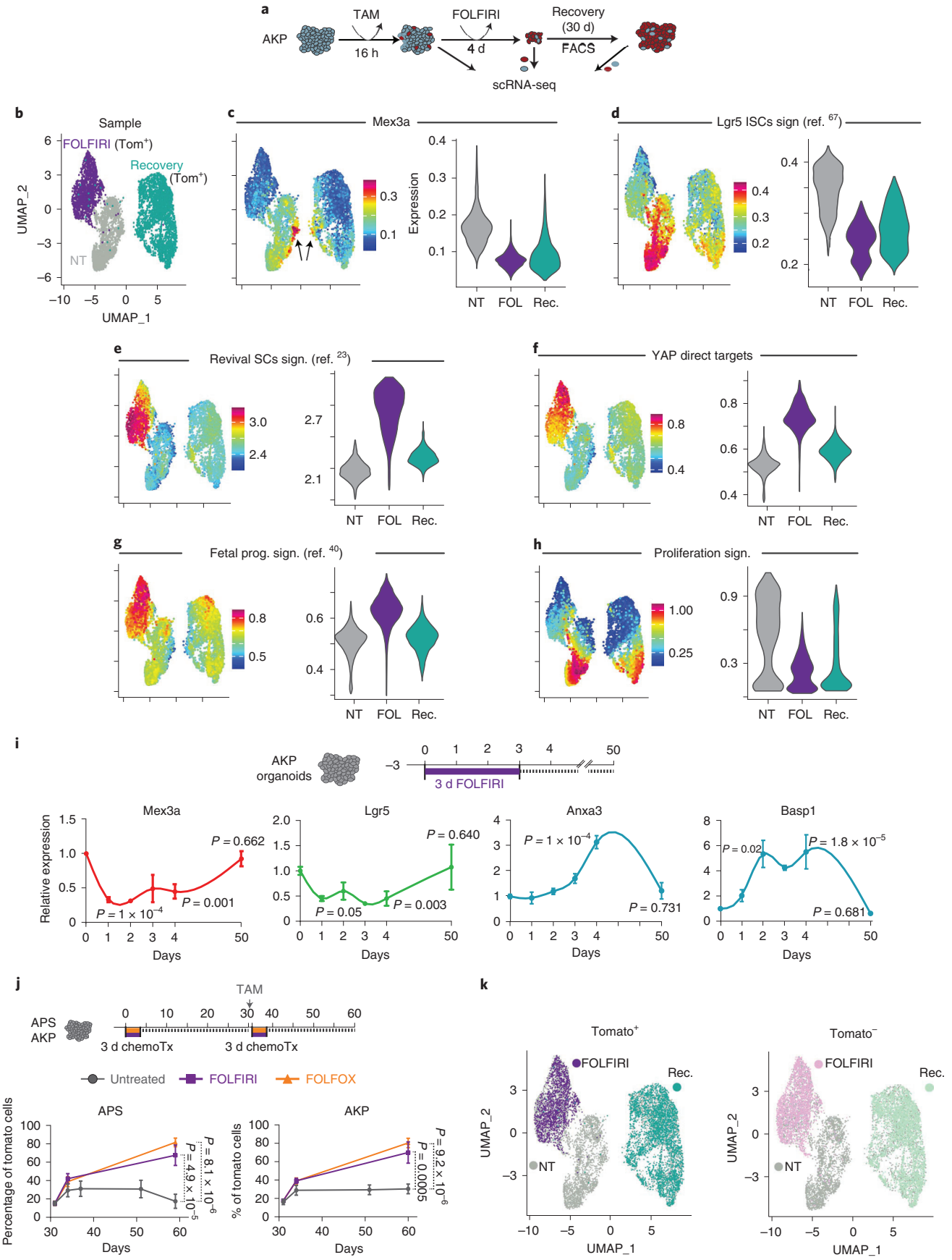
Fig. 6 | Lineage-tracing analysis of Mex3a$^+$ cells in liver metastasis. a, Graph shows longitudinal intravital bioluminescence imaging (BLI) measurements (photons per second), normalized to the day of MTO injection. Points and lines represent individual mice, trend lines (bold) were generated using a LOESS model. For AKPT, $n=2$; APS and AKP + S, $n=3$ and AKP, $n=5$ mice. **b**, Representative images and quantification of BrdU immunohistochemistry in liver metastasis of **a**. A mixed-effects linear model was fitted to percentages after a Box-Cox transformation, with no adjustment for multiple comparisons; $n=3$, 27, 5 and 16 mice for APT, AKP, AKP + S and AKPT, respectively. Boxes represent the first and third quartiles. The line indicates the median. Whiskers indicate maximum and minimum values. Scale bar, 250 µm. **c**, BLI quantification of the liver upon intrasplenic injection of Mex3a-CreERT2 APS MTOs in C57BL/6J mice. Mice were administered with a single dose of tamoxifen before initiation of chemotherapy treatment. NT, not treated. **d**, Representative images of Mex3a-derived clones in metastasis. Scale bar, 50 µM. **e**–**g**, Percentage of tdTomato$^+$ from total tumor epithelial area (**e**), average size (**f**) and size distribution (**g**) of Mex3a-derived clones in both treated and untreated conditions; $n=192$ individual metastasis (from three mice), 199 (3) and 141 (3) for non-treated conditions at d2, d46 and d117 after TAM induction, respectively; and $n=205$ (3) and 45 (3) for FOLFIRI-treated metastasis at d46 and 117 after TAM. *P* value of **e** was calculated using a linear mixed model with no adjustment for multiple comparisons. *P* value in **g** was calculated using a Pearson's chi-squared test. **h**, Mex3a-inducible caspase-9-tdTomato allele engineered in PDO7. **i**, Quantification of MEX3A^{hi} cells (defined as the top 10% of the tdTomato population); $n=10$ xenografts in NT, $n=9$ in Dim-treated condition; paired two-tailed Wilcoxon test. **j**, PDO7 subcutaneous xenograft growth; $n=13$ xenografts per condition; mixed linear model with no adjustment for multiple comparisons. **k**, Kaplan–Meier analysis of the experiment in **j**. Mice were killed when xenograft size reached 300 mm^3. A two-tailed Wilcoxon test was used to compare differences in the average time to event. Data in **e**, **f**, **i** and **j** are presented as mean ± s.e.m.

randomly recombined subset of tumor cells by the CMV-CreERT2 driver declined over time after chemotherapy (Fig. 5e). This result further highlights the unique properties of the pre-existing Mex3a$^+$ cell population to resist a lethal chemotherapy dose and regenerate the MTO culture after treatment. It also rules out the possibility that any tumor cell within a CRC is able to adopt a DTP cell state as recently suggested by a study that used a lentiviral barcoding approach[6].

Lineage-tracing of Mex3a$^+$ DTP cells in metastasis. We have previously shown that intrasplenic inoculation of quadruple-mutant (AKPT) MTOs in c57BL/6 mice produces rapidly growing liver metastases that kill the host in only 3–4 weeks[21]. Triple-mutant MTOs (APS and AKP) also colonized the liver efficiently yet they generated micrometastases that proliferated slower (Fig. 6a,b). Only after a prolonged latency time, triple-mutant metastases entered an exponential growth phase shown by in vivo bioluminescence analysis (Fig. 6a). In the case of APS MTOs, the period between inoculation and the onset of rapid growth spanned 60–80 d (Fig. 6a,c). We leveraged this model of metastatic latency to investigate the contribution of Mex3a$^+$ cells to chemotherapy resistance and disease relapse. By giving a single injection of tamoxifen 88 d after APS MTO inoculation, we marked ~4% of all tumor cells in micrometastases (Fig. 6d,e). These Mex3a$^+$ cells contributed to a reduced amount of progeny in untreated mice as shown by the quantification of tdTomato$^+$ cells in tumors during exponential growth phase (day 46 after tamoxifen induction) and at experimental end points (day 117 after tamoxifen induction) (Fig. 6d,e). A continuous FOLFIRI regime during 46 d aimed at eliminating micrometastases caused a reduction of bioluminescence (Fig. 6c); however, 50% of mice relapsed and eventually succumbed to overt liver metastatic disease. Tracking of Mex3a$^+$ cells that were recombined just before chemotherapy treatment demonstrated striking clonal amplification of their progeny at experimental end points (Fig. 6d–g and Extended Data Fig. 6a–c) to the extent that some metastases were completely formed by tdTomato$^+$ cells (Extended Data Fig. 6a). At experimental end points, tdTomato$^+$ clones contained both Mex3a$^+$ and Mex3a$^-$ cells shown by in situ hybridization (Fig. 6d and Extended Data Fig. 6d). It was also evident the presence of residual tdTomato$^+$ cell clusters, which resisted therapy but did not expand further (Extended Data Fig. 6b). Therefore, Mex3a$^+$ cells present in micrometastases resist chemotherapy and, after treatment cessation, undergo a large expansion that regenerates the disease.

To analyze the role of Mex3a$^+$ cells in human CRC, we devised an in vivo cell ablation strategy. Using CRISPR/Cas9, we introduced an inducible suicide cassette composed of a FKBP–caspase-9 fusion (iCaspase-9) plus a tdTomato reporter[28,37] into the MEX3A locus of PDO7 (MEX3A-iCT; Figs. 6h–k and 2d,e) and, subsequently, inoculated this genome-edited PDO into immunodeficient mice (Fig. 6h–k). In these experiments, treatment of mice with AP20187 (dimerizer or DIM), a compound that induces dimerization of the FKBP–caspase-9 fusion, reduced by half the 10% Mex3a-tdTomato brightest cell population present in xenografts (Fig. 6i). The limited cell-killing activity of the FKBP–caspase-9 knock-in cassette is probably due to low expression levels of the Mex3a locus. Reduced numbers of MEX3A^{hi} cells did not affect xenograft growth rates implying that this population is dispensable in unperturbed tumors (Fig. 6j). A 10-d FOLFIRI regime halted tumor growth yet, 2 weeks after treatment ended, all xenografts resumed their expansion (Fig. 6j). Despite suboptimal MEX3A$^+$ cell ablation efficiency in these

Fig. 7 | Fate mapping of Mex3a$^+$ tumor cells during chemotherapy and recovery. a, Experimental design. Tamoxifen was added for 16 h before treatment to AKP organoids to induce recombination in Mex3a$^+$ cells; then FOLFIRI treatment was applied for 4 d, after which recovery (Rec.) was allowed for 30 d. Viable Tom$^+$ and Tom$^-$ cells were isolated by flow cytometry after FOLFIRI treatment (4 d) or the recovery period (30 d) and subsequently analyzed by scRNA-seq. RNA-seq of AKP before treatment were used as a reference. **b**, UMAP representation of untreated MTOs and sorted tdTomato$^+$ cells 4 d after treatment and upon recovery (30 d). **c**, Representation of Mex3a expression in UMAP (left) and quantification of expression levels as violin plots. Arrow indicate Mex3a^{hi} cells. **d**–**h**, Individual cells in UMAPs (left) are colored according to average expression of LGR5 ISC signature[67] (**d**), revival stem cell signature[23] (**e**), yap direct targets (**f**) (Supplementary Table 6), (**g**) fetal progenitor signature[40] and proliferation signature (**h**) (Supplementary Table 6). Accompanying violin plots (right) next to the UAMPs depict the expression scores of the indicated gene signatures across each population: not treated (NT, in gray), 4 days after FOLFIRI treatment (FOL, in purple), after 30 days of recovery (Rec., in turquoise). Only tdTomato$^+$ cells are shown upon treatment. Signatures are detailed in Supplementary Table 6. **i**, RT-qPCR analysis of the expression of Mex3a, Lgr5 and two fetal progenitors/revival stem cell marker genes (*Anxa3* and *Basp1*) during FOLFIRI treatment and recovery; $n=4$ independent experiments. Unpaired two-tailed Student's *t*-test. **j**. Experimental design to assess contribution of Mex3a$^+$ cells upon re-exposure to chemotherapy in APS and AKP MTO cultures regenerated from a first round of treatment (Tx). Percentage of tdTomato$^+$ cells during APS (left) or AKP (right) organoid regeneration after chemotherapy assessed by flow cytometry; $n=6$ independent samples for APS and AKP organoids; *P* value was calculated by a linear model, with no adjustment for multiple comparisons. **k**, UMAP representation showing overlapping distributions of tdTomato$^+$ and tdTomato$^-$ cells in AKP MTOs from the experiment shown in **b**–**h**. Error bars in **i** and **j** indicate mean ± s.e.m.


a
AKP
TAM
16 h
FOLFIRI
4 d
Recovery
(30 d)
FACS
scRNA-seq
b
Sample
FOLFIRI (Tom+)
Recovery
(Tom+)
NT
UMAP_2
UMAP_1
c
Mex3a
Expression
NT FOL Rec.
d
Lgr5 ISCs sign (ref. 67)
NT FOL Rec.
e
Revival SCs sign. (ref. 23)
NT FOL Rec.
f
YAP direct targets
NT FOL Rec.
g
Fetal prog. sign. (ref. 40)
NT FOL Rec.
h
Proliferation sign.
NT FOL Rec.
i
AKP organoids
3 d FOLFIRI
Mex3a
Lgr5
Anxa3
Basp1
Relative expression
P = 0.662
P = 0.001
P = 1 × 10−4
P = 0.640
P = 0.05
P = 0.003
P = 1 × 10−4
P = 0.731
P = 0.02
P = 1.8 × 10−5
P = 0.681
Days
j
APS
AKP
TAM
3 d chemoTx
3 d chemoTx
Untreated
FOLFIRI
FOLFOX
APS
AKP
Percentage of tomato cells
% of tomato cells
P = 8.1 × 10−6
P = 4.9 × 10−5
P = 9.2 × 10−6
P = 0.0005
Days
k
Tomato+
Tomato−
FOLFIRI
Rec.
NT
UMAP_2
UMAP_1

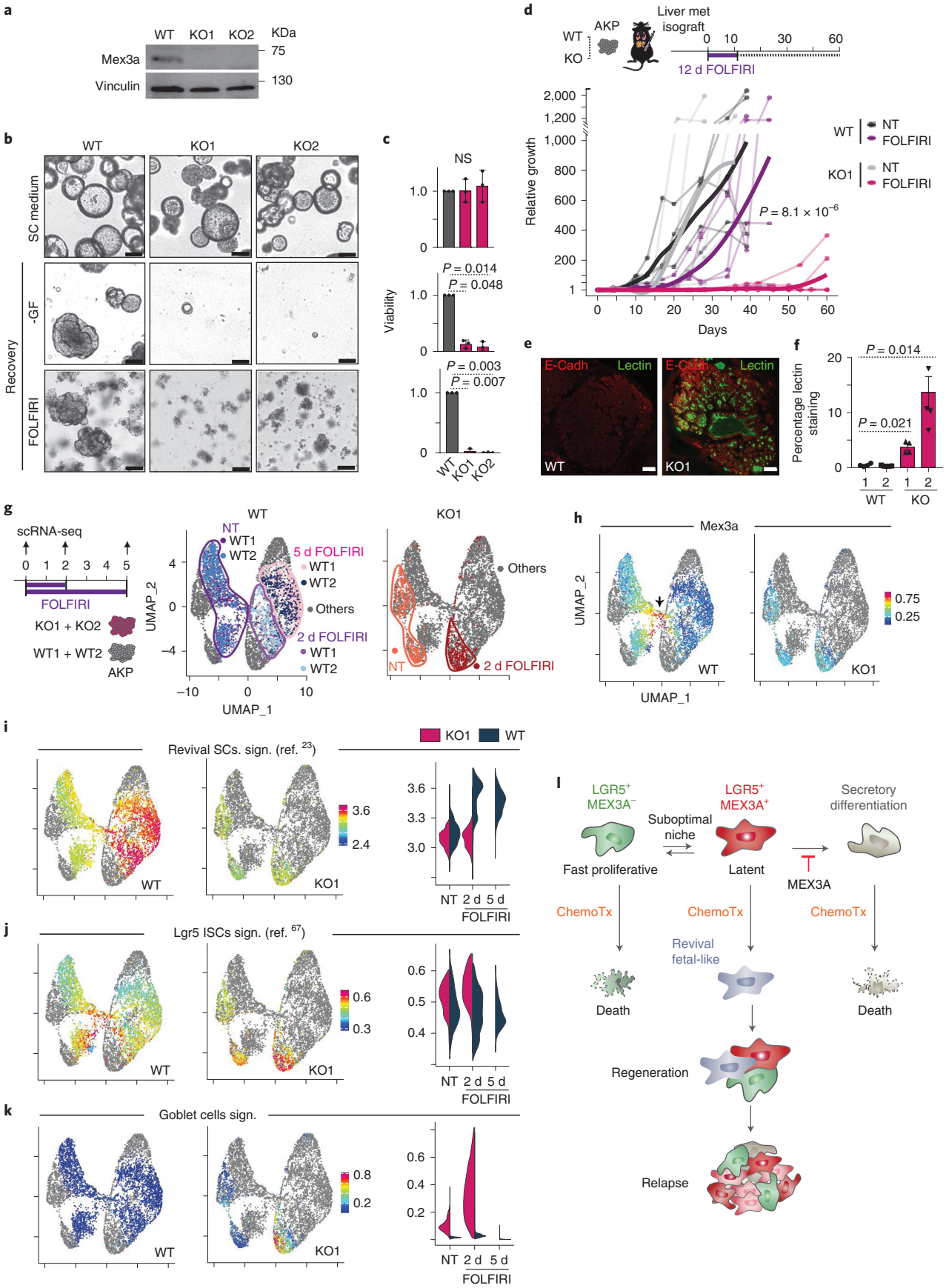
experiments, the combination of dimerizer plus FOLFIRI prolonged substantially the therapeutic effects of the treatment (Fig. 6j,k).

Conversion of Mex3a$^+$ DTP cells to fetal-like cells. We next sought to understand how Mex3a$^+$ cells adapt to chemotherapy. To this end, we combined scRNA-seq and lineage-tracing analysis as described in Fig. 7a,b. We identified discrete subsets of Mex3a-expressing cells in untreated AKP MTOs (arrow Fig. 7c). Four days after FOLFIRI treatment, we sorted viable tdTomato$^+$ cells that survived the treatment and sequenced their transcriptomes. At this early time point, the residual DTP cell population silenced Mex3a expression (Fig. 7c). The global Lgr5$^+$ ISC gene program, including bonafide ISC marker genes such as *Lgr5* and *Smoc2*, was also downregulated (Fig. 7d and Extended Data Fig. 7a,b). Instead, there was a remarkable upregulation of the transcriptional program of revival stem cells (Fig. 7e and Extended Data Fig. 7c). Revival stem cells are specified by activation of the transcriptional regulator YAP and mediate the regeneration of the intestinal epithelium after radiation[23]. They share features with the fetal-like developmental program expressed in the crypts after injury[24,38–40]. Indeed, we found increased levels of multiple markers of fetal intestinal progenitors and YAP target genes in the chemotherapy-resistant residual population such as *Anxa1*, *Ctgf* (Ccn2) or *Basp* (Fig. 7f,g and Extended Data Fig. 7d). We validated changes in the expression of marker genes for these populations by RT–qPCR (Fig. 7i). The Mex3a-to-revival/fetal conversion was consistent in both AKP (Fig. 7a–i and Extended Data Fig. 7) and APS MTOs (Extended Data Fig. 8) implying a common mechanism of adaptation after chemotherapy irrespectively of the genotype. tdTomato$^+$ cells present 30–50 d after FOLFIRI treatment in fully regenerated MTO cultures downregulated the revival/fetal stem cell program, regained expression of WNT target genes and contained again a small proportion of Mex3a$^+$ cells (Fig. 7a–i and Extended Data Fig. 7). In an independent experiment, we tested whether these newly formed Mex3a$^+$ cells retained a drug-persister phenotype (Fig. 7j). Indeed, after a second round of FOLFIRI, 60–80% of MTOs cells originated from Mex3a$^+$ cells demonstrating reversibility of the DTP cell phenotype (Fig. 7j). Finally, in these experiments we sorted and sequenced equivalent numbers of tdTomato$^+$ and tdTomato- cells, which enable the analysis of the minority of cells that originated from the non-Mex3a$^+$ cell population after chemotherapy. We found that the cell states present in tdTomato$^-$ cells completely overlap with that of tdTomato$^+$ cells in both AKP (Fig. 7k and Extended Data Fig. 7f–m) and APS MTOs (Extended Data Fig. 8). Thus, either non-Mex3a$^+$ drug-tolerant cells adopt the same transcriptional states than Mex3a$^+$ cells during treatment or, more likely, tdTomato$^-$ cells reflect failure of the knock-in CreERT2 cassette to recombine a proportion of Mex3a$^+$ cells.

Mex3a-deficient CRC cells are sensitive to chemotherapy. To obtain definitive evidence of the role of Mex3a$^+$ cells during relapse after chemotherapy, we generated Mex3a knockout AKP MTOs using CRISPR/Cas9 (Fig. 8a). Lack of Mex3a did not affect growth kinetics in complete stem cell conditions yet these MTOs were unable to survive in no-EGF/+TGF-β medium (Fig. 8b,c). Mex3a-deficient MTOs also showed susceptibility to chemotherapy and were unable to regenerate the culture after FOLFIRI treatment (Fig. 8b,c). Two independent Mex3a knockout AKP MTOs showed similar behaviors in these conditions (Fig. 8b,c). To assess chemotherapy sensitivity in vivo, we inoculated wild-type and Mex3a KO1 AKP MTOs into the liver of c57BL/6 mice (Fig. 8d). We measured liver bioluminescence and initiated a chemotherapy regime at the onset of exponential growth phase. Echoing our findings in adenomas, Mex3a KO metastases were enriched in secretory-like cells shown by UEA-1 lectin staining (Fig. 8e). FOLFIRI treatment delayed growth of wild-type AKP tumors but all mice eventually succumbed to metastatic disease (Fig. 8d). In contrast, all mice bearing Mex3a-deficient metastases developed robust responses to chemotherapy (Fig. 8d); three showed no signs of relapse over the follow-up period (2 months) and in two other mice, metastatic disease relapsed with delayed kinetics (Fig. 8d).

Prompted by these results, we applied scRNA-seq to investigate the fate of Mex3a-deficient cells during FOLFIRI treatment. We characterized the response of two independent wild-type and Mex3a knockout AKP MTOs at 2 d and 5 d after treatment (data for KO1 in Fig. 8g–k and Extended Data Fig. 9 and for KO2 in Extended Data Fig. 10). Mex3a KO1 and KO2 MTOs clustered together in the Uniform Manifold Approximation and Projection (UMAP) but occupied a distinct space than their wild-type counterparts implying the existence of different cell states in the two genotypes (Fig. 8g and Extended Data Figs. 9a and 10a). Consistent with our previous findings, wild-type AKP CRC cells adopted a regenerative revival/fetal-like stem cell state after chemotherapy shown by global upregulation of gene signatures and well-known marker genes for these populations (Fig. 8i and Extended Data Figs. 9b,c and 10c,d). A cell subset also showed elevated expression of YAP targets genes (Extended Data Figs. 9d,e and 10e,f). The upregulation of revival/fetal-like markers genes started 2 d after chemotherapy and was particularly evident 5 d after treatment. As we demonstrated previously through lineage-tracing analysis (Fig. 7a–i), this population originated from chemoresistant Mex3a$^+$ cells. Indeed, a string of Mex3a$^+$ cells linked untreated and treated wild-type cell clusters in the UMAP (arrow in Fig. 8h). In contrast, the residual Mex3a-deficient population that we recovered 2 d after chemotherapy in Mex3a KO MTOs showed residual expression of the regenerative/fetal-like program and instead upregulated the Lgr5$^+$ ISC gene signature and

Fig. 8 | Effects of Mex3a deficiency on AKP MTOs and liver metastasis. a, Western blot analysis of MEX3A and in wild-type (WT) and two independent CRISPR-generated AKP KO MTOs. Vinculin is used as a loading control. **b**, Representative images of Mex3a WT and KO AKP MTOs upon recovery of no-EGF/+TGF-β condition or FOLFIRI treatment. Scale bar, 50 µm. **c**, Relative viability (ATP measurements) of AKP MTOs in the experiment shown in **b**. MTOs always in complete SC medium (top), MTOs recovered from no-EGF/+TGF-β condition (middle) and MTOs recovered from FOLFIRI treatment (bottom); $n=3$ independent experiments; *P* value calculated using a linear model with no adjustment for multiple comparisons. **d**, WT or Mex3a KO1 AKP MTOs were inoculated through the spleen of C56BL/6 mice. Mice were randomized in two arms (untreated or treatment with FOLFIRI as indicated). Graph shows liver bioluminescence over time as a surrogate of metastatic growth. Individual mice (thin lines) and average growth (thick lines) in each treatment group are indicated; $n=5$ mice per treatment. A mixed-effects linear model was applied to calculate *P* values between WT FOLFIRI and KO1 FOLFIRI at day 39, with no adjustment for multiple comparisons. **e**, Representative UEA-1 lectin staining of liver metastases generated by WT and Mex3a$^-$ KO1 MTOs. Scale bar, 200 µm. **f**, Quantification of UEA-1 lectin staining in liver metastases; $n=5$ mice for WT2 and 4 mice for WT1, KO1 and KO2. An average of 30 metastatic nodules were quantified per mice; unpaired two-tailed Wilcoxon test. **g**, Cell distribution in UMAPs of two independent WT (WT1 and WT2) and Mex3a KO1 AKP MTO cultures in untreated conditions and treated with FOLFIRI for 2 d or 5 d. Note that no viable cells were recovered in KO1 at 5 d after treatment. **h**, Mex3a comparing WT1+2 versus KO1 AKP MTOs in UMAPs. Arrow indicates Mex3a^{hi} cells. **i**–**k**, Individual cells in UMAPs (left) are colored according to average expression of the revival stem cell signature[23] (**i**), the LGR5 ISC signature[67] (**j**), and goblet cell signature (**k**) (Supplementary Table 6). Accompanying violin plots (right) next to the UMAPs depict the expression scores of the indicated gene signatures in WT cells (dark blue) or Mex3a KO1 cells (in fuchsia) in each population: not treated (NT), 2 days after FOLFIRI treatment (2 d), and 5 days after FOLFIRI treatment (5 d). Signatures are detailed in Supplementary Table 6. **l**, Graphical summary of the key findings of the study. Data in **c** and **f** are presented as mean ± s.e.m.


a
WT KO1 KO2 KDa
Mex3a
Vinculin
75
130
b
WT KO1 KO2
SC medium
Recovery
-GF
FOLFIRI
c
NS
P = 0.014
P = 0.048
P = 0.003
P = 0.007
Viability
WT KO1 KO2
d
Liver met isograft
AKP
WT
KO
0 10 30 60
12 d FOLFIRI
Relative growth
P = 8.1 × 10−6
WT NT FOLFIRI
KO1 NT FOLFIRI
Days
e
E-Cadh Lectin
WT
KO1
f
P = 0.014
P = 0.021
Percentage lectin staining
1 2 1 2
WT KO
g
scRNA-seq
0 1 2 3 4 5
FOLFIRI
KO1 + KO2
WT1 + WT2
AKP
WT
NT
WT1
WT2
5 d FOLFIRI
Others
2 d FOLFIRI
KO1
UMAP_1
UMAP_2
h
Mex3a
0.75
0.25
i
KO1 WT
Revival SCs. sign. (ref. 23)
3.6
2.4
NT 2 d 5 d
FOLFIRI
j
Lgr5 ISCs sign. (ref. 67)
0.6
0.3
k
Goblet cells sign.
0.8
0.2
l
LGR5+ MEX3A−
LGR5+ MEX3A+
Secretory differentiation
Suboptimal niche
Fast proliferative
Latent
MEX3A
ChemoTx
Revival fetal-like
Death
Regeneration
Relapse

key ISC marker genes such as *Lgr5* and *Smoc2* (Fig. 8i,j and Extended Data Figs. 9b–g and 10c–h). Many of these Lgr5$^+$ cells adopted a goblet-like cell phenotype demonstrated by elevated expression of Atoh1 and Spdef1, two master transcription factors of this lineage and of effector genes *Muc2* and *Tff3* (Fig. 8k and Extended Data Figs. 9h,i and 10i,j). We also noticed that the proliferation signature remained elevated after chemotherapy in Mex3a-deficient MTOs compared to their wild-type counterparts (Extended Data Fig. 9j,k). All these findings were consistent in both Mex3a-deficient AKP MTOs. At 5 d after treatment, we could not isolate sufficient cells from Mex3a KO1 MTOs but residual KO2 cells expressed high levels of apoptosis-related genes at this time point (Extended Data Fig. 10k,i).

Discussion

Our data supports that not all functionally important cells contribute to CRC growth;[28,41,42] in unchallenged tumors, some cells expand continuously, whereas others are held in reserve. Slow-cycling drug-tolerant cells have been observed in multiple tumor types[4,5], including CRC[41,43,44]; however, cellular diversity represented a major challenge to define the precise identity and features of this cell population[4,5]. We identify a latent chemotherapy-tolerant cell state that results from suboptimal adaptation of LGR5$^+$ stem cell-like cells to their niche. This population emerges in CRCs bearing an incomplete set of driver mutations such as those carrying wild-type EGFR or TGF-β signaling pathways. We speculate that other constraints such as a limited nutrient availability may also induce MEX3A$^+$ latent cells. Nevertheless, the majority of CRCs are characterized by genetic alterations in only two or three driver pathways[12,16,22]. Our own analysis of a large CRC patient cohort corroborated that dual- and triple-mutant pathways are the most common genotypes in full-blown primary CRCs and metastases (Supplementary Table 5). Furthermore, recent studies support the early dissemination of these tumors[45]. Therefore, the MEX3A$^+$ cell population described herein may contribute to therapy failure in a substantial proportion of patients and may be of particular relevance for disseminated tumor cells during the metastatic latency period.

The MEX3A cell state is not hardwired as shown by the observation that MEX3A$^+$ cells isolated from tumors give rise to organoids when cultured in optimal stem cell conditions. Lineage-tracing also demonstrates that MEX3A$^+$ cells can reactivate and generate progeny in intact tumors. Our data fit in well with a model whereby LGR5$^+$ cells adopt active (MEX3A$^-$) and latent (MEX3A$^+$) states depending on occupancy of optimal or suboptimal niches respectively. Upon chemotherapy, DTP MEX3A$^+$ cells are transcriptionally reprogrammed; they switch-off genes that characterize LGR5$^+$ ISCs and adopt a state reminiscent of that of fetal epithelium and of revival stem cells. In healthy epithelium, a rare cell subset named revival stem cells are capable of resisting radiotherapy and regenerating the damaged mucosa[23]. A fetal reversion has been described in the intestinal mucosa upon injury linked to cell survival and tissue regeneration[24,38–40]. Both transcriptional programs depend on the activity of the transcription factor YAP[23,24,38–40]. We show that fetal, revival and YAP expression programs are transiently co-opted by chemoresistant Mex3a$^+$ CRC cells. It is tempting to speculate that adaptive mechanisms of resistance such as adaptive mutability[3] or epigenetic remodeling[4,5] are facilitated by the acquisition of this primitive state. In agreement with the reversible nature of the DTP cell phenotype observed in other tumor types[46,47], these programs are switched off upon regeneration and the DTP MEX3A$^+$ cell population is reconstituted. Whereas the targets of the RNA-binding protein MEX3A$^+$ remain to be determined, our data indicate that its function regulates the transition toward the fetal/revival stem cell state. MEX3A-deficient cells upregulate a secretory program and are unable to survive the treatment.

We recognize that our study has several limitations. First, organoid cultures do not reproduce the complete pathophysiology of drug responses and, in particular, the key role of the tumor stroma in modulating chemotherapy resistance[4,5]. This may explain the observation that MEX3A$^+$ DTP cells contribute to relapse after therapy more substantially in organoids than in in vivo experimental models. Second, our data predict that targeting the MEX3A$^+$ cell population may help improve the outcome of current treatments in KRAS or TGF-β pathway wild-type CRCs. Yet, we cannot rule out that particular cell populations lacking MEX3A expression may also behave as DTP cells. A more extensive characterization of the role of the MEX3A$^+$ cell population in preclinical models, such as PDXs, is warranted. Nonetheless, our findings reveal how pre-existing cell heterogeneity imposed by adaptation to different stem cell niches shape chemotherapy responses in CRCs.

Methods

This study complies with all relevant ethical regulations. The use of patient information, tissues and genomic data was performed according to the guidelines of the European Network of Research Ethics Committees, following European, national and local laws as detailed below for each case. Experiments with mice were approved by the Animal Care and Use Committee of Barcelona Science Park under protocols CEEA-PCB-14-000053, CEEA-PCB-14-000275, CEEA-PCB-14-000276, CEEA-PCB-14-000285 and CEEA-PCB-15-000075.

Mouse tumor organoids. We described elsewhere the establishment of MTOs from primary tumors arising in genetically engineered mouse models (GEMMs) with combined distinct genetic alterations[21]. In brief, triple-mutant AKP MTOs were established from CRCs arising in Lgr5-CreERT2; Apc$^{fl/fl}$; Kras$^{LSL\text{-}G12D}$; Trp53$^{fl/fl}$ mice; ATP MTOs from Lgr5-CreERT2; Apc$^{fl/fl}$; Tgfbr2$^{fl/fl}$; Trp53$^{fl/fl}$ mice and AKPT in Lgr5-CreERT2; Apc$^{fl/fl}$; Kras$^{LSL\text{-}G12D}$; Trp53$^{fl/fl}$; Tgfbr2$^{fl/fl}$ mice. AKP + Smad4 were generated by introducing a mutation in the SMAD4 locus in AKP MTOs using CRISPR/Cas9 as previously described[21]. MTOs were cultured as detailed previously[21]. Cultures were checked bimonthly for mycoplasma contamination.

MTO latency and chemotherapy resistance. Complete stem cell medium corresponds to standard medium used for CRC organoid cultures[21]. It is supplemented with 10 mM HEPES, Glutamax, B-27 without retinoid acid (all Life Technologies), 50 ng ml^{-1} recombinant human EGF (Peprotech), 100 ng ml^{-1} recombinant human NOGGIN and 1 µM galunisertib (LY2157299) in advanced DMEM/F12. For no-EGF conditions, we used the same medium without supplementing EGF. In +TGF-β conditions, TGFBR1 inhibitor galunisertib was removed from the medium and 5 ng ml^{-1} of active TGF-β1 (Peprotech, 100-21-B) was added. To test chemotherapy resistance, MTOs were cultured in complete stem cell medium or in medium lacking EGF and supplemented with TGF-β1 (5 ng ml^{-1}) for 3 d. Organoids were then dissociated with trypsin and single cells were resuspended in BME (3533-010-02, Bio-Techne) (around 100 cells per µl). The 15-µl drops containing organoid cells were seeded on 96-well flat-bottom plates (153596, Cultek). The 5FU (sigma, F6627), SN38 (MedChem, HY-13704) and oxaliplatin (MedChem, HY-17371) were added in a range of concentrations (as described previously[8]) in both stem cell medium or in absence of EGF + TGF-β1 medium. Cell viability was assayed in triplicates using CellTiter-Glo (G9682, Promega) after 4 d of drug incubation according to manufacturer's instructions. The luminescence signal was quantified in a GloMax 96 Microplate Luminometer (Promega). Dose–response curves were fitted and corresponding IC_{50} values were estimated using a four-parameter log-logistic model[48] as implemented in the drc R package[49].

Animal experiments and maintenance. Mice were maintained in a specific-pathogen-free facility with a 12-h light–dark cycle, under controlled temperature and humidity (18–23 °C and 40–60% respectively) and given ad libitum access to standard diet and water.

Genetically modified mouse strains. Lgr5–EGFP–IRES–CreERT2 (B6.129P2-Lgr5tm1(Cre/ERT2)Cle/J; stock 008875), reporter strain Rosa26-Stop-tdTomato (B6.Cg-Gt(ROSA)26Sortm14 (CAG-tdTomato)Hze/J, stock no. 007914) and Apc$^{Min/+}$ mice were obtained from the Jackson Laboratory. The intestinal epithelium specific Cre (villin-CreERT2)[50] was a kind gift of S. Robine. Conditional Apc$^{fl/fl}$ was a kind gift of C. Perret[51]. Mex3a-tdTomato reporter mice were previously generated in the laboratory[27]. All GEMMs used in this study had been backcrossed onto a C57BL/6 background at least for six generations.

Generation of Mex3a-CreERT2 knock-in and Mex3a-floxed mice. The targeting vectors, mouse embryonic stem cell (mESC) culture, blastocyst injection of selected clones and generation of chimeras were conducted in the Mouse Mutant Core Facility at IRB Barcelona following institutional guidelines. We previously described the generation of Mex3a-tdTomato-CreERT2 mice[27]. Although Tomato is expressed in these mice, the efficiency of recombination is very low[27] possibly due to poor expression of the bi-cistronic knock-in cassette. To overcome this

drawback, we generated a new knock-in mouse strain carrying the CreERT2 cassette without the tdTomato reporter. In brief, a cassette bearing a Cre-ERT2/WPRE/bGH-polyA was inserted in frame with the Mex3a start codon in exon 1 by standard homologous recombination into B6/C2 ES cells. For the Mex3a-flox allele, a gene-targeting vector was made consisting of 5,999 bp of 5′ homology and 5,054 bp of 3′ homology. An Frt flanked PGK-neo cassette was inserted into the region upstream of exon 2 CDS and loxP sites were placed at either side of the coding region of exon 2. The linearized targeting vector was electroporated into W4 mESCs.

ESC clones were expanded, confirmed by PCR and further analyzed by Southern blot to assure correct and unique genomic integration. Finally, two positive clones were selected for injection into blastocysts and generation of chimeras. Pups with the correct color were PCR tested for the presence of the Mex3a-CreErt2 cassette or the Mex3a-floxed allele. Positive offspring were then crossed with FlpO-bearing mice to remove the neomycin-resistance cassette. Mice were backcrossed to C57BL/6 background at least for six generations.

Adenoma formation in mice. $Apc^{fl/fl}$ $Lgr5^{eGFPCreERT2}$ recombination was induced by a single 8 mg kg^{-1} intraperitoneal injection of tamoxifen (Sigma, T5648) followed by treatment with 2% dextran sulfate sodium (DSS) in drinking water for 5 d consecutively to enhance the number of colonic adenomas. APC fl/+; villin-CreERT2 recombination was induced with four consecutive injections of tamoxifen (80 mg kg^{-1}), followed by 2% DSS treatment for 5 d. The general condition of animals was monitored using animal fitness, weight controls and tests to detect blood in stools throughout the experiment. When deteriorating clinical alterations were observed, the animals were killed.

In vivo lineage-tracing and clone quantification. To ensure single-cell clone recombination in adenomas, we performed the experiments administrating diluted tamoxifen (sigma) at 8 mg kg^{-1} in Mex3a-CreErt2/+; Rosa26-Stop-tdTomato/tdTomato and at 1.6 mg kg^{-1} in Lgr5-CreErt2/+; Rosa26-Stop-tdTomato/tdTomato. Animals were killed 2, 7 or 15 d after tamoxifen induction, and the samples were fixated as previously described[27].

Once samples were collected, we performed immunohistochemical staining of 3-mm histological sections using the primary antibody against tdTomato. Images for clone quantification were acquired using a NanoZoomer-2.0 (Hamamatsu). For clonal quantification, images were processed using Fiji/ImageJ by a custom-made macro. In brief, Tomato-positive objects were detected and thresholds were calculated automatically. The algorithm grouped objects that resided in close proximity (<12 µm) so that they were considered as unique clones. Clonal area was relativized to the total area (µm^2) of the tumor section. For Figs. 3i and 6e, total tomato size (regardless of the number of clones) was calculated per adenoma area, to identify the overall contribution of tomato$^+$ clones to adenoma growth. In Figs. 3j and 6f,g and Extended Data Fig. 3l, the area per clone is plotted, to show differences in the size of individual clones.

For Fig. 6e, Tomato area was relativized to total epithelial area. Epithelium was detected using an ImageJ macro that includes size of nuclei and cells as parameters to create a mask in which the area of Tomato-positive pixels was identified. To calculate number of cells, we estimated the size of single cells in histological sections and interpolated the number of cells per each clone and per section.

BrDU and EdU detection in $Lgr5^{GFP/+}$;$Mex3a^{tdTomato/+}$ adenomas. A total of 40 mg kg^{-1} EdU (Click Chemistry Tools 1149–100) was administered 16 h before the mice were killed. A minimum of 50,000 Mex3a$^+$ and Mex3a$^-$ in Lgr5–GFP$^+$ cells were isolated by FACS. EdU was detected using Click-iT EdU Alexa Fluor 647 Flow Cytometry Assay kit (C10419). BrdU (125 mg kg^{-1}, Medchemtronica, HY-15910) was detected in paraffin sections by immunohistochemistry using anti-BrdU antibody and was quantified using Qupath v.0.2.0 or a custom-made ImageJ macro. Images were visualized with NIS v.5.21.00.

Mex3a$^+$ cell lineage-tracing in MTOs. Mex3a-CreERT2 MTOs were seeded as single cells in BME drops with fully supplemented control medium. After 2 d of growth, organoids were induced for 16 h with 4-OH-tamoxifen (1 µM Sigma-Aldrich/Merck, H7904). The 4-OH-tamoxifen was thoroughly washed out with PBS and organoids were treated as indicated. FOLFOX was a mixture of 5FU (50 µg ml^{-1}, Sigma) and oxaliplatin (100 µM MedChem HY-17371). FOLFIRI was a combination of 5FU (50 µg µl^{-1}) and SN38 (100 nM, MedChem HY-13704). After 3 d of treatment, drugs were removed, followed by addition of fresh fully supplemented control medium. Organoids in control conditions were passaged every 3–4 d as clumps using 1–2 min treatment with trypsin. For analysis, organoids were dissociated to single cells with trypsin for 20 min, followed by mechanical disaggregation and Tomato$^+$ cells were analyzed by flow cytometry in a FACS Aria 2.0 (BD). Data were analyzed with FACSDiva software (BD) or FlowJo v.10. For experiments retracing persister cells (Fig. 7j), before performing lineage-tracing, organoids were first treated with FOLFIRI/FOLFOX for 3 d and allowed to recover for 30 d.

Mex3a$^+$ cell tracing in APS metastases. For all injections, male C57BL/6J mice were purchased from Janvier Laboratories at 6 weeks of age and injected at 7 weeks. Intrasplenic vein injections were used for liver colonization by the introduction of intact small organoids into the portal circulation as previously described;[21] MTOs were cultured in standard six-well plates for 4 d and resuspended in cold HBSS to dissociate remaining BME. Cells were counted and suspended in HBSS for injection, using 1–7.5 × 10^5 cells in 70 µl of HBSS per mouse. Tumor growth was inferred by emission of luciferase using IVIS Spectrum In Vivo Imaging System upon intraorbital injection of luciferin (50 µl of a 15 mg ml^{-1} solution). Quantifications were made with Living Image v.4.5.2.

APS (Apc^{mut}, $p53^m$ and $Smad4^{KO}$) Mex3a-CreERT2 organoids were inoculated in the liver of C57BL/6J. A single shot of tamoxifen (200 mg kg^{-1}) was administrated to induce recombination of Mex3a cells at clonal level. Mice were killed 2, 46 or 117 d after tamoxifen induction. Two days after recombination, three doses per week of FOLFIRI were injected intraperitoneally over the period of 46 d. The FOLFIRI cocktail contained 5FU (30 mg kg^{-1}), irinotecan (73 mg kg^{-1}) and leucovorin (90 mg kg^{-1}). Control animals were treated with vehicle following the same regimen.

Patient-derived organoids. All PDO lines have been previously described. PDO7, DA13 and CC09 were established by the Stassi laboratory as described elsewhere[52] under approval of the University of Palermo ethics committee. SP42, SP5 and SP43 were established by the Batlle laboratory as described elsewhere[53] under approval of approval of the Tumor Bank Committees of Hospital del Mar and Hospital de la Santa Creu i Sant Pau. P7, P19b, C31M, P26 and P18 were established by the Clevers laboratory and HUB Organoids™ as described elsewhere[16,30] under approval of the Biobank Research Ethics Committee of the UMC Utrecht (TCBio). In all cases, donors signed informed consent.

Genome editing of PDO7. For the generation of Mex3a-iCaspase-9-T2A-tdTomato in PDO7 we followed the same strategy as reported elsewhere[31]. In brief, guides were designed as explained above (hsMex3a_g1: CCCGTGCCCCATGCC TCCG) and donor Hs_Mex3a homology arms (HAs) flanking the knock-in insertion cassettes were synthetized by gene synthesis (Thermo Fisher) and cloned into pShuttle or pDONOR vectors. RT–qPCR was used to determine Mex3a levels in Tom^{hi} versus Tom^{lo}, analyzed with StepOne software v.2.3. Primers are listed in Supplementary Table 7.

In vivo ablation in PDO7 xenografts. For subcutaneous injections, PDO7 was cultured in standard six-well plates for 4–5 d and recovered in cold HBSS as undissociated organoids. To determine the cell number, a 10% fraction was trypsinized and counted using a TC20 automated cell counter (Bio-Rad). Organoids were suspended in 100 µl HBSS with 50% BME and 1.5 × 10^5 cells per flank were injected in Beige (CB17.Cg-PrkdcscidLystbg-J/Crl) female mice of around 2 months of age. Growth was scored by palpation and volume was measured with a caliper (multiplying three orthogonal diameters and dividing by two) twice per week. The maximum individual xenograft volume allowed by the ethics committee (350 mm^3) was generally not exceeded, except in 8 out 47 patient-derived xenografts (PDXs) due to very fast tumor growth between the last measurement and the day of being killed (Fig. 6j). For mice injected with multiple xenografts, the maximum tumor burden allowed by the ethics committee (1,200 mm^3) was never reached. For inducible ablation experiments we treated animals with 5 mg kg^{-1} of dimerizer (AP20187, Medchem Express, HY-13992) when tumors reached an average size of 40 mm^3. Dimerizer was administrated intraperitoneally for 3 d consecutively before chemotherapy and then maintained throughout the experiment on a daily basis until the end of FOLFIRI treatment. Animals received five doses of FOLFIRI (5FU (30 mg kg^{-1}), irinotecan (73 mg kg^{-1}) and leucovorin (90 mg kg^{-1})) over a period of 10 d. Then, 13 xenographs were analyzed for NT, FOLFIRI and DIM$^+$ FOLFIRI conditions (five mice per group) and 13 xenographs for the DIM condition (four mice).

Single-cell RNA-seq data of CRC patient samples. We analyzed a previously described cohort[32] of patients with CRC that included samples from Medical Samsung Center (Seoul, South Korea) and the Katholieke Universiteit Leuven (Belgium) deposited at the Gene Expression Omnibus (GEO) under codes GSE132465, GSE132257 and GSE144735 and at ArrayExpress under code E-MTAB-8107. This study was approved by the institutional review board of Commissie Medische Ethiek UZ KU Leuven/Onderzoek. The study was carried out in accordance with ethical guidelines and all patients provided written informed consent as described previously[32]. The study participants were patients diagnosed with CRC who underwent surgery without previous treatment. After resection, tumor samples were collected and immediately transferred for tissue preparation and processing in the 10X Chromium platform as described elsewhere[32]. We only analyzed scRNA-seq data corresponding to epithelial tumor cells. Signatures for LGR5$^+$ cells and KI67$^+$ are detailed in Supplementary Table 6.

For processing, count matrices were downloaded from ArrayExpress (accession no. E-MTAB-8107 for EXT samples) and GEO (accession no. GSE132465 for all SMC.-T samples)[54]. Seurat (v.4.0.3) was used for all future analyses. Cells with mitochondrial content higher than 20%, fewer than 1,000 counts, more than 6,000 or fewer than 200 genes were discarded. Samples were processed individually following Seurat V3 recommendations:[55] SCTransform regressing mitochondrial

content and principal-component analysis dimensionality reduction followed by UMAP computation with 20 components. Expression imputation was computed using MAGIC (v.2.0.3)[56].

For epithelial classification, gene signatures for epithelial, endothelial, CD45$^+$ cells and fibroblasts were extracted from reference[57]. Whenever used, signature scores were computed as the mean of MAGIC expression values of their genes. Non-epithelial scores were summed up. Hierarchical clustering (function hcluster of the amap R package) was applied to the two-dimensional matrix containing the epithelial and non-epithelial scores (scaled to the interval [0–1]) with parameters method, 'euclidean' and link, 'average'. Epithelial clusters were decided by visual inspection in each tumor.

For sample integration, samples were processed individually with SCTransform with parameters to return all genes and min_cells to 1. A total of 3,000 integration features were selected through SelectIntegrationFeatures and the dataset was prepared with PrepSCTIntegration. To maintain as much heterogeneity as possible only stromal cells from the sample with larger stromal content (SMC07-T) were used to find anchors. The function FindIntegrationAnchors was applied to the processed dataset with 30 dimensions on the aforementioned sample. The IntegrateData function was used setting the features.to.integrate parameter to all common genes.

For the epithelial component in the integrated dataset, UMAP and MAGIC expression were computed on epithelial cells without any further normalization or scaling. Cells were classified into cell cycle phase using Seurat's CellCycleScoring. The score for the G0 phase was computed using the AddModuleScore function and the G0 signature extracted from elsewhere[58]. Cells classified as G1 with a G0 score larger than 0 were assigned to G0.

Generation of Mex3a KO AKP organoids. We used CRISPR–Cas9 technology to insert a puromycin resistance cassette that disrupted the first exon of the Mex3a gene. The donor plasmid construction was synthetized by gene synthesis (Thermo Fisher) and consisted of a 2,890-bp DNA fragment containing 935 bp of 5′ Mex3a HA, 1022 bp of SV40 promoter and puromycin cassette followed by 933 bp of 3′ Mex3a HA. Small guide RNAs (GGAGAACTAGGATGTTTCGG) were designed <10 bp around the desired site of knock-in insertion using the https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design web tool and were cloned into a pX330-hSpCas9 plasmid as previously described[31]. To generate Mex3a KO organoids, 2×10^6 of single-cell suspensions of trypsinized mouse cells were nucleofected with 7 µg of donor plasmid and 2 µg of px330-Cas9-corresponding plasmids using Lonza nucleofector kit V (VVCA-1003) and program A32 in an Amaxa-II nucleofector following manufacturer instructions. Culture medium was exchanged with medium containing puromycin (0.5–1 µg ml^{-1}) 3 d after nucleofection. For clonal expansion, puromycin-resistant organoids were seeded one cell per well in a 96-well plate and expanded. KO was verified by western blot.

Mouse single-cell RNA-seq tracing data analysis. Recombination of Mex3a cells in AKP and APS organoids carrying a Mex3a KI Cre-ERT2 locus was achieved by a 16-h pulse of 4-OH-tamoxifen after 2 d of growth. Following 4-OH-tamoxifen removal, FOLFIRI was added to the medium. After 3 d of treatment, drugs were removed, followed by addition of fresh fully supplemented control medium. FACS was used to isolate live cells (DAPI negative) after 16 h of tamoxifen (before FOLFIRI), 4 d (after FOLFIRI treatment) and 30 d after the addition of FOLFIRI (recovery time point). For the last two time points, cells were pre-sorted based on their tracing reporter (tdTomato) expression. Data were deposited in ArrayExpress under accession nos. E-MTAB-11145 and E-MTAB-11146.

To handle class imbalance, 7 parental FOLFIRI and 8 WT1 FOLFIRI samples were randomly downsampled to 1,000 cells. Cells with more than 20% mitochondrial content were discarded. Sample-specific thresholds were set upon inspection for the minimum number of unique molecular identifiers. Count matrices contained 20,506 and 6,581 cells for the Tomato-labeled and loss-of-function experiments, respectively, after removing a cluster of cells with relatively high mitochondrial content. Genes belonging to the RPL and RPS ribosomal components, whose expressions constituted a median of 23.38% in the Tomato-labeled and 20.5% in the loss-of-function samples, were discarded form the expression matrix. SCTransform was applied with mitochondrial content regression. AKP and APS Tomato-labeled samples were normalized independently. All samples from the loss-of-function experiment were normalized together and WT1 and parental samples were considered WT.

The first 15 principal components were used to obtain UMAPs for AKP and APS Tomato-labeled and the first 5 for the loss-of-function experiment. Cells were assigned to clusters using FindClusters (resolution of 1.2). Cell populations were defined as cells having a score above the 90th percentile unless otherwise stated. The Seurat function FindMarkers was used for differential expression analysis from the raw counts (RNA slot). The top 150 genes in terms of average log fold change (with Bonferroni-adjusted $P<0.05$) were selected for functional enrichment analysis with a hypergeometric test. Terms with a Benjamini–Hochberg-adjusted $P<0.05$ were considered significant. The average log fold change was used to perform pre-ranked GSEA[59]. All statistical analyses were performed using R v.4.0.5.

Creation of CRC patient transcriptomic meta-cohort. Public CRC transcriptomic datasets were downloaded from GEO, NCI GDC commons and Synapse data repository and were pre-processed and integrated to create a unique meta-cohort for analysis. The meta-cohort includes 656 tumor samples from three different datasets: TCGA, GSE14333 and GSE39582. Notably, all GEO datasets and the colon TCGA data (TCGA-COAD) include disease-free survival information with a median follow-up ranging from 2.5 to 5 years. Following the TCGA PanCanAtlas Clinical Data Resource (TCGA-CDR) guidelines, TCGA-COAD data were included in the disease-free survival analyses despite the short follow-up available and rectum TCGA samples (TCGA-READ) were excluded from these analyses according to the same guidelines. When not available, microsatellite instability (MSI) status was imputed using the transcriptomic signature reported by Jorissen et al[60]. Sample annotation according to consensus molecular subtypes (CMS) classification was obtained from the original publications.

Datasets were processed separately and merged after gene-wise standardization to the GSE39582 dataset according to their distribution of sex, age, tumor site, MSI and stage. Such standardization was carried out by inverse probability weighting using a logistic regression model. For doing so and for each dataset D, a model was fitted on D and GSE39582 that included the clinical parameters as predictors of the dataset affiliation. Clinical parameters were included in the model according to the Akaike Information Criterion after running a forward step-wise algorithm for variable selection. Conditional probabilities of belonging to datasets D and GSE39582 (Pd and Pr, respectively) were retrieved from the model; weights were computed as the ratio of Pd/Pr after truncating to 5 to avoid overrepresentation of any samples. Expression values were truncated to the maximum and minimum values observed in this reference dataset.

Genotype distributions in patients with mCRC. Mutational status and clinical data were obtained from the cBioPortal[61,62] v.crc_msk_2018. Pathway definitions were obtained from elsewhere[63]. Only homozygous deletions and high amplifications were considered as copy number alterations.

RNA-seq data. RNA-seq was used to compare AKP, APS and APT organoids in complete stem cell medium compared to +TGF-β or medium without EGF, respectively. Single-end reads were aligned to the mouse genome version mm10 using STAR (v.2.3.0e). SAM files were converted to BAM files and sorted using sambamba. The count matrix was generated with Rsubread[64] with the built-in annotation for mm10. DESEq2 was used for differential expression with fold change shrinkage as implemented in the lfcShrink function.

Gene set enrichment analysis. Functional enrichment analysis was performed over gene sets defined in the Hallmarks collection and Gene Ontology database. The rotation-based approach for enrichment[65] implemented in the R package limma was used to represent the null distribution. The max-mean enrichment statistic proposed elsewhere[66], under restandardization, was considered for competitive testing. All custom gene sets are detailed in Supplementary Table 6.

Statistics and reproducibility. Only data that we were able to replicate in at least two independent experiments were included in the manuscript. Animals were excluded only if they died or had to be killed according to protocols approved by the animal experimental committees. No statistical method was used to predetermine sample size but our sample sizes are similar to those reported in previous publications[21,28,31]. Data collection and analysis were not performed blinded to the conditions of the experiments. For in vivo experiments, animals were randomized and treated as indicated in each experiment. Images from a minimum of three mice were quantified in each experiment and each condition. For in vitro experiments, all samples were analyzed equally with no subsampling; therefore, there was no requirement for randomization. The following representative experiments were repeated the indicated number of times with similar results: Fig. 3a,g shows representative images of $n=5$ mice; Figs. 8a,b and 3f were performed three times with similar results; Extended Data Fig. 5f,h shows three independent organoids; and Fig. 5c was repeated twice with similar results. Data distributions were assumed to be normal but this was not formally tested. Transformations were applied whenever needed. For percentage data, normality was assumed for values far from 0 or 100. Statistical analyses were performed using R software package (v.4.0.5) and GraphPad Prism (v.7.03).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

RNA-seq data and microarray expression data that support the findings of this study have been deposited at GEO under accession numbers GSE163171, GSE187650, GSE163035 and GSE187512. scRNA-seq experiments were deposited in ArrayExpress under accession numbers E-MTAB-11145 and E-MTAB-11146. Previously published scRNA-seq data of CRC patient samples that were reanalyzed here are available at GEO under accession codes GSE132465, GSE132257 and GSE144735 and at ArrayExpress under code E-MTAB-8107. Previously published bulk RNA-seq and microarray expression data of CRC patient samples reanalyzed here are available at GEO under accession codes GSE14333 and GSE39582

and at TCGA Research Network (http://cancergenome.nih.gov/) under code TCGA-COAD. Previously published driver mutation data of CRC patient samples from the MSK-Impact cohort (v.crc_msk_2018) reanalyzed here are available at cBioPortal (https://www.cbioportal.org/). All other data supporting the findings of this study are available from the corresponding author on reasonable request. Source data are provided with this paper.

Received: 21 April 2022; Accepted: 20 May 2022;
Published online: 30 June 2022

References

1. Van Cutsem, E., Cervantes, A., Nordlinger, B. & Arnold, D. The ESMO Guidelines Working Group. Metastatic colorectal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **25**, iii1–iii9 (2014).
2. Argilés, G. et al. Localised colon cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **31**, 1291–1305 (2020).
3. Russo, M. et al. Adaptive mutability of colorectal cancers in response to targeted therapies. *Science* **366**, 1473–1480 (2019).
4. Boumahdi, S. & de Sauvage, F. J. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nat. Rev. Drug Discov.* **19**, 39–56 (2020).
5. Shen, S., Vagner, S. & Robert, C. Persistent cancer cells: the deadly survivors. *Cell* **183**, 860–874 (2020).
6. Rehman, S. K. et al. Colorectal cancer cells enter a diapause-like DTP state to survive chemotherapy. *Cell* **184**, 226–242 (2021).
7. Dhimolea, E. et al. An embryonic diapause-like adaptation with suppressed myc activity enables tumor treatment persistence. *Cancer Cell* **39**, 240–256 (2021).
8. Ooft, S. N. et al. Patient-derived organoids can predict response to chemotherapy in metastatic colorectal cancer patients. *Sci. Transl. Med.* **11**, eaay2574 (2019).
9. Vlachogiannis, G. et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science* **359**, 920–926 (2018).
10. Ganesh, K. et al. A rectal cancer organoid platform to study individual responses to chemoradiation. *Nat. Med.* **25**, 1607–1614 (2019).
11. Yao, Y. et al. Patient-derived organoids predict chemoradiation responses of locally advanced rectal cancer. *Cell Stem Cell* **26**, 17–26 (2020).
12. Fujii, M. et al. A colorectal tumor organoid library demonstrates progressive loss of niche factor requirements during tumorigenesis. *Cell Stem Cell* **18**, 827–838 (2016).
13. Sato, T. et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262–265 (2009).
14. Jung, P. et al. Isolation and in vitro expansion of human colonic stem cells. *Nat. Med.* **17**, 1225–1227 (2011).
15. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
16. van de Wetering, M. et al. Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161**, 933–945 (2015).
17. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat. Med.* **23**, 1124–1134 (2017).
18. Matano, M. et al. Modeling colorectal cancer using CRISPR–Cas9-mediated engineering of human intestinal organoids. *Nat. Med.* **21**, 256–262 (2015).
19. Fumagalli, A. et al. Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. *Proc. Natl Acad. Sci. USA* **114**, E2357–E2364 (2017).
20. O'Rourke, K. P. et al. Transplantation of engineered organoids enables rapid generation of metastatic mouse models of colorectal cancer. *Nat. Biotechnol.* **35**, 577–582 (2017).
21. Tauriello, D. V. F. et al. TGF-β drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
22. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
23. Ayyaz, A. et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125 (2019).
24. Nusse, Y. M. et al. Parasitic helminths induce fetal-like reversion in the intestinal stem cell niche. *Nature* **559**, 109–113 (2018).
25. Han, T. et al. Lineage reversion drives WNT Independence in intestinal cancer. *Cancer Discov.* **10**, 1590–1609 (2020).
26. Lupo, B. et al. Colorectal cancer residual disease at maximal response to EGFR blockade displays a druggable Paneth cell-like phenotype. *Sci. Transl. Med.* **12**, eaax8313 (2020).
27. Barriga, F. M. et al. Mex3a marks a slowly dividing subpopulation of Lgr5$^+$ intestinal stem cells. *Cell Stem Cell* **20**, 801–816 (2017).
28. Morral, C. et al. Zonation of ribosomal DNA transcription defines a stem cell hierarchy in colorectal cancer. *Cell Stem Cell* **26**, 845–861 (2020).
29. McFall, T. et al. A systems mechanism for KRAS mutant allele-specific responses to targeted therapy. *Sci. Signal.* **12**, eaaw8288 (2019).
30. Herpers, B. et al. Functional patient-derived organoid screenings identify MCLA-158 as a therapeutic EGFR × LGR5 bispecific antibody with efficacy in epithelial tumors. *Nat. Cancer* **3**, 418–436 (2022).
31. Cortina, C. et al. A genome editing approach to study cancer stem cells in human tumors. *EMBO Mol. Med.* **9**, 869–879 (2017).
32. Lee, H.-O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
33. Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).
34. Schepers, A. G. et al. Lineage tracing reveals Lgr5$^+$ stem cell activity in mouse intestinal adenomas. *Science* **337**, 730–735 (2012).
35. Kozar, S. et al. Continuous clonal labeling reveals small numbers of functional stem cells in intestinal crypts and adenomas. *Cell Stem Cell* **13**, 626–633 (2013).
36. Pereira, B. et al. MEX3A regulates Lgr5$^+$ stem cell maintenance in the developing intestinal epithelium. *EMBO Rep.* **21**, e48938 (2020).
37. Shimokawa, M. et al. Visualization and targeting of LGR5$^+$ human colon cancer stem cells. *Nature* **545**, 187–192 (2017).
38. Qu, M. et al. Establishment of intestinal organoid cultures modeling injury-associated epithelial regeneration. *Cell Res.* **31**, 259–271 (2021).
39. Yui, S. et al. YAP/TAZ-dependent reprogramming of colonic epithelium links ECM remodeling to tissue regeneration. *Cell Stem Cell* **22**, 35–49 (2018).
40. Mustata, R. C. et al. Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep.* **5**, 421–432 (2013).
41. Kreso, A. et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543–548 (2013).
42. de Sousa e Melo, F. et al. A distinct role for Lgr5$^+$ stem cells in primary and metastatic colon cancer. *Nature* **543**, 676–680 (2017).
43. Puig, I. et al. TET2 controls chemoresistant slow-cycling cancer cell survival and tumor recurrence. *J. Clin. Invest.* **128**, 3887–3905 (2018).
44. Touil, Y. et al. Colon cancer cells escape 5FU chemotherapy-induced cell death by entering stemness and quiescence associated with the c-Yes/YAP axis. *Clin. Cancer Res.* **20**, 837–846 (2014).
45. Hu, Z. et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **51**, 1113–1122 (2019).
46. De Conti, G., Dias, M. H. & Bernards, R. Fighting drug resistance through the targeting of drug-tolerant persister cells. *Cancers* **13**, 1118 (2021).
47. Marine, J.-C., Dawson, S.-J. & Dawson, M. A. Non-genetic mechanisms of therapeutic resistance in cancer. *Nat. Rev. Cancer* **20**, 743–756 (2020).
48. Seber, G. A. F. & Wild, C. J. *Nonlinear Regression* (John Wiley & Sons, 1989).
49. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose–response analysis using R. *PLoS ONE* **10**, e0146021 (2016).
50. El Marjou, F. et al. Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis* **39**, 186–193 (2004).
51. Colnot, S. et al. Liver-targeted disruption of Apc in mice activates β-catenin signaling and leads to hepatocellular carcinomas. *Proc. Natl Acad. Sci. USA* **101**, 17216–17221 (2004).
52. Lombardo, Y. et al. Bone morphogenetic protein 4 induces differentiation of colorectal cancer stem cells and increases their response to chemotherapy in mice. *Gastroenterology* **140**, 297–309 (2011).
53. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
54. Qian, J. et al. A pan-cancer blueprint of the heterogeneous tumor micro-environment revealed by single-cell profiling. *Cell Res.* **30**, 745–762 (2020).
55. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
56. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
57. Calon, A. et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer Cell* **22**, 571–584 (2012).
58. Cheung, T. H. & Rando, T. A. Molecular regulation of stem cell quiescence. *Nat. Rev. Mol. Cell Biol.* **14**, 329–340 (2013).
59. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
60. Jorissen, R. N. et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin. Cancer Res.* **14**, 8061–8069 (2008).
61. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
62. Yaeger, R. et al. Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. *Cancer Cell* **33**, 125–136 (2018).
63. Sanchez-Vega, F. et al. Oncogenic signaling pathways in the Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
64. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
65. Wu, D. et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).
66. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).
67. Muñoz, J. et al. The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *EMBO J.* **31**, 3079–3091 (2012).

Acknowledgements

We thank the Clevers laboratory (Hubrecht laboratorium, Utrecht) and the Hub for organoids (Utrecht) for sharing PDOs. We are grateful to the patients who donated their samples for the study. We thank all members of the laboratory for support and discussions. We are grateful for the outstanding assistance by the IRB Barcelona core facilities for histopathology, functional genomics, mouse mutants and advanced digital microscopy as well as the flow cytometry, animal facilities of the UB/PCB and the CRG genomic unit. We thank A. Berenguer for expert support on statistics. C.M. and A.A.-V. have held La Caixa predoctoral fellowships. A.C.-S. and L.J.-G. acknowledge an FPU fellowship from the Spanish Ministry of education. H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III and the Agencia Estatal de Investigación and FEDER (SAF2017-89109-P). L.N. holds a Beatriu de Pinos fellowship from Generalitat de Catalunya. This work has been supported by ERC advanced grant 884623 (residual CRC), IMI grant PERSIST-SEQ, Spanish Ministry of Science PID2020-119917RB-I00 and 2017-SGR-698 (Generalitat de Catalunya). IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO. The single-cell profiling of CRC samples was supported by the Belgian Federation against Cancer grant nos. 2018-127 and 2016-133 and by a grant from Fondation Roi-Baudouin. S.T. is supported by a Senior Clinical Investigator award of the Research Foundation, Flanders.

Author contributions

E.B. and A.A.-V. conceived the study, coordinated experiments and wrote the manuscript. A.A.-V. designed and performed key experiments including in vitro and in vivo lineage-tracing, generated Mex3a-CreErt2 knock-in mouse model. L.N. generated and characterized Mex3a KO MTOs. F.M.B. generated tdTomato-CreERT2 mice and Mex3a-floxed mice. A.C.-S. performed experiments of latency in triple-mutant MTOs in vitro and in vivo. S.C.-C. performed experiments of chemotherapy resistance in MTOs. L.J.-G., G.C. and H.H. provided support with scRNA-seq experiments. X.H.-M. performed all mice work. A.A.-V., C.M., D.S., C.C., S.C.-C., F.S. and G.T. generated CRISPR-engineered organoids. C.S.-O.A. and L.M. analyzed scRNA-seq data and performed statistical analysis. M.S. performed IHC. E.S. provided strategic support and helped with figures and manuscript writing. D.V.F.T. generated MTOs. S.T. granted access to scRNA-seq data. G.S. generated PDO7, CC09 and DA13. P.J. contributed expertise regarding chemotherapy in PDOs. E.B. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-022-00402-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-022-00402-0.

Correspondence and requests for materials should be addressed to Eduard Batlle.

Peer review information *Nature Cancer* thanks Silvia Fre and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Declarations

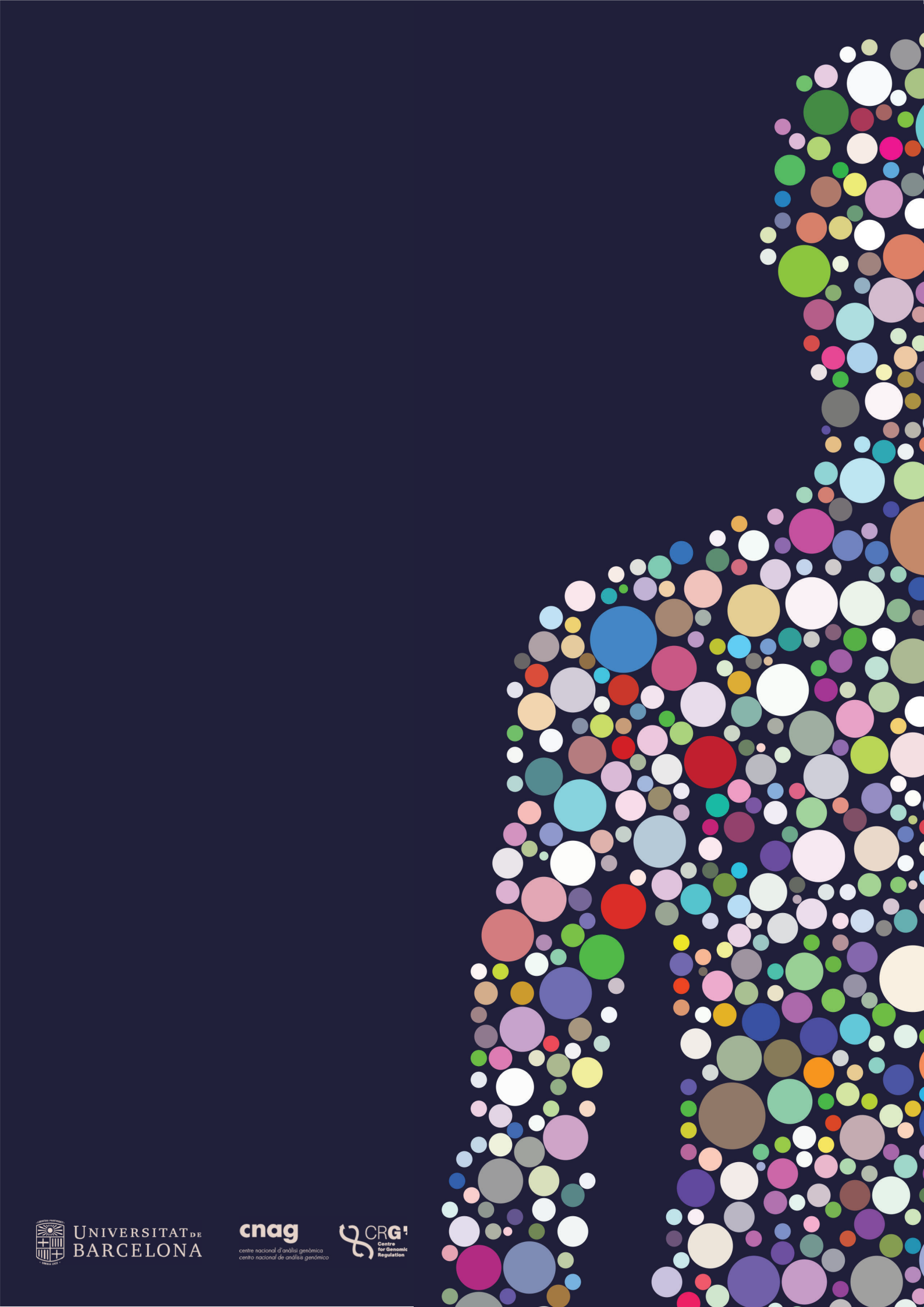
- This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where stated otherwise by reference or acknowledgment, the work presented was generated by myself under the supervision of my advisors during my doctoral studies.

- Whenever a figure or a table is identical to one from a previous publication, it is explicitly stated in the thesis, and the appropriate citations are provided in the captions.

- All references cited in this document have been utilized to derive ideas and support conclusions. These citations aim to acknowledge and give credit to the original authors.

- I acknowledge the use of ChatGPT [https://chatgpt.com/] in assisting me with the final review of the thesis. I critically reviewed the feedback provided by ChatGPT and made modifications to enhance clarity and correctness of the writing.

- The cover design concept was developed by myself, Laura Jiménez-Gracia, with the final design executed by Davide Maspero.



UNIVERSITAT DE BARCELONA

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation