UNIVERSITAT DE BARCELONA

Advancing precision medicine in cancer and COVID-19 with bioinformatics: a multifaceted affair

Carlos Antonio García Prieto

Ph.D. Program in Biomedicine

University of Barcelona

UNIVERSITAT DE BARCELONA

Advancing precision medicine in cancer and COVID-19 with bioinformatics: a multifaceted affair

Carlos Antonio García Prieto

Thesis director and tutor: Dr. Manel Esteller Badosa

Josep Carreras LEUKAEMIA Research Institute

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona, 2024

A mi Sagrada Familia particular.

Por seguir construyendo y creciendo juntos.

Acknowledgements

Quiero agradecer a todas las personas que han formado parte de este camino y que me han acompañado hasta hoy, empujándome y animándome a seguir adelante para alcanzar mis objetivos. Este trabajo no es solo el resultado de mi esfuerzo, sino de muchas personas que trabajan con la ilusión de aportar conocimiento para mejorar la sociedad en la que vivimos. Me siento muy afortunado de haber compartido momentos y vivencias que han marcado mi desarrollo, tanto personal como profesional. Aunque no me gusta personalizar, es necesario agradecer y mencionar a las personas que han tenido un impacto especial en mi trayectoria en estos últimos años.

En primer lugar, quiero agradecer a Manel la oportunidad que me ha dado de trabajar en su laboratorio durante estos años. En especial, por la confianza que ha depositado en mí para formar parte de una gran variedad de proyectos que han contribuido a mi desarrollo científico, y que sobre todo me han permitido aprender de muchas vertientes diferentes de la ciencia, enseñándome a valorar y poner en contexto el trabajo que realizamos. De la misma forma, quiero agradecer a mis compañeros y compañeras del laboratorio por haberme acogido y, sobre todo, por su plena disposición para ayudar en lo que fuera necesario. El capital humano del grupo es lo que nos permite trabajar con confianza y serenidad, siendo clave para el éxito de nuestro trabajo.

A este respecto, quiero mencionar a las personas que, de alguna forma, me acogieron cuando llegué por primera vez al laboratorio. En especial a David, Manu y Pedro, por enseñarme y ayudarme en mi adaptación durante mi primer año. También debo agradecer a quienes ya no forman parte del laboratorio, pero con quienes compartí momentos tanto dentro como fuera del trabajo que hicieron mucho más fácil mi integración en el grupo. Gracias Lorea, Lida y Laia. También quiero tener una mención especial para Alberto y Fer, con quienes he compartido muchas tardes-noches en los viajes de vuelta a Barcelona. Muchas gracias por estar siempre dispuestos a ayudar en lo que sea necesario. Gracias Nacho por todas las comidas y charlas de sobremesa que hemos compartido, por tu serenidad, tu coherencia, y por ser un ejemplo de profesionalidad. Gracias Gerardo, por enseñarme de qué va esto de la ciencia, y junto a Vero y Eva, por ilustrarme la importancia del trabajo detrás de las cámaras para conseguir los recursos necesarios que nos permiten llevar a cabo los proyectos, algo que no se valora hasta que no te toca escribir un proyecto. A Ines y Lucas por su gran trato y compañerismo. A mis compañeras y compañeros con

los que comparto el camino del doctorado, Laura, Marta, Yoana, Eloy y Carlos, mucho ánimo que ya queda menos. A Aleix y Eva con los que no solo comparto despacho, sino también batallas, muchas gracias por vuestra gran ayuda y labor diaria, sois un pilar fundamental del laboratorio. A Espe por estar pendiente de todo lo que necesitamos. A Anna, Bea y Marta por toda la labor de gestión del grupo. A todos los miembros del laboratorio, con los que he compartido tantos momentos, va también mi agradecimiento.

Igualmente quiero recordar a mis colegas del BSC, quienes me acogieron cuando llegué a Barcelona y con quienes compartí infinidad de momentos memorables durante mi estancia. Agradezco especialmente a Alfonso por la oportunidad que me brindó de trabajar en su laboratorio, mostrando gran confianza en mí y permitiéndome formarme en un campo que era completamente ajeno a mis conocimientos. También quiero destacar la labor de Vera y de Eduard, quienes me guiaron durante mis primeros pasos en este nuevo y desconocido universo, llevándome paradójicamente hasta el IJC. Tengo que mencionar también a quienes han compartido muy buenos momentos conmigo, en especial a François y Hugo, con quienes las jornadas de trabajo transcurrían entre risas, y a Iker, compañero de aventuras, con el que he disfrutado de muchos momentos memorables. También a José, gran descubrimiento por su sentido del humor y sabiduría.

Me hace especial ilusión agradecer a través de estas líneas a las personas que sin ninguna duda han sido grandes responsables de que haya continuado mi camino en Barcelona lejos de casa, haciéndome sentir precisamente como si estuviera en ella, y con las que he compartido vivencias, experiencias y enseñanzas que me han ayudado a desarrollarme como persona. Muchas gracias, Davide, Ester y Jon, por haber compartido infinidad de planes que han sido capitales para poder mantenerme a flote durante esta etapa, y por mostrarme una calidad humana y un aprecio que es muy difícil de encontrar hoy en día. Me siento muy honrado de haber podido compartir con vosotros estos años y de haber sido partícipe y testigo de vuestro gran desarrollo profesional y personal.

Quiero agradecer también a todo el profesorado que ha contribuido a mi formación desde mi etapa escolar. Es un gremio al que no se le valora su labor, por lo que desde aquí quiero reconocer todo su esfuerzo y dedicación para formarme, no solo a nivel académico, sino también como ser humano. En la misma línea, quiero agradecer al profesorado

universitario que ha contribuido a que adquiera los conocimientos necesarios para poder trabajar en busca del bienestar social.

En último lugar, os quiero dedicar unas palabras familia. Gracias abuelas y abuelos, tías, tíos, primas y primos, por ayudarme a creer en mí mismo, por confiar en mis capacidades, y por empujarme a afrontar nuevos retos. Gracias, mamá y papá, no solo por vuestro apoyo incondicional, al que no hacen justicia estas líneas, sino por educarnos y transmitirnos valores fundamentales como el respeto, el esfuerzo y la honestidad, que nos permiten afrontar y gestionar los retos a los que nos enfrentamos, valorando también los obstáculos que vamos superando. Gracias Pablo, por mostrarme la importancia de luchar por tus sueños y de seguir tu propio camino. Me siento muy afortunado de poder compartir mi vida con vosotros, sois mis referentes y mi inspiración. Estar fuera de casa me ha hecho valorar aún más la importancia de la familia como eje fundamental de la vida. No puedo más que agradeceros por todo lo que me habéis dado y me seguís dando. Gracias por enseñarme a ser, y querer ser, mejor persona. Mis logros como individuo son el reflejo de vuestra labor como madre, padre y hermano.

Para finalizar, me gustaría recordar a todas aquellas personas anónimas que, en su día a día, dedican su tiempo a ayudar a quienes lo necesitan. Este trabajo no es más que el reflejo de un propósito mayor, el de generar el conocimiento necesario para intentar mejorar las condiciones de vida de aquellas personas que viven afectadas por diversas patologías que les impiden disfrutar de una calidad de vida óptima.

Gracias a todos y todas por haberme acompañado y por haber sido partícipes de este propósito.

Abstract

Abstract

Advancing precision medicine requires integrating bioinformatics to unravel complex biological data and translate these insights into clinical applications. This thesis explores the role of bioinformatics in enhancing our understanding of cancer and infectious diseases through studies focused on cancer genomics, immunotherapy, and COVID-19. In cancer genomics, a comparative analysis of variant calling tools revealed significant variability in their ability to identify cancer driver genes and clinically actionable variants, underscoring the need for tailored strategies across different cancer types. Combining mutations from multiple callers proved more effective in cancer driver gene detection, while MuTect2 identified more subclonal and actionable mutations linked to therapeutic outcomes. In the context of immunotherapy, we developed the EPICART signature, a DNA methylation-based classification model that successfully predicted complete clinical response in patients receiving CD19-targeted chimeric antigen receptor (CAR) T-cell therapy for relapsed or refractory B-cell malignancies. EPICART-positive CAR T-cell products, characterized by higher proportions of naïve and central memory T-cells, were associated with improved clinical outcomes. Importantly, the EPICART signature has since been licensed to a pharmaceutical company for validation in diverse patient cohorts, representing a key step toward potential clinical implementation. Extending the application of DNA methylation profiling to COVID-19, we identified the EPIMISC signature, which differentiated multisystem inflammatory syndrome in children (MIS-C) from pediatric COVID-19 cases without MIS-C. The presence of EPIMISC in Kawasaki disease further suggested shared immune mechanisms, likely triggered by viral infections such as SARS-CoV-2 in MIS-C. To deepen our understanding of COVID-19 pathology, we applied spatial transcriptomics to investigate diffuse alveolar damage in fatal cases, revealing key contributors to lung fibrosis, including aberrant myeloid activation, peribronchial fibroblast proliferation, and activation of the TGF-β/SMAD3 pathway. These findings highlight the critical role of bioinformatics in advancing precision medicine and emphasize the importance of multisectoral collaboration for clinical translation.

Table of contents

Introduction

Introduction

The origins of modern medicine can be traced back to the late 18th century with the discovery of the smallpox vaccine (**Figure 1**), a breakthrough that shifted the focus from merely treating symptoms to actively preventing infectious diseases, the leading cause of death during the 19th and beginning of the 20th centuries[1–3]. The identification of bacteria and viruses as causative agents of infectious diseases[4] led to significant improvements in public health measures, including sanitation and quarantine methods, and to the landmark discovery of antibiotics[5,6] (**Figure 1**). These milestones drastically reduced mortality rates and enhanced infection control, shifting the medical focus to non-communicable diseases, particularly cancer, which became a major health threat in the latter half of the 20th century, partly due to the rise in smoking-related lung cancer[7].

Early cancer treatments were limited to surgery[8] and rudimentary radiation[9]. After World War II, the introduction of chemotherapy marked the start of systemic cancer treatments[10,11] (**Figure 1**). However, traditional therapies often lacked specificity, resulting in significant toxicity and resistance[12], underscoring the need for more precise therapeutic strategies. The introduction of sequencing technologies prompted the identification of cancer-specific targets[13,14], setting the stage for the development of precision medicine and targeted therapies[15] (**Figure 1**).

The development of next-generation sequencing (NGS) technologies[16,17] has revolutionized biomedical research, allowing for the comprehensive analysis of genetic mutations in cancer and other diseases. This technological leap has enabled the identification of novel biomarkers and therapeutic targets that inform personalized medicine. Complementing these genetic insights, the field of epigenetics has further expanded our understanding of disease mechanisms by revealing how chemical modifications to Deoxyribonucleic acid (DNA), such as methylation, affect gene expression without altering the underlying genetic code. Epigenetics has provided new avenues for understanding cancer development and progression[18,19].

In addition to transforming cancer research, this precision-driven approach played a critical role in addressing global health challenges, including the coronavirus disease

2019 (COVID-19) pandemic[20] by facilitating rapid genomic analysis and vaccine development (**Figure 1**).

Building on these advancements, spatial transcriptomics (ST)[21] has emerged as a powerful tool for studying cellular interactions within their native tissue context. By preserving spatial information, this technology allows us to deepen our understanding of how cells communicate and respond within their microenvironments, providing crucial insights into both normal physiology and disease states.

A multifaceted approach that utilizes various types of biological data, including genetic, epigenetic, spatially resolved, and clinical information, is essential for advancing our understanding of complex diseases and for the development of precision medicine, which tailors treatment to the unique (epi-)genetic, molecular and clinical characteristics of individual patients.

Despite these advances, several key questions remain unanswered: How can we enhance the detection of oncogenic and clinically actionable variants (CAVs) in cancer to improve treatment decisions? What drives successful responses to immunotherapy, and how might epigenetic markers help improve its efficacy? Additionally, how can we better understand the dysregulated inflammatory response in COVID-19, particularly in the subset of children who develop severe post-infectious complications? Furthermore, what are the molecular changes driving the progression of lung tissue damage that led to fatal outcomes in adults? These are the central questions explored in the subsequent sections of this thesis.

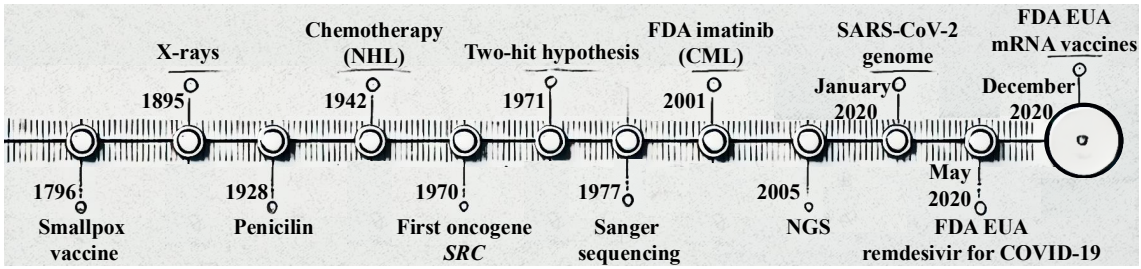


Figure 1. Timeline of major breakthroughs in modern medicine, highlighting pivotal advancements that have shaped current medical practices. Emphasis is placed on the rapid development of COVID-19 vaccines. NHL: non-Hodgkin lymphoma; *SRC*: sarcoma; CML: chronic myelogenous leukemia; NGS: next-generation sequencing; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; FDA: Food and Drug Administration; EUA: emergency use authorization; mRNA: messenger ribonucleic acid; COVID-19: coronavirus disease 2019.

1. Cancer

1.1 Definition of cancer

The definition of cancer has evolved significantly as our understanding of this complex disease increased. In the late 19th and early 20th centuries, cancer was described focusing mostly on observable phenomena like the uncontrolled proliferation of cells and tumor formation[22,23].

In contemporary times, organizations such as the National Cancer Institute and the World Health Organization (WHO) have defined cancer more specifically as a disease or group of diseases, respectively, characterized by uncontrolled cell growth with the potential to spread to other parts of the body[23]. While these definitions recognized a systemic affection, they do not account for the dynamic nature of cancer.

A more nuanced characterization is provided by the recent definition of cancer as 'a disease of uncontrolled proliferation by transformed cells subject to evolution by natural selection'[23]. This updated definition acknowledges that cancer cells are transformed entities undergoing continuous evolution. This evolution is mostly driven by genetic mutations, epigenetic modifications, and selective pressures within the tumor microenvironment (TME).

By considering cancer through this evolutionary lens, researchers and clinicians can better appreciate the complexities of the disease. This approach emphasizes the continuous and adaptive evolution of cancer cells, which drives disease progression and presents significant challenges for treatment. The comprehensive characterization of this dynamic process is crucial for developing more effective therapeutic strategies to combat this multifaceted disease.

1.2 Classification of cancer

Classification of cancer has traditionally been based on the tissue or organ of origin, providing a foundation for cancer diagnosis and management. The WHO Classification of Tumors employs this approach, classifying cancers by their primary site, such as breast, head and neck, or central nervous system tumors[24], which helps guide initial clinical decisions. Additionally, cancers can be classified by histological type, including

carcinomas, sarcomas, leukemias or lymphomas, based on the type of tissue involved, like epithelial and connective tissues, bone marrow or lymphatic nodes.

Importantly, another traditional aspect of classification is tumor grading[25], which evaluates the differentiation degree of cancer cells, ranging from well differentiated (low grade) to poorly differentiated (high grade). Furthermore, the TNM classification system[26] stages cancer by assessing the size of the tumor (T), involvement of lymph nodes (N), and the presence of metastasis (M). These approaches further inform treatment planning and prognosis.

With advances in cancer genomics, molecular classification has emerged as a complementary approach. Recent studies have shown that tumors can also be grouped based on genetic and molecular characteristics[27,28], uncovering new cancer subtypes across different tissues, revealing more precise insights into tumor biology and informing personalized treatment strategies. Integrating molecular characteristics with traditional classification systems enhances the ability to tailor cancer treatments, ultimately improving patient outcomes.

1.3 Cancer statistics

In 2022, an estimated 20 million new cancer cases were reported globally, alongside 9.7 million cancer deaths[29]. Approximately one in five individuals develop cancer during their lifetime, with one in nine men and one in twelve women dying from the disease. As a major public health and socioeconomic challenge of the 21st century, cancer accounts for one in six deaths worldwide and is responsible for three out of ten premature deaths caused by non-communicable diseases[29].

Lung cancer remains the most frequently diagnosed cancer, representing 12.4% of all new cancer cases, followed by breast cancer (11.6%) and colorectal cancer (9.6%). Lung cancer is also the leading cause of cancer-related deaths, responsible for 1.8 million deaths (18.7%), followed by colorectal cancer (9.3%) and liver cancer (7.8%)[29]. Additionally, breast cancer is the most frequent cancer among women, while lung cancer is the most common among men, both in terms of incidence and mortality[29].

The global cancer burden is expected to rise to 35 million new cases by 2050, driven by demographic shifts and population growth[29]. To mitigate this increasing burden, prioritizing prevention strategies is essential. These strategies include reducing smoking rates, addressing obesity, and increasing access to vaccination and early screening programs. Effective global cancer control requires a multifaceted approach focusing on prevention, early detection, and equitable access to treatment[30].

According to the 2024 American Association for Cancer Research Cancer Progress Report[31], 40% of all cancers in the United States are linked to modifiable risk factors, underscoring the importance of public health initiatives aimed at reducing cancer risk. Nearly 20% of cancer diagnoses in the United States are associated with excess body weight, unhealthy dietary habits, alcohol consumption, and physical inactivity. These statistics highlight the potential for lifestyle changes to reduce cancer incidence and improve public health outcomes.

In Spain, the Spanish Statistical Office reported that in 2023, for the first time, cancer became the leading cause of death, accounting for 26.6% of all deaths in the country. This highlights the pressing public health challenge cancer presents in Spain, reinforcing the need for sustained efforts in prevention, early detection, and treatment to address the rising cancer burden.

1.4 Genetic basis of cancer

The uncovering of the causes of cancer and the development of genetics are closely intertwined[14]. Initial records of the causes of cancer date back to the late 18th century, when the occupational exposure of chimney sweeps to soot was first linked to an increased risk of cancer[32], marking one of the earliest demonstrations of a direct connection between environmental exposure and cancer. In the 19th century, even before Mendel's laws of inheritance were recognized, reports began to document familial patterns of breast cancer[33], suggesting a genetic predisposition to the disease.

The early 20th century marked a significant turning point with the identification of chromosomal abnormalities associated with cancer[34], laying the groundwork for cancer genetics. The discovery of cancer driver genes, defined as genes whose mutations can initiate and promote cancer, was pivotal. By the late 1960s, the term 'oncogene' was

introduced[13], reflecting the acquired ability of certain genes to drive cell transformation when mutated. The identification of *HRAS* as the first human oncogene demonstrated that a single nucleotide change could convert normal cells into cancerous ones[35], even when the non-mutated allele was still present. This discovery highlighted that oncogenes act in a dominant fashion, meaning only one mutated allele is necessary for cancer development.

The two-hit hypothesis[36], proposed in 1971 (**Figure 1**), further advanced our understanding of cancer genetics by suggesting that both alleles of a tumor suppressor gene must be inactivated for a tumor to form, being later confirmed with the identification of *RB1*[37], the first tumor suppressor gene. This model explained the mechanism of familial cancers where a germline mutation serves as the first 'hit' with a subsequent somatic mutation acting as the second.

In the following decades, numerous cancer driver genes were discovered, including key oncogenes such as *MYC*[38] and *EGFR*[39], and tumor suppressor genes like *TP53*[40] and *PTEN*[41]. Additionally, the identification of germline mutations in certain tumor suppressor genes, such as *BRCA1*[42] and *BRCA2*[43] in breast and ovarian cancers, and *APC* in colorectal cancer[44], revealed genetic predispositions to these cancers. These discoveries underscored that cancer could result from disruptions in a range of key cellular functions, leading to a comprehensive exploration of cancer genomics. This growing field continues to refine our understanding of the genetic mechanisms underlying cancer and guides the development of targeted therapies.

1.4.1 Cancer genomics

The field of cancer genomics advanced rapidly in the early 21st century with the advent of NGS technologies[16] and the completion of the Human Genome Project[45,46]. These breakthroughs enabled large-scale initiatives like The Cancer Genome Atlas (TCGA)[47] and the International Cancer Genome Consortium (ICGC)[48], which conducted whole-exome sequencing (WES) and whole-genome sequencing (WGS) on hundreds to thousands of tumor samples across multiple cancer types. These projects aimed to characterize the molecular basis and mutational landscape of cancer genomes, including the identification of oncogenic mutations with unprecedented resolution to help define new clinically relevant subtypes and enable the development of new targeted therapies.

These projects have generated vast amounts of sequencing data, revealing the complexity and diversity of somatic mutations involved in tumorigenesis[49,50]. A major challenge emerging from these efforts have been distinguishing between somatic 'driver' mutations in cancer genes, which directly contribute to cancer development (**Figure 2A**), and the remaining somatic mutations called 'passengers', which do not. In cancer genomes, somatic passenger mutations outnumber somatic drivers.

The accurate identification of these mutations, a process known as variant calling, is a critical step that underpins all downstream analyses in cancer genomics. The precision and reliability of variant calling directly impact the insights derived from these studies, highlighting the importance of standardizing the tools and strategies used to ensure consistency and reproducibility in cancer research.

1.4.1.1 Variant calling

Cancer genomes accumulate a wide range of somatic mutations[51], including single nucleotide variants (SNVs), short (<50 base pairs) insertions and deletions (indels), chromosomal rearrangements, and copy number alterations. Accurately identifying these mutations through variant calling is particularly challenging in cancer due to several factors such as tumor heterogeneity, varying levels of sequencing coverage, the presence of subclonal mutations, and the significant computational resources required[52,53]. Despite its critical role in cancer research, variant calling has not been fully standardized across different studies, even within major initiatives like TCGA.

The computational analysis of cancer genome sequencing data is a complex process involving multiple steps, each of which can be carried out using various tools that significantly influence the results[54,55]. The initial step is preprocessing, which includes mapping sequencing reads to a reference genome and performing quality assessments. In a typical tumor-normal paired design (**Figure 2B**), variant calling identifies candidate somatic mutations as genomic positions where an alternate allele present in the tumor reads is absent in the matched normal sample (**Figure 2C**). A critical factor in this process is the variant allele frequency (VAF), which represents the fraction of reads supporting the mutation. The VAF is influenced by factors such as tumor purity (proportion of cancer cells in the sequenced sample), tumor ploidy (amount of DNA in cancer cells), and intra-

tumor heterogeneity. Detecting mutations with low VAFs often requires high-depth sequencing or sophisticated algorithms.

Most variant calling tools use a matched normal sample to differentiate true mutations from sequencing artifacts. Advanced algorithms such as Mutect2[56], MuSE[57] or SomaticSniper[58], employ probabilistic models to refine mutation detection, while other methods like VarScan2[59] rely on heuristic thresholds. Each of these tools varies in sensitivity, specificity, and the types of mutations they detect, leading to inconsistencies in mutation calls across different studies[60,61]. For instance, MuTect2 is often favored for its high sensitivity in detecting low VAF mutations, whereas MuSE, VarScan2 and SomaticSniper may be used to balance sensitivity and specificity. Additionally, to filter out germline variants, databases of genetic variation such as dbSNP[62] and gnomAD[63] are used to exclude known single nucleotide polymorphisms from variant calls. The selection of tools and parameters at each stage critically impacts the consistency and accuracy of the variant calling results.

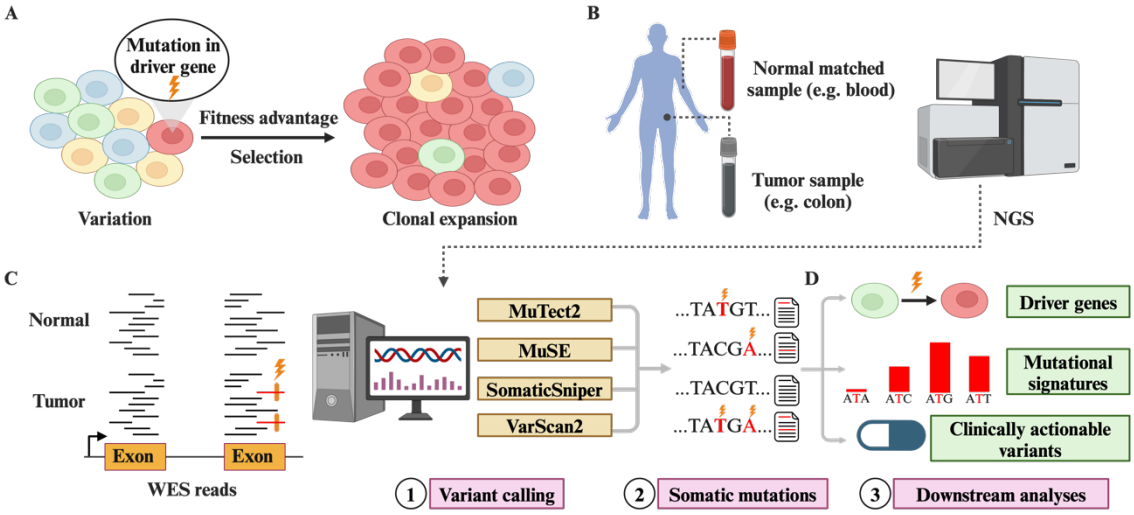


Figure 2. Workflow for cancer genome analysis, illustrating somatic variant detection as a basis for downstream analyses. A | Clonal selection during tumor development, highlighting how somatic mutations in cancer driver genes confer a selective advantage, leading to clonal expansion and tumor evolution. **B |** Overview of matched tumor-normal analysis, involving DNA extraction and sequencing from both the tumor and a matched normal sample (e.g., peripheral blood) to distinguish somatic from germline mutations. **C |** NGS data (e.g., WES) is processed for variant calling analysis and utilized by tools such as MuTect2, MuSE, SomaticSniper, and VarScan2, to detect somatic mutations like point mutations. **D |** Downstream analyses include the detection of cancer driver genes, mutational signatures, and clinically actionable variants, providing insights into tumorigenesis, mutational processes, and therapeutic opportunities. This figure summarizes key methodologies and analyses presented in Chapter I, focusing on the impact of variant calling decisions on the results of important downstream analyses in cancer genomics. NGS: next-generation sequencing; WES: whole-exome sequencing. Created in BioRender.com.

This variation in tool usage and performance highlights a fundamental challenge: without a standardized approach to variant calling, the mutation profiles generated by different studies are not directly comparable.

Efforts to harmonize variant calling processes have been made through initiatives like the Multi-Center Mutation Calling in Multiple Cancers (MC3)[64] project and the Pan-Cancer Analysis of Whole Genomes (PCAWG)[65]. These projects sought to standardize variant calling by combining outputs from multiple tools to improve consistency and reproducibility. However, the impact of this approach on downstream analyses remains uncertain, as it often prioritizes specificity at the cost of sensitivity. To address these limitations, strategies that leverage machine learning (ML) have been proposed to enhance both aspects[66].

The Genomic Data Commons[67] further supports this effort by providing a data platform that ensures reproducibility and comparability across studies through standardized variant calling pipelines for each tool. Despite these efforts, no universally accepted standard for variant calling in cancer genomics has been established, as different tools and methods continue to be used.

While several benchmarking studies have assessed the performance of variant callers[68,69], including efforts like the DREAM Challenge[70], these studies typically focus on the ability of tools to recall specific synthetic or known variants under controlled conditions. Such benchmarks are valuable but limited because they do not address the broader consequences of variant calling decisions on downstream analyses. The impact of choosing one variant caller over another or using different strategies to combine outputs from multiple callers, on critical downstream analyses such as the detection of cancer driver genes, mutational signatures, and CAVs, remains underexplored (**Figure 2C**).

Understanding these downstream impacts is crucial because variant calling is a foundational step that influences all subsequent analyses and interpretations in cancer genomics. This thesis aims to address this gap by examining how variant calling decisions affect key secondary analyses that characterize cancer biology and inform patient treatment. The findings and implications of this research are presented in detail in **Chapter I**.

1.4.1.2 Cancer driver genes

TCGA and ICGC initiatives provided a wealth of data on somatic mutations in tumors, facilitating the identification of cancer driver genes. These genes harbor mutations that confer a selective advantage to cells, leading to their clonal expansion in tumors (**Figure 2A**). To identify cancer driver genes, computational methods look for deviations in mutation patterns from what would be expected under neutral mutagenesis[71]. These deviations indicate positive selection. For example, cancer driver genes often show an unusually high frequency of mutations[72,73], a bias towards mutations with high functional impact[74,75], clustering of mutations in specific regions of the protein[76,77], or a skew in the frequency of trinucleotide changes[78]. Over time, various computational methods have been developed to detect these signals of positive selection.

The application of these methods to large tumor datasets, such as those from TCGA and ICGC, has paved the way for the identification of a comprehensive compendium of cancer driver genes[14,79–81]. However, this task is challenging due to several factors. Early analyses of these datasets revealed that mutation types vary significantly across tumors of different origins, and mutation rates across the genome can be highly heterogeneous, even among samples of the same cancer type. This variability can stem from technical differences, such as sequencing technologies, depth, and variant calling methods, as well as biological factors, including varying exposures to mutational processes. As a result, accurately modeling the background mutation rate of genes is crucial for identifying mutational patterns that are under positive selection. Moreover, the vast amount of genomic data produced by these projects, coupled with current computational limitations, complicates the consistent recall of mutations across all cohorts and studies, necessitating separate analyses for each tumor type.

The development of comprehensive computational pipelines like IntOGen[14] has been crucial in addressing these challenges. The IntOGen pipeline integrates outputs from various methods designed to detect signals of positive selection, reducing false positives and spurious results for a more accurate identification of cancer driver genes. Given the diverse mutational landscapes across cancer cohorts, identifying less frequently mutated driver genes requires large cohort sizes to enhance statistical power and detection sensitivity[82].

The resulting compendium of driver genes and their specific mutational features provides valuable insights into their roles in cancer development[14]. The types of mutations prevalent in these genes can indicate whether a gene functions as an oncogene or a tumor suppressor. Oncogenes are often characterized by an excess of missense mutations, which result in a different amino acid being encoded in the protein sequence, without a corresponding increase in nonsense mutations, which introduce a premature stop codon. This pattern suggests a gain-of-function role. In contrast, tumor suppressor genes typically exhibit a higher frequency of nonsense mutations, reflecting their loss-of-function role in cancer progression.

The Cancer Gene Census (CGC)[83] is a comprehensive catalogue of validated cancer driver genes curated from the scientific literature. As part of the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, the CGC serves as an essential 'ground truth' resource for cancer genomics. Currently, the catalogue includes more than 700 cancer driver genes, and its completion is crucial for advancing our understanding of tumor biology and uncovering the roles these genes play in tumorigenesis across various cancer types.

1.4.1.3 Mutational signatures

The somatic mutations in tumor genomes serve as a historical record of the mutational processes that have shaped tumor development, reflecting the various events and exposures that tumor cells have experienced throughout a patient's life. By examining these mutational patterns, known as mutational signatures, we can gain valuable insights into the environmental and biological factors that have influenced the evolution of the tumor[84,85]. Each mutational signature is characterized by a unique combination of mutation types, including single base substitutions (SBS), short indels, and genomic rearrangements. For SBS, the specific pattern is determined by the nucleotide change, one of six possible substitutions (C>A, C>G, C>T, T>A, T>C, T>G), along with the 5' and 3' bases immediately flanking the mutated site. When considering all possible sequence contexts, these combinations result in 96 distinct mutation types (4x6x4). These patterns can indicate specific mutational processes, such as exposure to ultraviolet (UV) light or tobacco smoke, or intrinsic factors like DNA replication errors and defects in DNA repair mechanisms.

Mutational signatures can be broadly classified into those common across many cancer types and those specific to exposures or biological processes. For example, signatures associated with aging[86] and spontaneous deamination of cytosines[87] are seen across various cancers, while UV-induced damage and smoking-related mutations are more specific to skin cancers and lung cancers, respectively[84,85]. Certain signatures also reveal biases in DNA repair activities, suggesting disruptions in normal cellular repair functions.

The identification and analysis of mutational signatures rely on computational algorithms like non-negative matrix factorization (NMF), which decompose the mutational catalogue of a tumor into distinct signatures, each corresponding to a different underlying mutational process, and its relative contribution to the tumor's mutational burden[84,88]. This approach has led to the construction of the COSMIC Mutational Signatures catalogue, which include over 50 distinct SBS processes[85]. With this catalogue, researchers can estimate mutational signatures in new tumor samples without needing a large cohort for *de novo* signature extraction, utilizing tools that apply algorithms like non-negative least squares to refit the known signatures[89].

This detailed analysis not only enhances our understanding of cancer biology but also informs treatment strategies. Recognizing the active mutational processes in a tumor can reveal its underlying causes and guide preventive and therapeutic approaches. For instance, tumors with specific DNA repair deficiencies, such as those with homologous recombination repair defects, may be more susceptible to targeted treatments like poly (ADP-ribose) polymerase (PARP) inhibitors[90]. Similarly, tumors with deficient mismatch repair (dMMR), which often lead to microsatellite instability (MSI) in colorectal cancer, can respond particularly well to immunotherapy[91]. Moreover, mutational signatures reflecting the impact of previous treatments, such as chemotherapy regimens, have been described[92]. Thus, integrating mutational signature analysis into cancer research provides a powerful tool for unraveling the diverse mechanisms of tumorigenesis and developing more personalized and effective therapies.

1.4.1.4 Clinically actionable variants

The integration of genomic profiling into cancer care has been a cornerstone of precision medicine, an approach that tailors treatment to the individual characteristics and needs of

each patient. In oncology, this involves using detailed tumor genetic information to identify CAVs, meaning genetic alterations that guide treatment decisions and predict therapeutic responses[93]. This approach represents a shift from the traditional one-size-fits-all model to a more personalized strategy, where treatments are customized based on the unique molecular profile of each tumor, moving beyond traditional histological classifications and enabling the identification of subtypes according to a molecular taxonomy[27,28].

CAVs identified through NGS platforms play a crucial role in guiding the choice of targeted therapies and immunotherapies. For example, the identification of the *BCR-ABL* fusion gene in chronic myelogenous leukemia[94] led to the development of imatinib (**Figure 1**), the first targeted therapy approved by the Food and Drug Administration (FDA), which specifically inhibits this fusion protein[95]. Similarly, *EGFR* mutations in lung cancer and *BRAF* mutations in melanoma have become critical biomarkers for selecting appropriate targeted treatments, such as *EGFR* inhibitors[96,97] and *BRAF* inhibitors[98].

Advances in NGS technology and reduced costs have accelerated the adoption of larger panel NGS-based diagnostic platforms, facilitating the discovery of new drug targets and companion diagnostics[99], which are tests designed to identify patients most likely to benefit from specific therapies. This approach has also enabled innovative clinical trial designs, such as basket trials[100], which enroll patients based on specific molecular alterations regardless of tumor type, leading to FDA approval of tumor-agnostic biomarkers like MSI[101,102], tumor mutational burden (TMB)[103], and *NTRK* gene fusions[104].

However, interpreting these variants can be challenging, particularly because many are variants of unknown significance with insufficient evidence to guide clinical decisions. Moreover, different mutations within the same oncogene can exhibit distinct biological properties, resulting in varied drug responses. For instance, the response to vemurafenib in patients with *BRAF*-V600E mutations varies depending on the tumor type[98,105,106], posing challenges for clinicians in selecting the most appropriate treatments.

To improve variant interpretation, tools such as the Molecular Oncology Almanac (MOAlmanac)[107], Cancer Genome Interpreter[108], PanDrugs[109], and OncoKB[110], catalogue the biological properties and clinical implications of specific mutations. MOAlmanac, for example, stands out by integrating both first-order and second-order molecular features to guide treatment decisions. First-order features include specific gene variants, copy number alterations, and fusions, while second-order features cover broader molecular characteristics such as mutational signatures, TMB, and whole-genome doubling. This comprehensive approach provides a more nuanced interpretation of a tumor's genomic landscape. Moreover, MOAlmanac incorporates a diverse range of data sources and uses an evidence-based framework to classify genomic alterations. It evaluates the clinical relevance of each marker by linking molecular features to therapeutic sensitivity, resistance, and prognosis, categorizing these features based on the strength of clinical evidence, including FDA-approved markers, clinical trial data, and preclinical studies.

These tools underscore the dynamic evolution of cancer genomics, emphasizing the need to integrate multiple data types and leverage sophisticated computational frameworks to advance cancer treatment strategies and improve patient outcomes.

1.5 Complexities of cancer

Cancer is driven by mutations in cancer driver genes, which affect essential cellular functions and confer selective advantages to tumor cells. These advantages, known as the hallmarks of cancer[111,112], enable tumor cells to maintain proliferative signaling, evade growth suppressors, resist apoptosis, sustain replicative immortality, induce angiogenesis, initiate invasion and metastasis, reprogram energy metabolism, and avoid immune destruction. Genomic instability and tumor-promoting inflammation further facilitate the acquisition of these traits, making cancer a multifaceted disease.

Beyond mutations in cancer driver genes, the TME plays a crucial role in cancer progression. The TME, a diverse ecosystem comprising various cancerous and non-cancerous cells, such as fibroblasts, inflammatory immune cells, and endothelial cells, can support and promote tumor growth[112–114]. These recruited normal cells form the tumor stroma interact with cancer cells through complex signaling networks, actively participating in tumorigenesis. This dynamic interplay is critical in promoting hallmark

capabilities, such as inducing angiogenesis, facilitating tissue invasion and modulating immune responses, underscoring the importance of the TME in cancer progression[112].

Immune evasion, another hallmark of cancer, is critical for tumor survival and growth within the host organism. Cancer cells avoid detection and destruction by the immune system through various strategies, including secreting immunosuppressive molecules and recruiting regulatory immune cells[113,115,116]. The emergence of immunotherapy as a treatment modality has underscored the significance of this hallmark, as targeting immune evasion mechanisms can reinvigorate the host's immune response against cancer.

Genomic instability, a key enabler of cancer, generates the diversity needed for acquiring other hallmarks. This instability arises not only from genetic mutations but also from epigenetic alterations, such as DNA methylation and histone modifications[19,117]. These changes can silence tumor suppressor genes or activate oncogenes without altering the DNA sequence, indicating that some clonal expansions are driven by non-mutational changes affecting gene expression regulation. This perspective introduces the role of epigenetic reprogramming in cancer, highlighting the importance of both genetic and epigenetic alterations in driving tumor progression.

Recent advancements in cancer biology have expanded our understanding of these processes[118]. Phenotypic plasticity allows cancer cells to adapt to environmental cues, altering their state to enhance metastatic potential and treatment resistance[119]. Epigenetic reprogramming, often driven by the TME, further support this adaptability[120,121]. For instance, hypoxic conditions within the TME can lead to widespread alterations in the epigenome[122], fostering an environment that supports malignancy.

Another example is the epithelial-to-mesenchymal transition (EMT), driven by epigenetic alterations in response to TME signals, which facilitates metastasis by enabling cancer cells to invade other tissues[123]. EMT is frequently initiated by signals from the TME, such as cytokines and growth factors, which trigger epigenetic changes that stabilize the mesenchymal state[124]. Moreover, the TME contains various cell types that can be epigenetically reprogrammed to support tumor growth, creating a niche that promotes

cancer cell survival and invasion, highlighting the profound impact of the TME on cancer progression through epigenetic mechanisms[118].

Additionally, cellular senescence and the microbiome play emerging roles in cancer progression. Senescent cells within the TME can paradoxically promote tumorigenesis through secretory factors[125,126], while the gut microbiome has been shown to modulate immune responses and influence cancer development[127,128], further underscoring the multifaceted nature of cancer.

In light of these expanded hallmarks of cancer, it is clear that cancer is not just a genetic disease, but a dynamic system profoundly influenced by its surrounding environment and epigenetic regulation.

1.6 Epigenetics in cancer

Epigenetic modifications are essential regulators of gene expression, playing a critical role in cancer biology by influencing cellular behavior and phenotype. Unlike genetic mutations that alter the DNA sequence permanently, epigenetic changes are reversible modifications that allow for a dynamic response to environmental cues, contributing to cancer's adaptability and heterogeneity. Recently, the importance of these modifications has been increasingly recognized, leading to their classification as epigenetic hallmarks of cancer[129], further highlighting their pivotal role in tumor development and evolution. Additionally, large-scale collaborative projects, such as the Encyclopedia of DNA Elements[130] and the International Human Epigenome Consortium[131], have significantly advanced our understanding of the epigenome by mapping regulatory elements and epigenetic markers across different tissues, cell types, and conditions.

One of the most well-characterized epigenetic modifications is DNA methylation, which involves the addition of a methyl group to the fifth carbon of cytosine residues within CpG (cytosine followed by guanine) dinucleotides. These CpG sites are often clustered in regions known as CpG islands, frequently found in gene promoter regions where they play a crucial role in regulating gene activity[132]. In somatic cells, DNA methylation typically displays a bimodal pattern where most of the genome is heavily methylated, while CpG islands, particularly those in the promoters of housekeeping genes, remain unmethylated to allow consistent gene expression. DNA methylation is also vital for

normal cell differentiation[133,134] and various developmental processes, such as genomic imprinting, X-chromosome inactivation, and the suppression of transposable elements, which are essential for maintaining genomic stability[135].

The balance of DNA methylation is maintained by two key types of enzymes: DNA methyltransferases (DNMTs), which establish and maintain methylation marks, and TET enzymes, which facilitate active demethylation and enable gene reactivation. This interplay between methylation and demethylation is crucial for regulating gene expression and maintaining cellular identity, and disruptions in these processes can contribute to cancer development.

In cancer, the DNA methylation landscape is markedly altered, characterized by global hypomethylation and hypermethylation of CpG islands in the promoter regions of tumor suppressor genes[117]. This aberrant methylation pattern silences tumor suppressor genes that regulate critical cellular processes such as the cell cycle, apoptosis, and DNA repair, leading to uncontrolled cell growth and tumor progression[136]. For example, hypermethylation of the promoter regions of the *RB1* gene in retinoblastoma and the *MLH1* gene in colorectal cancer leads to the inactivation of these tumor suppressors, thereby promoting cancer cell survival and proliferation[132].

While promoter hypermethylation is typically associated with gene silencing, gene body methylation often correlates with active transcription[137]. Methylation within gene bodies can enhance transcriptional activity and prevent inappropriate initiation of transcription, contributing to the fine-tuning of gene expression[135,136]. This pattern is frequently associated with specific histone modifications, such as H3K36me3 (trimethylation of lysine 36 on histone H3), which marks actively transcribed regions and facilitates DNMT recruitment, thereby reinforcing transcriptional activity[132].

Epigenetic regulation is also influenced by histone modifications. Histone acetylation typically promotes gene expression by relaxing chromatin (euchromatin), while histone methylation can either activate or repress transcription, depending on the specific mark. Repressive marks like H3K27me3 (trimethylation of lysine 27 on histone H3) are often associated with condensed chromatin states (heterochromatin). Together with DNA methylation, these histone modifications regulate gene accessibility and expression in

cancer, often in concert with Polycomb repressive complexes that maintain repressive chromatin states[138].

To explore DNA methylation patterns in detail, various DNA methylation assays have been developed. Array-based technologies, like the Illumina Infinium microarrays[139,140], are widely used for their cost-effectiveness and extensive coverage of CpG sites, enabling precise detection of methylation at single-CpG resolution. This is achieved through bisulfite-treated DNA, a chemical conversion method where unmethylated cytosines are converted to uracil while methylated cytosines remain unchanged. The treated DNA then hybridizes to specific probes designed to distinguish between methylated and unmethylated CpG sites[139,140]. DNA methylation levels are represented as beta values, ranging from 0 (fully unmethylated) to 1 (fully methylated), which quantify the ratio of the methylated probe signal intensity to the combined total signal intensity at the CpG locus, providing a biologically interpretable quantification of methylation status.

To validate array findings, targeted methods such as bisulfite sequencing of multiple clones[141] and pyrosequencing[142] provide high-accuracy measures of methylation status at specific sites. Whole-genome bisulfite sequencing[143] offers comprehensive, high-resolution data across the entire genome, though it is more expensive and resource-intensive. To balance cost and coverage, reduced representation bisulfite sequencing[144] focuses on CpG-rich regions. As technology advances, single-cell DNA methylation sequencing has emerged[145], enabling high-resolution mapping of methylation patterns and offering unprecedented insights into cellular heterogeneity in tumors. While this approach provides a deeper understanding of cell-specific epigenetic regulation, it remains expensive and requires sophisticated computational analysis[146].

The targeting of these epigenetic modifications has led to the development of epigenetic drugs for cancer treatment, particularly in hematologic malignancies. DNMT inhibitors such as 5-azacitidine and decitabine are approved for treating myelodysplastic syndromes and acute myeloid leukemia, reactivating silenced tumor suppressor genes[147,148]. Additionally, histone deacetylase inhibitors like vorinostat and romidepsinare are used to treat cutaneous T-cell lymphoma by promoting an open chromatin structure and gene reactivation[149–151].

The study of DNA methylation and its role in cancer not only deepens our understanding of tumorigenesis but also offers promising avenues for early detection and treatment. By targeting aberrant epigenetic modifications, therapies can reactivate key regulatory genes, offering a promising approach to cancer treatment.

1.7 Hematologic malignancies

Building upon advancements in cancer treatment, both acute lymphoblastic leukemia (ALL) and B-cell non-Hodgkin lymphoma (B-NHL) have been central to the development of novel treatment strategies for hematologic malignancies[152,153]. ALL, the most common pediatric cancer, represents about 25% of childhood malignancies[154] and is characterized by significant genetic heterogeneity. This includes various somatic mutations and chromosomal rearrangements leading to the malignant transformation of B-cell (B-ALL) or T-cell (T-ALL) progenitors, resulting in uncontrolled proliferation of immature lymphoid cells in the bone marrow. B-ALL constitutes around 80% of cases, while T-ALL accounts for approximately 20% of ALL cases[152]. Advances in chemotherapy, including risk-adjusted protocols and monitoring of measurable residual disease (MRD), have improved outcomes significantly, particularly in children, where 5-year survival rates exceed 90%[152,155].

The genetic complexity of ALL has driven the use of NGS and multi-omics approaches, integrating genomic, transcriptomic, epigenomic, and single-cell analyses[156]. These methods have refined ALL subtypes and facilitated personalized treatment strategies. Epigenetic profiling, especially the analysis of DNA methylation patterns, has been crucial in understanding ALL's pathobiology. Unlike most cancers, ALL is marked by significant CpG island hypermethylation and minimal global hypomethylation, reflecting its highly methylated genome[157], which may be attributed to the higher baseline methylation levels in younger patients[158]. Key epigenetic alterations, such as *KMT2A* rearrangements and *TET2* promoter hypermethylation, are associated with poor prognosis and highlight potential therapeutic targets for DNMTs inhibitors[159].

NHL accounts for approximately 7% of pediatric cancers, with nearly 90% of these cases classified as B-NHL, which includes diverse subtypes that originate from the clonal expansion of B-cells and display significant genetic and epigenetic diversity[153]. Diffuse large B-cell lymphoma (DLBCL) is the most common and aggressive form, comprising

30-40% of B-NHL cases. DLBCL typically responds well to initial immunochemotherapy, with about two-thirds of adult patients and more than 80% of pediatric patients achieving long-term remission[160].

Despite good overall survival rates in both B-ALL and B-NHL, particularly in pediatric patients, the outcomes for relapsed or refractory (R/R) cases remain poor, representing a significant unmet clinical need. These patients face limited treatment options and lower survival rates, underscoring the urgent need for novel therapeutic approaches. In recent years, immunotherapies, including monoclonal antibodies[161,162], antibody-drug conjugates[163,164], and bispecific T-cell engagers[165,166], have provided new treatment options for both ALL and B-NHL patients.

Notably, a transformative advancement in the treatment of both B-ALL and B-NHL has been the development of chimeric antigen receptor (CAR) T-cell therapies[167,168]. These therapies have demonstrated remarkable success in R/R cases, offering a promising new avenue for patients who have not responded to conventional treatments.

1.7.1 CAR T-cell therapy

CAR T-cell therapy has become a groundbreaking treatment for R/R B-ALL and B-NHL, addressing significant unmet clinical needs. With complete response (CR) rates ranging from 70% to 90% in R/R B-ALL and around 50% in R/R DLBCL, CAR T-cell therapy provides new hope to patients who have exhausted conventional treatment options[152,153]. However, 30% to 60% of patients still experience relapse, underscoring the need for further research to optimize long-term success[169]. Given the variability in patient outcomes, this prompts critical questions: What drives differential responses, and how can we predict durable remission?

A deeper understanding of the epigenetic landscape of CAR T-cells may offer valuable insights and reveal novel biomarkers of response. The epigenetic landscape, particularly DNA methylation, plays a key role in T-cell differentiation and function, making it a promising area for uncovering the determinants of effective CAR T-cell therapy This thesis aims to explore these questions by profiling the DNA methylation landscape of CAR T-cells, seeking to identify biomarkers that can predict clinical outcomes and

differentiate responders from non-responders, with findings presented in detail in **Chapter II**.

1.7.1.1 Overview of CAR T-cell therapy

CARs are engineered receptors designed to provide immune effector cells, typically T-cells, with a specific antigen-targeting ability. These receptors enhance T-cell function, enabling them to recognize and attack tumor cells. After infusion into the patient, CAR T-cells engraft, proliferate extensively, and promote immune surveillance by targeting and eliminating cancer cells[170].

CARs consist of an extracellular antigen-binding domain, a single-chain variable fragment derived from immunoglobulin heavy and light chain regions, providing the receptor with its targeting specificity. This domain is fused to a spacer and transmembrane region that anchors the CAR to the T-cell surface. The intracellular portion contains costimulatory domains, such as CD28 and 4-1BB, which enhance T-cell activation, and the CD3ζ chain, which triggers T-cell signaling, driving proliferation and persistence[170]. Notably, CAR T-cells recognize antigens independently of human leukocyte antigen, allowing broader patient application.

The manufacturing process for CAR T-cells begins with the collection of autologous peripheral blood mononuclear cells from the patient via unstimulated leukapheresis. T-cells are then enriched using magnetic beads or density-based methods. Following enrichment, the T-cells are activated with antibodies targeting CD3 and CD28, mimicking natural T-cell activation signals, and then transduced *in vitro* using a lentiviral or gamma-retroviral vector, which encodes the CAR construct. These vectors integrate the CAR gene into the T-cells' genome, allowing them to express the engineered receptor. Once transduced, the T-cells undergo *ex vivo* expansion using cytokines, typically IL-2, or a combination of IL-7 and IL-15, which enhance T-cell proliferation and survival. Lastly, the cells are expanded to reach the necessary therapeutic dose, after which they are infused back into the patient to target tumor cells. This general protocol[*], with minor modifications, has been employed over the last decade in numerous clinical trials and for commercial CAR T-cell production[171,172].

[*] This protocol is illustrated in detail in Figure 4, presented later in the discussion of Chapter II results.

CAR T-cells targeting CD19 were the first to demonstrate potent efficacy in early clinical trials for patients with R/R B-cell malignancies[173,174]. Currently, three autologous CAR T-cell therapies are FDA-approved for patients with R/R DLBCL after two prior lines of systemic therapy: axicabtagene ciloleucel[167], tisagenlecleucel[175], and lisocabtagene maraleucel[176]. These therapies represent a paradigm shift, establishing CAR T-cells as the new standard of care for transplantation-eligible DLBCL patients in early first relapse[153]. Furthermore, CD19-targeted CAR T-cell therapies have also been approved for R/R B-ALL, providing a standalone treatment option or a bridge to allogeneic hematopoietic stem cell transplantation[152]. In particular, tisagenlecleucel[168] and brexucabtagene autoleucel[177] have received FDA approval for treating pediatric and adult patients with R/R B-ALL, further expanding the use of CAR T-cells in B-cell malignancies.

While CAR T-cell therapy has shown significant promise, it is also associated with notable adverse effects, including B-cell aplasia, cytokine release syndrome (CRS), and immune effector cell-associated neurotoxicity syndrome (ICANS)[178,179]. B-cell aplasia occurs as a result of CD19-targeting therapies eliminating both malignant and healthy B-cells, which increases the risk of infections. However, this condition is typically manageable with immunoglobulin replacement therapy and generally resolves once CAR T-cells are ablated[170].

CRS is a systemic inflammatory response characterized by fever, hypotension, hypoxia, and elevated cytokines such as IL-6 and IFN-γ. Its severity often correlates with tumor burden, and in severe cases, CRS can progress to life-threatening complications. It is typically managed with the IL-6 receptor antagonist tocilizumab[170,180]. ICANS often occurs alongside CRS, manifesting as neurological symptoms, including confusion, headache, and in severe cases, seizures and cerebral edema. Management of ICANS involves corticosteroids to reduce inflammation, along with intensive supportive care in more severe cases[180].

Given these adverse effects, and the high costs and complexity of the manufacturing of CAR T-cell therapy, assessing long-term outcomes is crucial. Early studies in R/R B-NHL patients reported CR rates of around 55%, with approximately 60% of these patients remaining in remission at five years[169,181]. Similarly, long-term data from CD19-targeted

CAR T-cell therapy in R/R B-ALL patients show high initial CR rates, particularly in pediatric cases (~80%)[182], and adults (~65%)[183]. This underscores the superior outcomes in pediatric B-ALL patients. However, despite these high CR rates, long-term remission remains challenging, even in pediatric patients, where fewer than 50% achieve long-term event-free survival[169,182]. Thus, while B-ALL patients are more likely to achieve a CR compared to those with B-NHL, fewer patients sustain remission without additional therapies.

1.7.1.2 Biomarkers of CAR T-cell therapy response

The success of CAR T-cell therapy in R/R B-NHL and B-ALL is promising, but achieving durable remissions remains a challenge[169]. Identifying biomarkers associated with long-term outcomes is crucial for refining treatment strategies. While various factors influence therapeutic efficacy[169], recent research suggests that epigenetic profiling could provide new insights into patient responses, offering a more comprehensive understanding of CAR T-cell function[184].

Epigenetic changes, particularly DNA methylation, play a pivotal role in T-cell differentiation and functional state. These modifications are more reflective of long-term cell fate than transient transcriptional changes, offering a stable phenotypic identity[185]. In T-cell responses, DNA methylation defines effector and memory programs that contribute to long-term immune function[186]. During acute immune responses, naïve T-cells undergo extensive epigenetic reprogramming, facilitating cytotoxic function and the development of memory cells[187,188]. In chronic antigen exposure, T-cells can acquire new methylation patterns leading to functional exhaustion, highlighting the importance of studying these changes in CAR T-cell therapies to enhance treatment durability[189].

Moreover, chromatin accessibility and histone methylation play important roles in T-cell differentiation, particularly in response to cancer and chronic infections[190]. Targeting epigenetic regulators like DNMTs and chromatin remodeling complexes has been shown to alter T-cell behavior, further emphasizing the therapeutic potential of modulating these pathways[191,192]. Thus, incorporating epigenetic profiling, especially DNA methylation, into CAR T-cell therapy offers a chance to identify new biomarkers and optimize therapeutic outcomes by better understanding the mechanisms that govern T-cell function.

While epigenetic factors play a significant role, other clinical and biological factors are also important to the success of CAR T-cell therapy[169,184]. The depth of the initial response to treatment, typically assessed through MRD-negativity within the first few months post-infusion, is a critical predictor of long-term success. While early responses are crucial, relapses can still occur, suggesting that further biomarker identification is essential to fully understand long-term efficacy. Additionally, the type of malignancy and baseline tumor burden play significant roles in predicting response durability. Pre-infusion lymphocyte-depleting chemotherapy and high post-infusion CAR T-cell levels also correlate strongly with improved long-term outcomes[169].

The characterization of the CAR T-cell infusion product through genomic and transcriptomic analyses has identified specific T-cell subsets, such as less-differentiated naïve and central memory T-cells, being associated with better responses[193,194]. However, epigenetic profiling of these cells has been largely overlooked. Viral transduction used to introduce CAR constructs can alter the DNA methylation landscape, potentially impacting clinical outcomes by disrupting genes like *TET2*, which regulate T-cell differentiation[195,196]. Moreover, methylation of the viral vector promoter region, as seen in other adoptive cell therapies, could also lead to silencing of CAR transgene expression[197]. By incorporating DNA methylation analysis, we can gain deeper insights into these phenomena, providing a more complete understanding of the factors influencing CAR T-cell persistence and efficacy, while also helping guide modifications to enhance therapeutic efficacy.

2. COVID-19

The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been a defining global health crisis, rapidly spreading since its identification in December 2019 in Wuhan, China[198]. With over 700 million confirmed cases and 16 million deaths attributable to COVID-19 in just the first two years (2020 and 2021), its rapid transmission led to unprecedented social, economic, and healthcare disruptions[199]. The pandemic resulted in the most severe drops in life expectancy seen in over 50 years, reversing decades of global health progress. Life expectancy declined in 84% of countries and territories during this time[199]. Excess mortality rates, coupled with

long-lasting health sequelae such as long COVID, affecting more than 10% of those infected[200], underscore the persistent burden of the disease on health systems worldwide.

The swift global response to the pandemic, particularly through vaccine development, highlighted the critical role of biomedical research. Within weeks of the first outbreak, the virus genome was sequenced, enabling the rapid development of diagnostic tools, vaccines, and treatments[201]. As a result of both vaccination and infection, population immunity significantly increased. By the end of 2022, an estimated 97% of people aged 16 years or older had infection or vaccination induced antibodies, leading to a 47% decrease in the age-adjusted COVID-19 associated death rate, from 115.6 per 100,000 persons in 2021 to 61.3 per 100,000 in 2022[202]. NGS technologies, supported by bioinformatics analysis tools, played a key role in tracking viral mutations, understanding pathogen behavior, and monitoring emerging variants such as Delta and Omicron. This real-time genomic surveillance was essential for guiding public health decisions, underscoring the importance of advanced bioinformatic and molecular biology tools in mitigating global health emergencies[201].

Vaccines, developed at unprecedented pace (**Figure 1**), relied on messenger RNA (mRNA) technology, a breakthrough in immunotherapy that leveraged decades of research in fields such as virology, genomics, and immunology. Collaborative efforts between governments, scientists, and pharmaceutical companies enabled the rapid production and scalability of vaccines, which were distributed at an extraordinary speed and saved millions of lives[203]. Despite these advances, the pandemic exposed significant gaps in global health infrastructure, particularly in the equitable distribution of vaccines. This experience underscored the need for sustained investment in biomedical research and the establishment of robust systems for global preparedness. Consequently, important lessons learned include the importance of building rapid-response frameworks, ensuring diverse participation in clinical trials, and fostering ongoing international collaboration to effectively combat future pandemics.

2.1 SARS-CoV-2 pathophysiology

The pathophysiology of SARS-CoV-2 infection reveals the complexity of COVID-19 impact on the human body, ranging from mild respiratory symptoms to severe systemic complications. SARS-CoV-2 primarily infects the respiratory tract, with initial viral entry

occurring in the upper airways. In most individuals, the infection remains limited to the upper respiratory tract, leading to mild or moderate symptoms such as cough, fever, and fatigue. However, in severe cases, particularly among those with risk factors like age, obesity, or pre-existing conditions, the virus spreads to the lower respiratory tract, contributing to the development of acute respiratory distress syndrome (ARDS), which is characterized by severe lung inflammation, diffuse alveolar damage (DAD), and hypoxemia[204].

Beyond lung involvement, SARS-CoV-2 infection can result in systemic effects, including coagulopathy, which leads to microvascular thrombosis in small lung vessels and other organs, further worsening outcomes. Severe cases are marked by an overactivation of the immune system, which contribute to widespread inflammatory response and tissue damage, underscoring the multi-organ nature of severe COVID-19, and leading to complications such as lung fibrosis, which remains a significant long-term concern[200].

2.2 COVID-19 in children

While adults are more likely to experience severe respiratory complications, children generally exhibit milder symptoms in response to COVID-19, which is unusual for a respiratory disease[205,206]. This difference is thought to be due to distinct immune responses between adults and children.

Children mount a more robust innate immune response, particularly in the nasal mucosa, which helps clear the virus before it progresses to the lower respiratory tract. This enhanced innate response may result from trained immunity due to more frequent respiratory infections in childhood, leading to higher baseline immune activity in children compared to adults[205,206]. Additionally, children exhibit a more tempered adaptive immune response, which may help avoid the hyperinflammation commonly seen in adults that contributes to severe outcomes like ARDS. For instance, children have an increased frequency of naïve T-cells, and a lower frequency of cytotoxic T-cells compared to adults, potentially preventing the overactive adaptive immune response characteristic of severe disease in adults[206].

Together, these differences in immune response contribute to the milder clinical course of COVID-19 in most pediatric cases[207]. However, despite their generally mild illness, some children develop a rare but severe post-infectious condition known as multisystem inflammatory syndrome in children (MIS-C), which can occur several weeks after infection, even if the initial infection was mild or asymptomatic[208,209]. The mechanisms behind why most children experience mild COVID-19 while some develop MIS-C remain unclear, posing significant questions about the variability in immune responses.

While much has been learned about the immune response in children with COVID-19, critical gaps remain in understanding the mechanisms behind severe post-infectious conditions such as MIS-C. What drives hyperinflammatory responses in some children, while others experience only mild illness? What factors contribute to this rare but severe syndrome? These questions are essential for improving diagnostic and therapeutic approaches to MIS-C. This thesis aims to address these unanswered questions by investigating the epigenetic landscape of immune cells in children affected by MIS-C. By profiling DNA methylation patterns, we seek to uncover biomarkers and molecular signatures that may explain the variability in disease severity and the development of MIS-C. The findings and their implications are presented in detail in **Chapter III**.

2.2.1 MIS-C

MIS-C is a severe post-infectious hyperinflammatory syndrome linked to SARS-CoV-2 infection, first reported in April 2020[208,209]. The formal definition of MIS-C includes individuals under 21 years of age presenting with fever, laboratory evidence of inflammation, hospitalization due to the involvement of at least two organ systems, and a confirmed SARS-CoV-2 infection or known exposure within four weeks prior to symptom onset, with no alternative plausible diagnosis[210].

MIS-C, though rare, reached a peak incidence of nearly 7 cases per million person-months by April 2021[211]. MIS-C has a mortality rate of approximately 2%, with over 60% of cases requiring intensive care[210,211]. One of the unique challenges in understanding MIS-C is its clinical and immunological overlap with Kawasaki disease (KD), a pediatric inflammatory vasculitis. While both conditions share symptoms like fever, rash, and multisystem involvement, MIS-C more frequently affects older children (6–12 years) and is more commonly associated with gastrointestinal and cardiovascular symptoms,

compared to KD, which predominantly affects children under five and is more prevalent in East Asian populations[210].

Although the incidence of MIS-C has decreased with COVID-19 vaccination[205], the underlying drivers of hyperinflammation remain poorly understood. Genetic factors that impair immune regulation have been proposed as potential risk contributors[210,212]. However, the specific genetic factors that predispose certain children to develop MIS-C after SARS-CoV-2 exposure remain undetermined.

Given the significant role of DNA methylation in regulating immune system homeostasis and responses to viral infections, we propose profiling these epigenetic changes as a promising approach to uncover the mechanisms driving MIS-C. Methylation is pivotal for immune cell differentiation and may help explain the hyperinflammatory state observed in MIS-C patients. Moreover, this approach has already been used to identify susceptibility loci for respiratory failure in COVID-19 patients, providing insights into why some individuals are more prone to severe outcomes[213].

In addition, other mechanisms have been suggested for MIS-C pathogenesis. For example, it has been hypothesized that the SARS-CoV-2 spike protein may act as a superantigen, triggering non-specific T-cell activation and causing a hyperinflammatory state[214]. Specific autoantibodies have also been detected in children with MIS-C, suggesting that molecular mimicry could contribute to the hyperinflammation[215], although their exact role in disease progression remains unclear. Current treatments such as intravenous immunoglobulin and corticosteroids have proven effective in managing MIS-C, but the underlying causes of the syndrome remain elusive[210].

Future studies must resolve the paradox of why previously healthy children, often with mild or asymptomatic COVID-19 infections, later develop MIS-C. Our work aims to contribute to this effort by identifying biomarkers of hyperinflammation, which could aid in predicting its onset and guiding more targeted therapeutic approaches. By providing a deeper understanding of the DNA methylation profile associated with MIS-C, we aim to help explain why some children develop the syndrome while others do not.

2.3 Severe COVID-19

While most individuals infected with SARS-CoV-2 exhibit mild symptoms or remain asymptomatic, a subset develops severe respiratory failure, driven primarily by lung damage and ARDS. At the core of this severe lung pathology is DAD, a hallmark of severe COVID-19[204]. In some cases, persistent lung lesions are linked to prolonged clinical symptoms, such as in long COVID[200], highlighting the urgent need to understand the mechanisms driving DAD progression.

The drivers of severe COVID-19, particularly those leading to fibrosis and long-term lung damage, remain incompletely understood. Approaches such as bulk transcriptional analyses[216], and more recently, single-cell RNA sequencing[217], have provided valuable insights into the disease. However, these methods disrupt the spatial organization of tissues, which limits our ability to study the critical interactions between cells within their native context. To fully comprehend the complexity of lung injury and tissue remodeling in COVID-19, it is essential to preserve tissue architecture. ST addresses this need by enabling the study of cellular communication while maintaining tissue structure[218,219].

In the context of DAD, ST offers invaluable insights into the spatial distribution and interaction of different cell types, such as macrophages, fibroblasts, and alveolar type 2 cells, which are key players in lung damage[204]. By charting these interactions, ST enables the identification of molecular pathways and cell-cell communication networks that drive tissue remodeling and fibrosis in DAD, which progresses through two overlapping phases: the acute phase, marked by alveolar edema and epithelial and endothelial cell death, and the proliferative phase, characterized by alveolar type 2 cell hyperplasia, fibroblast proliferation, and tissue remodeling that leads to fibrosis[204,220].

These findings raise important questions: What are the cellular dependencies that facilitate the transition from acute to proliferative DAD? How do specific cell types and their spatial interactions contribute to tissue remodeling and fibrosis? Addressing these questions is essential for understanding the mechanisms of tissue damage in severe COVID-19 and identifying potential therapeutic targets. This thesis aims to bridge these knowledge gaps using ST to uncover the cellular and molecular drivers of DAD progression in fatal COVID-19 cases. The findings are presented in detail in **Chapter IV**.

2.3.1 Characterization of diffuse alveolar damage with spatial transcriptomics

The advancement of ST has provided a transformative tool for studying cellular communication in the context of tissue architecture. This capability is particularly valuable in dissecting complex pathologies like DAD in fatal COVID-19. However, the sophisticated datasets generated by ST require advanced bioinformatics tools to extract meaningful insights[221].

ST technologies are broadly categorized into two main approaches: imaging-based and NGS-based methods. Imaging-based techniques include *in situ* hybridization, which involves the hybridization of mRNA molecules to fluorescently labeled oligonucleotide probes, producing distinct on-off patterns that serve as optical barcodes for RNA identification[219].

In contrast, NGS-based methods involve extracting mRNA molecules from tissue while preserving their spatial information, followed by NGS profiling to generate genome-wide expression data. Spatial information can be preserved either through direct capture and location recording, such as with microdissection or microfluidics, or by ligating mRNAs to spatially barcoded probes on a microarray, as used in technologies like Visium ST[218] (**Figure 3**).

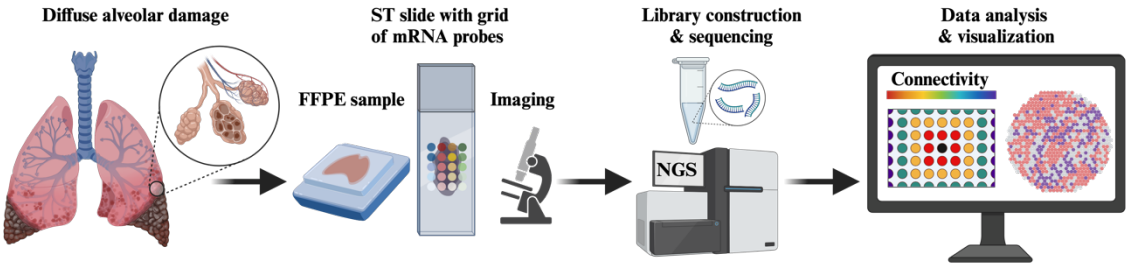


Figure 3. Workflow of Visium spatial transcriptomics data analysis in diffuse alveolar damage. FFPE tissue sections are placed onto Visium ST slides, which contain capture areas with 5,000 spatially barcoded spots to bind mRNA. The sections are stained (e.g., H&E) and imaged, followed by tissue permeabilization to release mRNA, which binds to capture oligonucleotides. The captured mRNA is synthesized into complementary DNA, and sequencing libraries are generated for NGS. Finally, sequencing data is analyzed, and visualizations are created to explore spatial gene expression, using spatial connectivity to gain insights into cellular communication within the native tissue. FFPE: formalin-fixed paraffin-embedded; ST: spatial transcriptomics; mRNA: messenger ribonucleic acid; H&E: hematoxylin and eosin; NGS: next-generation sequencing; Created in BioRender.com.

One of the key bioinformatics challenges in analyzing data from NGS-based methods like Visium ST is the process of deconvolution. This is particularly important because the spatial resolution of Visium (55 μm) means that each capture spot may contain multiple cell types, necessitating computational methods to disentangle mixed-cell populations. Bioinformatic tools have been developed to address this challenge by leveraging single-cell RNA sequencing reference datasets to map cell types onto the spatial grid, providing high-resolution spatial cell type mapping[222]. High-resolution single-cell atlases, like those developed by the Human Cell Atlas[223], are critical to this process, providing comprehensive cell-type profiles that enhance the precision of spatial deconvolution and allow for the accurate mapping of gene expression patterns across tissues.

Beyond mapping cell types, a critical aspect of ST analysis is uncovering cell-cell communication programs. These programs are essential for understanding the complex interactions that drive disease progression, requiring the integration of multiple layers of data to derive comprehensive insights into tissue architecture. To address this, bioinformatics tools have been developed to infer ligand-receptor interactions (LRIs) and cellular dependencies[224]. By functionally characterizing these interactions through pathway enrichment analysis and linking them to downstream intracellular signaling networks via transcription factor (TF) activity[224], a more holistic understanding of tissue remodeling and fibrosis in DAD can be achieved.

To infer cell-cell communication programs, spatially-informed bivariate metrics are utilized to pinpoint co-expressed ligand-receptor pairs[224]. These metrics include commonly used similarity measures like cosine similarity, and a bivariate extension of Moran's *I* called Moran's *R*[225]. When these metrics incorporate spatial connectivity weights, which assign higher weights to closer capture spots (**Figure 3**), they can assess both local and global interactions[224]. Local interactions reflect the spatial co-expression of ligand-receptor pairs between neighboring spots, while global interactions summarize the patterns across the entire tissue section, providing an overview of widespread cellular signaling events.

Moreover, advanced multi-view modeling approaches, such as MISTy[226], integrate spatial information from multiple features, like cell type abundance, LRIs, and TF activity, to reveal complex spatial relationships and dependencies across different

biological contexts. This approach provides a more holistic perspective on cellular communication, complementing local and global spatial metrics by learning relationships between different data layers. This combined approach offers deeper insights into how spatial relationships between cells contribute to DAD progression on both macro and micro scales.

Furthermore, the combination of local LRIs with the use of NMF-based methods can help reveal communication signatures across multiple samples or conditions[224]. These signatures can uncover coordinated signaling events that drive distinct phases of DAD progression, from acute alveolar damage to proliferative fibrosis, offering a comprehensive understanding of tissue dynamics in disease progression.

In summary, by integrating spatial insights from gene expression, cell type abundance, LRIs, pathway enrichment, and TF activity, ST provides a powerful framework for dissecting the molecular and cellular hallmarks of tissue damage, offering promising insights that could guide the development of potential therapeutic interventions in severe COVID-19.

Objectives

Objectives

1. To evaluate the impact of using different variant calling tools on the results of three important downstream analyses in cancer genomics: the detection of cancer driver genes and mutations, the quantification of mutational signatures, and the identification of clinically actionable variants.

2. To characterize the DNA methylation profile of pre-infusion CD19-targeted CAR T-cells and its association with therapy response in relapsed or refractory B-cell malignancies.

3. To characterize the DNA methylation profile in blood linked to multisystem inflammatory syndrome in children.

4. To leverage spatial transcriptomics to investigate the cellular and molecular mechanisms driving the progression of diffuse alveolar damage in fatal COVID-19.

5. To assess the role of bioinformatics in advancing precision medicine while considering the challenges of its clinical application.

Supervisor report

Supervisor report

In accordance with the requirements of the University of Barcelona for a thesis by article compilation, this report acknowledges the impact factor of the journals in which the presented research articles were published, as well as the contribution of the Ph.D. candidate to each publication. The impact factors and journal rankings have been extracted via Clarivate's Journal Citation Reports by Web of Science. The contributions of the Ph.D. candidate, including study design, data analysis, visualizations, and manuscript writing, are recognized as first or co-first authorships in the respective publications.

Chapter I presents the original research article titled **'Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools'**, published in *Bioinformatics*. The journal had a 2022 impact factor of 5.8 and is ranked in Quartile 1 within the *Biochemical Research Methods* category of the SCIE database. The Ph.D. candidate, as the first author of the article, participated in the design of the study, performed data curation, conducted the computational data analysis, created visualizations, and contributed to the drafting and writing of the manuscript.
Publication reference:
Garcia-Prieto, C. A., Martínez-Jiménez, F., Valencia, A. & Porta-Pardo, E. Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. *Bioinformatics* **38**, 3181–3191 (2022).

Chapter II presents the original research article titled **'Epigenetic profiling and response to CD19 chimeric antigen receptor T-cell therapy in B-cell malignancies'**, published in the *Journal of the National Cancer Institute (JNCI)*. The journal had a 2022 impact factor of 10.3 and is ranked in Quartile 1 within the *Oncology* category of the SCIE database. The Ph.D. candidate, as the first co-first author of the article, performed data curation, conducted the computational data analysis, created visualizations, and contributed to the drafting and writing of the manuscript.
Publication reference:
Garcia-Prieto, C. A. *et al.* Epigenetic Profiling and Response to CD19 Chimeric Antigen Receptor T-Cell Therapy in B-Cell Malignancies. *JNCI: Journal of the National Cancer Institute* **114**, 436–445 (2022)

Chapter III presents the original research article titled **'Epigenetic profiling linked to multisystem inflammatory syndrome in children (MIS-C): A multicenter, retrospective study'**, published in *EClinicalMedicine*. The journal had a 2022 impact factor of 15.1 and is ranked in Quartile 1 within the *Medicine, General & Internal* category of the SCIE database. The Ph.D. candidate, as the second co-first author of the article, participated in the design of the study, performed data curation, conducted the computational data analysis, created visualizations, and contributed to the drafting and writing of the manuscript.

Publication reference:

Davalos, V. *et al.* Epigenetic profiling linked to multisystem inflammatory syndrome in children (MIS-C): A multicenter, retrospective study. *EClinicalMedicine* **50**, 101515 (2022).

Chapter IV presents the preprint of the original research article titled **'Spatial transcriptomics unveils the *in situ* cellular and molecular hallmarks of the lung in fatal COVID-19'**, currently under revision at *Nature Communications*. The journal had a 2023 impact factor of 14.7 and is ranked in Quartile 1 within the *Multidisciplinary Sciences* category of the SCIE data base. The Ph.D. candidate, as the first author of the article, participated in the study design, performed data curation, conducted the computational data analysis, created visualizations, and contributed to the drafting and writing of the manuscript.

An earlier version of the preprint is available on *bioRxiv*:

Garcia-Prieto, C. A. et al. Spatial transcriptomics unveils the in situ cellular and molecular hallmarks of the lung in fatal COVID-19. bioRxiv 2024.07.03.601404 (2024).
Available at: https://doi.org/10.1101/2024.07.03.601404.

I confirm that none of the articles presented in this thesis have been used for any other doctoral thesis.

Dr. Manel Esteller
Thesis director
Josep Carreras Leukaemia Research Institute

Chapter I | Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools

Bioinformatics, 38(12), 2022, 3181–3191
https://doi.org/10.1093/bioinformatics/btac306
Advance Access Publication Date: 5 May 2022
Original Paper

OXFORD

Genome analysis

Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools

Carlos A. Garcia-Prieto[1,2], Francisco Martínez-Jiménez[3], Alfonso Valencia[1,4,*] and Eduard Porta-Pardo [1,2,*]

[1]Josep Carreras Leukaemia Research Institute (IJC), Badalona, Spain, [2]Barcelona Supercomputing Center (BSC), Barcelona, Spain, [3]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain and [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Can Alkan

Received on June 3, 2021; revised on February 9, 2022; editorial decision on March 25, 2022; accepted on May 1, 2022

Abstract

Motivation: The analysis of cancer genomes provides fundamental information about its etiology, the processes driving cell transformation or potential treatments. While researchers and clinicians are often only interested in the identification of oncogenic mutations, actionable variants or mutational signatures, the first crucial step in the analysis of any tumor genome is the identification of somatic variants in cancer cells (i.e. those that have been acquired during their evolution). For that purpose, a wide range of computational tools have been developed in recent years to detect somatic mutations in sequencing data from tumor samples. While there have been some efforts to benchmark somatic variant calling tools and strategies, the extent to which variant calling decisions impact the results of downstream analyses of tumor genomes remains unknown.

Results: Here, we quantify the impact of variant calling decisions by comparing the results obtained in three important analyses of cancer genomics data (identification of cancer driver genes, quantification of mutational signatures and detection of clinically actionable variants) when changing the somatic variant caller (MuSE, MuTect2, SomaticSniper and VarScan2) or the strategy to combine them (Consensus of two, Consensus of three and Union) across all 33 cancer types from The Cancer Genome Atlas. Our results show that variant calling decisions have a significant impact on these analyses, creating important differences that could even impact treatment decisions for some patients. Moreover, the Consensus of three calling strategy to combine the output of multiple variant calling tools, a very widely used strategy by the research community, can lead to the loss of some cancer driver genes and actionable mutations. Overall, our results highlight the limitations of widespread practices within the cancer genomics community and point to important differences in critical analyses of tumor sequencing data depending on variant calling, affecting even the identification of clinically actionable variants.

Availability and implementation: Code is available at https://github.com/carlosgarciaprieto/VariantCallingClinicalBenchmark.

Contact: eporta@carrerasresearch.org or alfonso.valencia@bsc.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The exponential growth in both, the generation and access to genomic data from tumor samples and cancer patients, is transforming all aspects of this disease, from basic research to its clinical care (Hyman *et al.*, 2017). For example, thanks to sequencing data, we are beginning to understand the etiology of the mutational processes that affect cancer cells (Alexandrov *et al.*, 2013). Furthermore, we are now able to track and reconstruct the phylogenetic tree of tumor evolution (Nik-Zainal *et al.*, 2012). Similarly, the large cohorts of cancer patients that have been sequenced so far, have helped us

identify germline and somatic mutations that predispose or drive carcinogenesis, respectively (Bailey *et al.*, 2018; Huang *et al.*, 2018), laying the foundations of personalized cancer care.

The first crucial step in analyzing cancer sequencing data is the identification of genetic variants, particularly those of somatic origin. In that sense, the research community has made great efforts to assess the performance of the many different somatic variant callers available (Alioto *et al.*, 2015; Cai *et al.*, 2016; Sandmann *et al.*, 2017; Wang *et al.*, 2013; Xiao *et al.*, 2021; Xu, 2018). However, so far, there has been no agreement on which variant caller, nor strategy to combine them, is the most suitable. For instance, The Cancer Genome Atlas (TCGA) implemented different variant callers on multiple papers throughout its history (Abeshouse *et al.*, 2017; Ciriello *et al.*, 2015; Robertson *et al.*, 2017). This eventually led to the Multi-Center Mutation Calling in Multiple Cancers (MC3) project (Ellrott *et al.*, 2018) to address standardization and reproducibility issues at the end of TCGA. During MC3, many groups worked together to define a clear and unique strategy to combine the output of multiple variant calling tools. Other groups have explored the use of machine-learning approaches to combine the output of different variant calling tools (Anzar *et al.*, 2019; Wood *et al.*, 2018). However, despite all these efforts, it is still unclear which variant calling tool, or combination of tools, is optimal to analyze cancer genomics data.

The biggest challenge in determining the optimal variant calling tool or strategy is the lack of gold standard sets of somatic variants. Another likely important reason is that it is difficult to define a metric in cancer genomics. At the end of the day somatic variant calling is a means to an end, as researchers and oncologists are interested not in the variant calling itself, but rather on the results of downstream analyses. Sequencing data from tumors can be used for many different secondary analyses, from finding cancer driver genes and mutations to determining the presence of clinically actionable mutations or quantifying the effects of mutational signatures. Since none of the somatic variant callers or strategies is perfect, it is possible that the answer to all these secondary analyses differs depending on which somatic variant calling tool or strategy is used.

While there have been benchmarking studies comparing how mutation callers find somatic mutations, to the best of our knowledge there has been no systematic study of the impact on variant calling tools in secondary analyses. In this article, we studied how decisions at the somatic variant calling stage of cancer genomics data affect the results of three different secondary analyses: detection of cancer driver genes and mutations, quantification of mutational signatures and identification of clinically actionable variants.

2 Materials and methods

2.1 Variant calling datasets

To compare the effects of different mutation calling approaches in secondary analyses, we analyzed the entire set of TCGA somatic mutation files comprising 10 189 patients from 33 different cancer types and spanning more than 3 500 000 unique somatic variants. We aimed to explore the impact of different somatic variant calling strategies in downstream analyses of cohorts with different sizes, mutational signatures and mutational burdens. The Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov) gives access to all the processed whole-exome sequencing (WXS) data for all the TCGA projects. In particular, the GDC created the DNA-Seq pipeline (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/) to process all TCGA samples in a uniform way (Grossman *et al.*, 2016). Briefly, this pipeline includes sample preprocessing, alignment to the human reference genome GRCh38.d1.vd1 followed by BAM cleaning and somatic variant calling with variant annotation and aggregation. Somatic variants were identified in WXS data by comparing allele frequencies in matched tumor-normal samples. The GDC used four different variant calling tools to identify somatic mutations: MuSE (Fan *et al.*, 2016), MuTect2 (Cibulskis *et al.*, 2013), SomaticSniper (Larson *et al.*, 2012) and VarScan2 (Koboldt *et al.*, 2012). After analyzing the WXS data for each individual sample, the GDC pipeline includes an aggregation step that combines variants from all cases of a cancer cohort into a single TCGA project mutation annotation format (MAF) file. For a detailed explanation of the GDC DNA-Seq pipeline see Supplementary Methods.

Therefore, for each of the 33 TCGA cancer types, we downloaded the four different Somatic aggregated MAF files with all the somatic mutations for each variant caller (MuSE, MuTect2, SomaticSniper and VarScan2). Additionally, we computed three extra mutation call sets per TCGA project: a Consensus of two variant callers (Consensus2) file with those variants that were called by at least two out of the four aforementioned variant callers, a Consensus of three variant callers (Consensus3) file with those variants that were called by at least three out of the four variant callers and a Union file with every somatic variant called by any variant caller.

2.2 Detecting cancer driver genes

To identify cancer driver genes, we used the IntOGen pipeline (https://bitbucket.org/intogen/intogen-plus/src/master/, March 20, 2020) (Gonzalez-Perez *et al.*, 2013). Specifically, we analyzed every somatic variant file (MuSE, MuTect2, SomaticSniper, VarScan2, Consensus2, Consensus3 and Union) of each of the 33 TCGA projects separately. We did not run IntOGen using PanCancer approaches on all samples combined. IntOGen integrates the result of seven driver discovery methods: OncodriveFML (Mularoni *et al.*, 2016), OncodriveCLUSTL (Arnedo-Pac *et al.*, 2019), dNdScv (Martincorena *et al.*, 2017), CBaSE (Weghorn and Sunyaev, 2017), HotMAPS (Tokheim *et al.*, 2016), smRegions (Martínez-Jiménez *et al.*, 2020a) and MutPanning (Dietlein *et al.*, 2020). The driver discovery methods integrated in IntOGen explore different signals of positive selection, such as clustering of mutations in protein structures or mutational functional bias, to pinpoint which driver genes deviate from the estimated neutral mutation rate using the set of input somatic mutations. The results of these tools are then combined by accounting each method credibility—the relative credibility for each method is based on the ability of the method to give precedence to well-known genes already collected in the Cancer Gene Census (Sondka *et al.*, 2018) catalogue of driver genes—to produce a consensus ranking of genes using a TIER based classification. Finally, IntOGen also provides a weighted combined *P*-value for each ranked gene. For the purpose of our analysis, we only considered true driver genes those within TIER 1 and TIER 2 (q-value <0.05). We, therefore, discarded genes classified in TIER 3 and TIER 4.

2.3 Benchmarking variant calling strategies with driver genes

We considered the curated set of known driver genes from IntOGen (https://www.intogen.org/download, release date February 1, 2020) as our reference set to benchmark how the different mutation call sets can be used to detect cancer driver genes. This set encompasses both, newly detected and previously annotated cancer driver genes in the Cancer Gene Census (https://cancer.sanger.ac.uk/census) of Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes *et al.*, 2017). To further assess and compare our results, we also benchmarked against a second reference set of cancer driver genes published by the PanCancerAtlas-MC3-project (Bailey *et al.*, 2018). We restricted our benchmarking analysis to only those genes annotated as known cancer driver genes in the 33 cancer types we analyzed (MC3 cancer driver genes uniquely identified using PanCancer approaches on all samples combined were not considered). Furthermore, in the case of IntOGen reference set, we only considered those driver genes identified within TCGA cohorts (i.e. driver genes uniquely identified by IntOGen in non-TCGA cohorts, such as ICGC or PCAWG cohorts were filtered out).

We used multiple metrics (Table 1) to assess the performance of the different variant calling strategies when detecting driver genes with IntOGen in downstream analyses. We defined our true positives (TP), false positives (FP) and false negatives (FN) as follows:

Table 1. Benchmarking metrics

Metric	Definition
Precision=TP/(TP+FP)	Also known as positive predictive value. It is the ratio of correctly detected driver genes among all driver genes detected by IntOGen with a given somatic variant call set.
Recall=TP/(TP+FN)	Also known as sensitivity. It is the ratio of correctly detected driver genes by IntOGen among all driver genes within the reference set.
*F*1-score =(2×Precision×Recall)/(Precision+Recall)	Harmonic average of precision and recall. The best value is 1 and the worst is 0.

- TP: those driver genes detected by IntOGen with a given variant call set that are within the reference set.
- FP: those driver genes detected by IntOGen with a given variant call set that are outside the reference set.
- FN: those driver genes within the reference set not identified by IntOGen with a given variant call set.

2.4 Mutational signature analysis

We used deconstructSigs (Rosenthal *et al.*, 2016) 1.8.0 R package to quantify the presence of different mutational signatures in the different mutation call sets. In brief, deconstructSigs accounts for the trinucleotide context of each mutation to classify the six different base substitutions (C > A, C > G, C > T, T > A, T > C and T > G) into 96 possible mutation types (Alexandrov *et al.*, 2013). The signature matrix with the number of times a mutation was found within each trinucleotide context was compared against COSMIC Single Base Substitution (SBS) signatures (available at https://cancer.sanger.ac.uk/signatures/sbs) (Alexandrov *et al.*, 2020).

Finally, deconstructSigs uses an iterative approach to assign different weights to each signature and estimate their contribution to the mutational profile of the tumor sample. We filtered out those samples with <50 mutations. Since we analyzed WXS samples, the signature matrix was normalized to reflect the absolute frequency of each trinucleotide context as it would have taken place in the whole genome. This way we adjusted for differences in trinucleotide abundances between exome and whole genome in order to compare our signatures to the ones extracted from whole genomes (available in synapse.org, ID syn12009743).

2.5 Clinically actionable variants analysis

We used the Molecular Oncology Almanac (Reardon *et al.*, 2021) (https://github.com/vanallenlab/moalmanac, November 4, 2021) (MOAlmanac) to detect alterations that might be therapeutically actionable. Briefly, MOAlmanac is a clinical interpretation algorithm paired with an underlying knowledge base for precision oncology to enable integrative interpretation of multimodal genomic data for point-of-care decision making and translational-hypothesis generation. The primary objective of MOAlmanac is to identify and associate molecular alterations with therapeutic sensitivity and resistance as well as disease prognosis. This is done for 'first-order' genomic alterations (individual events, such as somatic variants) as well as 'second-order' events [those that may be descriptive of global processes in the tumor, such as tumor mutational burden or microsatellite instability (MSI)]. In addition to clinical insights, MOAlmanac annotates and evaluates first-order events based on their presence in numerous other well established datasources as well as highlights connections between them. Overall, MOAlmanac is an open-source computational method for integrative clinical interpretation of individualized molecular profiles.

Since this method is currently geared toward hg19/b37 reference files, we needed to liftover genome coordinates between assemblies for all the Somatic MAFs using CrossMap (Zhao *et al.*, 2014) version 0.3.4 (99.99% of variants were successfully remapped).

2.6 Purity and ploidy dataset

We used purity and ploidy ABSOLUTE annotations (Hoadley *et al.*, 2018) for all TCGA samples available at https://gdc.cancer.gov/about-data/publications/pancanatlas. These annotations were used to adjust the variant allele frequencies (VAFs) by cancer DNA fraction and ploidy to use them in all the analyses.

Almost 97% of TCGA mutation call set cases (9871/10 189 samples) present purity and ploidy information. However, 85% of cases (8673/10 189 samples) match both mutation and purity/ploidy information at the TCGA analyte level (meaning both sources of information come from the same TCGA analyte). Thus, to ensure that the adjusted VAF information presented in our study was sufficiently accurate, we decided to report the adjusted VAF information for this 85% cases. However, when adjusting VAF information at the TCGA analyte level, 1% of variants ended up with adjusted VAFs >1. Therefore, we only used the unadjusted VAFs in our analyses for this 1% of variants and for the variants of the 15% aforementioned cases.

2.7 Clinical metadata

We retrieved tumor stage information from the TCGA-Clinical Data Resource (Liu *et al.*, 2018) file available at https://gdc.cancer.gov/about-data/publications/pancanatlas.

3 Results

3.1 Effects of variant calling in the detection of cancer driver genes

One of the most widespread uses of somatic mutation data from cohorts of cancer patients is the identification of cancer driver genes (Bailey *et al.*, 2018; Martínez-Jiménez *et al.*, 2020b). The tools to detect these genes are sensitive to which somatic mutations are included in the final analysis, as they can bias some aspects of the randomization in which most cancer driver detection tools rely (Arnedo-Pac *et al.*, 2019; Dietlein *et al.*, 2020; Martincorena *et al.*, 2017; Martínez-Jiménez *et al.*, 2020a; Mularoni *et al.*, 2016; Tokheim *et al.*, 2016; Weghorn and Sunyaev, 2017).

To assess to what extent variant calling affects the detection of cancer driver genes, we used IntOGen (Gonzalez-Perez *et al.*, 2013) to find driver genes in 231 different mutation call sets for the 33 different cancer types from TCGA. The seven mutation call sets of each cancer type are distributed as follows: one mutation set with all the calls from one of the four variant calling tools [MuSe (Fan *et al.*, 2016), MuTect2 (Cibulskis *et al.*, 2013), SomaticSniper (Larson *et al.*, 2012) and VarScan2 (Koboldt *et al.*, 2012)], another mutation set—Consensus2—with all those mutations found by, at least, two of the four variant callers, another consensus mutation set—Consensus3—with all those mutations found by, at least, three of the four variant callers and a final mutation set with all the mutations found by any mutation caller—Union (Fig. 1).

One of the main concerns while determining the optimal variant calling tool or strategy is the difficulty to classify mutation calls as TP due to the lack of gold standard sets of somatic variants. The best way to tackle this issue is by experimentally validating the mutation calls with an orthogonal technology. However, in the case of the TCGA somatic call set only 3% of unique somatic variants (110263/3592923) have been validated according to the information in 'GDC_Validation_Status' from the TCGA Somatic MAFs. Therefore, we considered including 'MC3_Overlap' information indicating whether a particular somatic variant overlaps with an MC3 variant for the same sample pair as proxy for bona fide calls. The 87% of unique somatic variants (3127800/3592923) in the

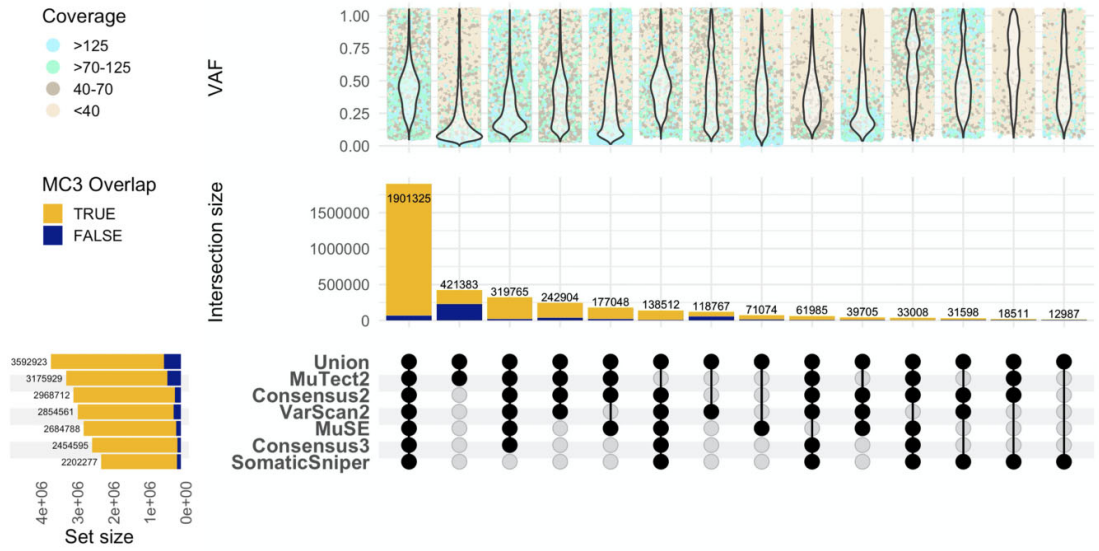


Fig. 1. Intersection of mutation calls across all variant calling strategies for the 33 TCGA cancer types. This UpSetR plot shows the number of variants uniquely identified by one variant calling tool (single point) and variants called by different tools (linked points). Bar-plot indicates intersection size and colors indicate the number of variants present in the PanCancerAtlas MC3 project. Violin plots represent VAF distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Bottom left plot indicates variant call set size

TCGA call set are included in the MC3 project (Ellrott *et al.*, 2018). Furthermore, we included VAF information adjusted by cancer DNA fraction and ploidy to better assess variant calling results. Variant callers tend to perform better when detecting clonal mutations (VAF = 0.5) whereas they struggle to call subclonal ones (VAF < 0.5).

The variant calling results (Fig. 1) show that the somatic mutation call sets from SomaticSniper and MuTect2 were, respectively, the smallest and largest from the individual variant callers. More importantly, 53% of somatic variants were shared among all variant calling strategies spanning a median VAF range around 0.5. Interestingly, MuTect2 uniquely identified 11.7% of all somatic variants, most of them with a very low VAF range. Thus, many of these variants are not included in the MC3 project call set. However, the very high coverage (read depth) across these loci prevents us from discarding these calls as TP and suggests that MuTect2 has high sensitivity to identify subclonal somatic variants.

We wondered whether the different capabilities of the variant calling strategy tools to detect mutations according to their VAF ranges may be clinically related to tumor stage as more advanced tumors tend to be more heterogeneous. However, we were not able to find any correlation in this regard in part due to the high rate of samples with missing American Joint Committee on Cancer (AJCC) stage information.

Having assessed the influence of various tumor properties in the number of mutations called by each tool and combination strategy, we next quantified the effect that they have when detecting cancer driver genes. To that end, we used IntOGen to detect cancer driver genes in the 231 somatic mutation call sets (Fig. 2A and B and Supplementary File 1).

Overall, we found that there are wide differences in the number of detected cancer driver genes in each cohort depending on which somatic variant calling tool or strategy we used. For example, in the case of prostate cancer [prostate adenocarcinoma (PRAD)], the set of mutations from MuTect2 leads to the detection of 33 cancer driver genes, whereas the set from VarScan2 leads to 62 driver genes. Similarly, in the case of bladder cancer [bladder urothelial carcinoma (BLCA)], the Union leads to the detection of 54 cancer driver genes, whereas the set of mutations from MuSE leads to 86 driver genes. Interestingly, the number of cancer driver genes detected in each mutation call set has a positive correlation with the median number of mutations per megabase (spearman rho = 0.56, *P*-value <2.2e-16), as already described in the final driver analysis of TCGA (Bailey *et al.*, 2018). Additionally, the number of cancer driver genes detected in each mutation call set positively correlates with the number of samples in each cohort (spearman rho = 0.36, *P*-value <2.1e-08).

To further assess the possible effects that different sample sizes may have on the performance of specific variant call sets upon detection of cancer driver genes, we conducted a downsampling experiment using the largest TCGA cohort available, the breast invasive carcinoma (BRCA) cohort with 986 samples (Supplementary Fig. S1). To this end, we created three new BRCA cohorts with different sample sizes by subsetting the 25%, 50% and 75% of all BRCA samples, respectively. Furthermore, to select the samples comprising each one of these three newly created BRCA cohorts, we conducted three iterations by selecting different samples for each cohort, obtaining a total of nine different cohorts (three with 25% samples, three with 50% samples and three with 75% samples) to better assess the robustness of the results. While conducting the three different iterations to select the samples, we adjusted for AJCC tumor stage to avoid any confounding effect this variable may have on the results. This analysis confirmed that the number of cancer driver genes detected positively correlates with the number of samples in each cohort (spearman rho = 0.38, *P*-value = 0.0012). Surprisingly, the Consensus3 proved to be the less robust of all strategies with very important differences in the number of cancer driver genes detected within each cohort. For example, in the 50% BRCA cohort ($n = 496$), 62 cancer driver genes were detected with the Consensus3 second iteration call set, whereas only 29 cancer driver genes were detected with the Consensus3 first iteration call set.

Next, we benchmarked our results against a reference set of known cancer driver genes from IntOGen. We also considered the set of cancer driver genes published by the PanCancerAtlas-MC3-project (Bailey *et al.*, 2018) as a second reference set to further assess our results. In both cases, we restricted our reference sets to only those genes annotated as cancer driver genes in the 33 tumor types we analyzed. For the IntOGen reference set, we only considered those cancer driver genes identified within TCGA cohorts. For the MC3 reference set we removed those cancer driver genes

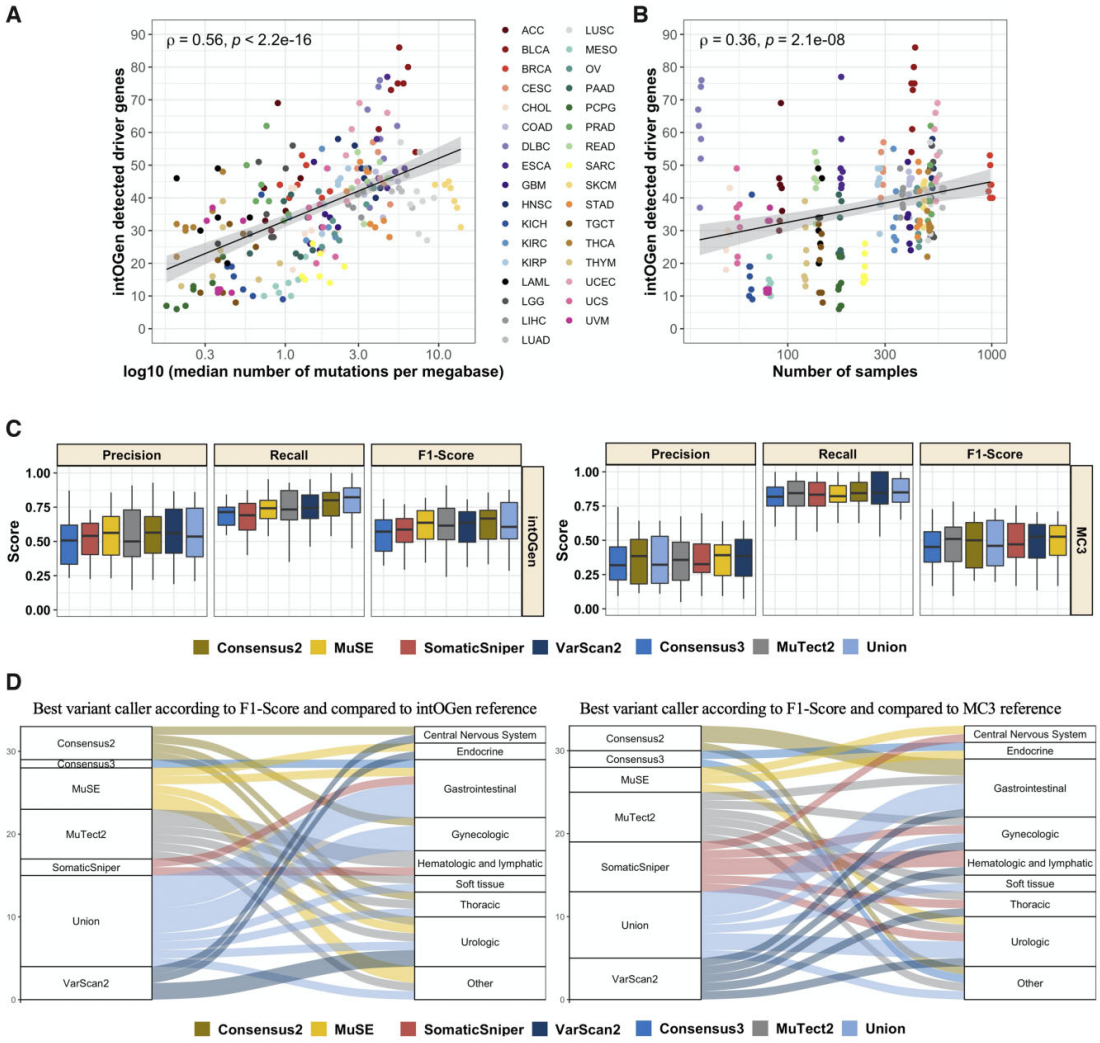


Fig. 2. Performance of different variant calling strategies when detecting cancer driver genes with IntOGen. (**A**) Correlation between cancer driver genes and median number of mutations per megabase. (**B**) Correlation between cancer driver genes and cohort sample size. The number of cancer driver genes detected by IntOGen with different call sets in each cancer type positively correlates with median number of mutations per megabase (A) and sample size (B). Shaded area indicates 95% bootstrapped confidence interval. (**C**) Boxplots represent the different performance metrics scores of the variant calling strategies when detecting cancer driver genes with IntOGen for the 33 TCGA cancer types. Boxplots are sorted by mean metric score. Metric scores when benchmarking against IntOGen (left panel) and PanCancerAtlas MC3 project (right panel) reference sets of known cancer driver genes are shown. (**D**) Alluvial plot indicating best performing variant calling strategy according to *F*1-score for each cancer type when benchmarking against IntOGen (left panel) and PanCancerAtlas MC3 project (right panel) reference sets of known cancer driver genes. *Y*-axis indicates number of cancers in each group

uniquely identified by PanCancer approaches on all samples combined.

The benchmarking results of the 33 cancer types showed, to our surprise, that the Union variant calling strategy is the best one when detecting cancer driver genes with IntOGen and benchmarking against IntOGen reference set (Fig. 2C left panel and Supplementary File 2). Also, when benchmarked against MC3 reference set (Fig. 2C right panel and Supplementary File 2), the Union call set remains the top performer according to recall score, being outperformed by MuSE, VarScan2 and SomaticSniper when looking at *F*1-score and precision results. Interestingly, Consensus3 proved to be amongst the lower performance strategies across all metrics when compared to both reference sets. Consensus2 showed to be pretty robust, being the second-best method when comparing against IntOGen reference set. However, it was outperformed by the Union in all cases. Regarding the four single variant caller performances, it is quite difficult to decide which one is the best one, as their performance depends on the metric and reference set used.

To further assess our results, and considering that Consensus2 performance seemed to be pretty robust, we benchmarked all possible two-caller intersections in a subset of five cancer types: adrenocortical carcinoma (ACC), BLCA, BRCA, PRAD and uterine corpus endometrial carcinoma (UCEC) (Supplementary Fig. S2 and Supplementary Files 3 and 4). According to F1-score metric, Consensus2 outperformed all other possible two-caller intersections when compared against both reference sets. Likewise, SomaticSniper and VarScan2 intersection proved to be the second-best two-caller intersection method.

We next wondered whether certain variant calling strategies are more suitable for specific cancer types. From a clinical point of view, knowing beforehand which variant caller is the best one for a particular cancer or group of cancers would be very helpful and could help inform patient treatment improving the clinical outcome. To this end, we classified all the 33 TCGA cancer types into different groups (Hoadley *et al.*, 2018): hematologic and lymphatic cancers include acute myeloid leukemia (LAML), lymphoid neoplasm diffuse large B cell lymphoma (DLBC) and thymoma (THYM); urologic cancers contain BLCA, PRAD, testicular germ cell tumors, kidney renal cell carcinoma, kidney chromophobe and kidney renal papillary cell carcinoma; gynecologic tumors comprise ovarian (OV), UCEC, cervical squamous cell carcinoma and endocervical adenocarcinoma and BRCA; endocrine cancers include thyroid carcinoma and ACC; central nervous system malignancies contain glioblastoma multiforme and brain lower-grade glioma; gastrointestinal tumors include esophageal carcinoma (ESCA), stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), liver hepatocellular carcinoma, cholangiocarcinoma and pancreatic adenocarcinoma; thoracic tumors contain lung adenocarcinoma, lung squamous cell carcinoma (LUSC) and mesothelioma; soft tissue cancers include sarcoma and uterine carcinosarcoma; finally the remaining cancer types were classified as 'other' including head and neck squamous cell carcinoma, pheochromocytoma and paraganglioma, skin cutaneous melanoma (SKCM) and uveal melanoma.

When analyzing the best variant calling strategy for each cancer type (Fig. 2D and Supplementary Fig. S3 and Supplementary File 2) we observed that the Union is still the best variant calling strategy for the majority of cancer types, especially for gastrointestinal tumors according to *F*1-score. Interestingly, MuTect2 showed very good results being the best variant caller in a variety of cancer types and being the best strategy alongside the Union when considering precision as the metric of interest. Surprisingly, SomaticSniper proved to be the best variant caller for hematologic and lymphatic malignancies, specifically for DLBC and THYM cancer types, but not for LAML malignancies where it was outperformed by other strategies. Consensus2 was the best strategy in the majority of cancer types when considering recall as the metric of interest.

Focusing on the total number of cancer driver genes detected by IntOGen with the different variant calling strategies across the different groups of cancer types (Supplementary Fig. S4 and Supplementary Files 1 and 2), we observed that in most of the cancers (gastrointestinal, gynecologic, urologic and 'other' cancer types) the majority of cancer driver genes detected were shared among all the variant calling strategies. Nevertheless, we found some exceptions in specific cancer types, such in the case of thoracic and hematologic and lymphatic malignancies where SomaticSniper uniquely identified 36 and 28 cancer driver genes respectively, in the latter case most of them from LAML malignancies. Furthermore, Consensus3 was the call set with the largest number of cancer driver genes identified by IntOGen in central nervous system and gastrointestinal cancers, including 41 and 43 uniquely identified cancer driver genes respectively. Overall, our results show important differences in the number and identity of the cancer driver genes detected in a cohort of patients depending on which tool is used to identify somatic variants.

3.2 Somatic mutations in cancer driver genes

Even if one can identify a gene as a driver in a cohort using a variant call set, it is possible that the variant caller misses some individual mutations of that gene in some samples. This could have important implications for patients, as the presence or absence of mutations in cancer driver genes can determine whether patients will receive certain treatments or not (Hyman *et al.*, 2017). To evaluate the impact of variant calling when finding mutations in cancer driver genes, we calculated the number of patients harboring missense and/or nonsense mutations in cancer driver genes depending on the mutation set used (Fig. 3 and Supplementary Fig. S5).

As expected, there is great variability in the detection of somatic mutations in cancer driver genes depending on the variant calling strategy used. Overall, there is a correlation between the total number of mutations called by each method and the number of mutations identified in cancer driver genes (Fig. 3). MuTect2, VarScan2 and, specially, Consensus2 detected more mutations in cancer driver genes than Consensus3 and SomaticSniper. Interestingly, we found that 61% of all missense and nonsense mutations in cancer driver genes were called by all variant callers. Furthermore, 56.5% of all missense and nonsense mutations were found in tumor suppressor genes with 30% of them being nonsense mutations. On the other hand, 37.5% of all missense and nonsense mutations in cancer driver genes were found in oncogenes with 96.5% of them being missense mutations. The remaining 6% of all the mutations affected genes with unknown roles. Importantly, none of the somatic variant call sets (except the Union) had all the mutations in all cancer driver genes, suggesting that we need to use multiple variant callers to ensure that we are detecting all missense and nonsense mutations in cancer driver genes.

We also found important differences when looking at the number of patients bearing at least one missense and/or nonsense mutation in specific cancer driver genes. Specifically, we quantified the number of missed mutations by each variant caller tool or strategy in the four most mutated cancer driver genes (TP53, KRAS, PTEN and PIK3CA) across the 33 cancer types (Supplementary Fig. S5). For example, depending on the variant caller used, up to 22% of UCEC patients (196 patients) differ their PTEN mutational status depending on the variant call set. Similarly, 6% of UCEC patients (32 patients) vary their PIK3CA mutational status when comparing Consensus3 and Union call sets. Importantly, up to 27% of PADD patients (49 patients) carrying a mutation in KRAS could be missing depending on the variant calling strategy used. Finally, regarding samples harboring TP53 mutations, up to 19% of ESCA patients (35 patients), 26% of LUSC patients (128 patients), 35% of OV patients (153 patients) and 20% of READ patients (27 patients) could be missing depending on the variant call set used.

3.3 Mutational signatures

The analysis of mutational signatures is important to understand the biological mechanisms underlying somatic mutations, such as defective DNA repair, mutagenic exposures, DNA replication infidelity or enzymatic DNA modifications. These mutational processes have implications in the understanding of cancer etiology and may inform patient treatment.

We analyzed the mutational signatures of five cancer types—ACC, BLCA, BRCA, PRAD and UCEC—so that they spanned a variety of mutational processes, ranges of purity, mutation rates and cohort sizes within TCGA. For example, ACC is one of the smallest cohorts within TCGA ($n = 92$), as well amongst those with the highest tumor purity (average purity 80%) (Aran *et al.*, 2015). On the other hand, BRCA is the largest cohort in TCGA ($n = 986$). Another factor that can alter the efficiency of tools to detect cancer driver genes is the mutation rate of the cohort, hence why we included UCEC, which is amongst the cancer types with highest mutation rates (Bailey *et al.*, 2018). Finally, BLCA and PRAD are amongst the cohorts that are closest to the TCGA average in all these aspects, making them good representatives of the average tumor sample.

We focused the mutational signatures analysis on those signatures that have been proved to contribute mutations to the corresponding cancer types (Alexandrov *et al.*, 2020) (Fig. 4 and Supplementary File 5). We detected all the expected mutational signatures in all cancer types regardless of the variant calling tool or strategy used. As expected, the mutational signatures contributing the most mutations to individual tumor genomes were SBS1, SBS2, SBS5, SBS13 and SBS40.

We observed SBS5 and SBS40 as flat signatures contributing to multiple types of cancer, although their proposed etiology remains unknown. Furthermore, SBS5, SBS40 and SBS1 mutations have been proved to correlate with age. Specifically, SBS1 may reflect the number of cell divisions a cell has undergone. On the other hand, cancers with high APOBEC activity, specially BLCA and to a lesser extent BRCA, show an increase in the mutational burden of SBS2 and SBS13, both of them related to the APOBEC family of cytidine deaminases activity.

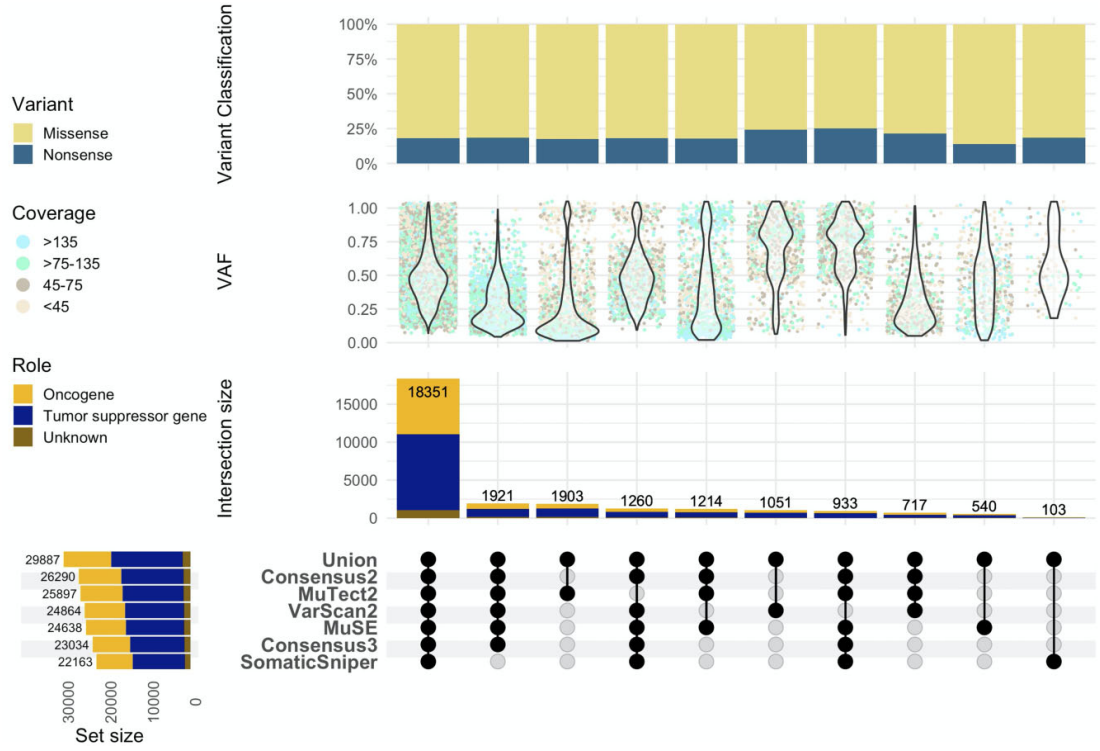


Fig. 3. Detection of somatic mutations in cancer driver genes. This UpSetR plot shows the number of somatic missense and nonsense variants in cancer driver genes uniquely identified by one tool (single point) and by different tools (linked points). Bar-plot indicates intersection size and colors indicate the cancer driver gene role. Violin plots represent VAF distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Top bar-plot indicates the ratio of missense and nonsense mutations. Bottom left plot indicates variant call set size

We found no differences in the quantification of mutational signatures regardless of the variant call set used in any of the five cancer cohorts analyzed. We would like to emphasize that one of the main sources of FP callings are germline mutations in CpG sites that are miscalled as somatic. Hence, the lack of significant differences in SBS1 (characterized by C > T mutations at NCG trinucleotides; N being any base) results is relevant. Overall, it seems that mutational signatures are pretty robust to variant calling decisions.

3.4 Differences in clinically actionable mutations depending on the variant calling strategy

Another important goal of the analysis of somatic cancer genomes is the identification of clinically actionable variants (CAVs). These are somatic variants that help oncologists and physicians decide whether they should give a treatment to a cancer patient, as they are associated with sensitivity, resistance or disease prognosis. Therefore, properly assessing the presence of such variants in the genome of cancer cells is of ultimate clinical importance. To that end, we used the Molecular Oncology Almanac (Reardon *et al.*, 2021) (https://github.com/vanallenlab/moalmanac, November 4, 2021) (MOAlmanac) to identify and associate somatic variants with therapeutic sensitivity and resistance as well as disease prognosis.

We found 36 874 CAVs (Supplementary Fig. S6 and Supplementary File 6) described as biomarkers for a selected tumor type, meaning that the disease for which the association has been reported coincides with the cancer type of the tumor under analysis. These somatic variants are classified according to different levels of clinical actionability or biological relevance depending on how closely they match an alteration–action relationship, as given by catalogued assertions. In total, 6% (2182/36 874) are putatively actionable variants (i.e. exact match between gene, variant classification and protein change with a catalogued variant), 71% (26 214/36 874) are classified as investigate actionability variants (i.e. gene and feature type—somatic variant—match but not either the variant classification or specific protein alteration) and 23% (8478/36 874) are classified as biologically relevant (i.e. gene match only).

Only a little over half of all CAVs were detected by all variant calling strategies (21 198 out of 36 874, 58%). Amongst variant callers, MuTect2 and VarScan2 identified 11% (4084/36 874) of CAVs that were missed by SomaticSniper and MuSE. Moreover, Mutect2 identified an additional 3536 CAVs (10% of all of CAVs). Importantly, all variant callers had some unique CAVs, highlighting the importance of using more than one variant caller when analyzing WXS data to ensure that no CAVs are missed.

MOAlmanac further classifies putatively actionable and investigate actionability somatic variants according to a predictive implication that describes the strength of clinical evidence for a given relationship between a somatic variant and a clinical action. Thus, these variants were matched independently on catalogued events associated with therapeutic sensitivity, therapeutic resistance and disease prognosis with different evidence levels: Food and Drug Association (FDA)-approved (the FDA recognizes an association between the alteration and recommend clinical action); Guideline (this relationship is catalogued as a guideline for standard of care treatment); Clinical trial (the alteration is or has been used as an eligibility criterion for a clinical trial); Clinical evidence (the relationship is reported in a clinical study that did not directly involve a clinical trial); Preclinical evidence (this relationship is reported in a study involving mice, cell line or patient derived models); Inferential evidence (the relationship is inferred as a result of mathematical modeling or an association between molecular features).

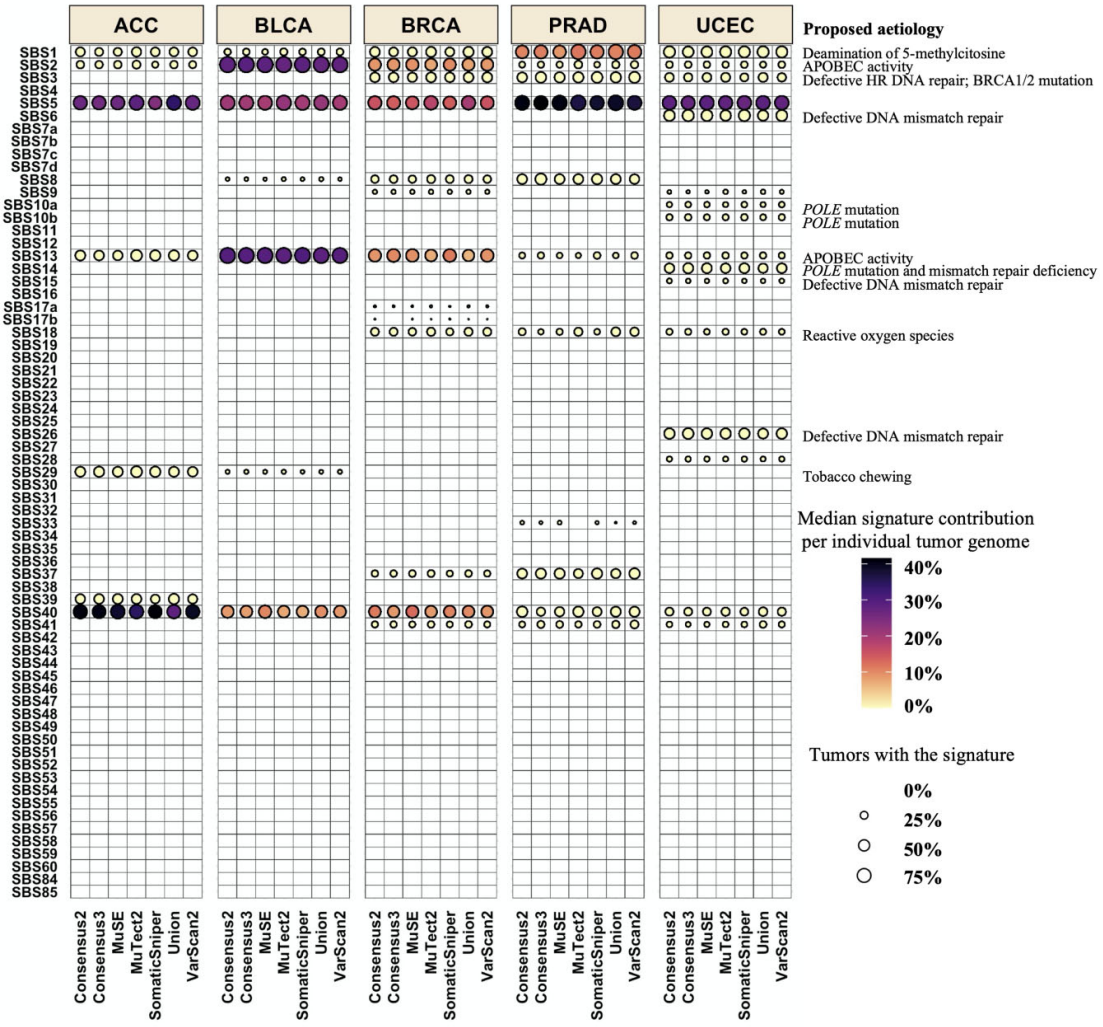


Fig. 4. The percentage of mutations contributed by each mutational signature to individual tumor genomes. The size of each dot represents the proportion of samples of each tumor type that shows the mutational signature. The color of each dot represents the median signature contribution per individual tumor genome in samples that show the signature. Tumors that had few mutations (<50) or that were poorly reconstructed by the signature assignment were excluded. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; PRAD, prostate adenocarcinoma; UCEC, uterine corpus endometrial carcinoma

In total, 5354 variants were associated with therapeutic sensitivity and 514 were associated with therapeutic resistance (Fig. 5 and Supplementary Files 7 and 8). Importantly, 55% (2935/5354) of variants associated with therapeutic sensitivity and 59% (304/514) of variants associated with therapeutic resistance were detected by all variant calling strategies, respectively.

Interestingly, 11.3% (607/5354) of the variants associated with therapeutic sensitivity were found to have FDA-Approved evidence level associations and 27.7% (1483/5354) have a Clinical evidence level. Most of the variants, 44.8% (2396/5354), have a Preclinical evidence level and finally 12.6% (675/5354) have an Inferential evidence level. More importantly, 15.1% (809/5354) were uniquely detected by MuTect2 and VarScan2 (and Consensus2), comprising 12.7% (77/607) of all the variants with FDA-Approved evidence level association. Likewise, MuTect2 uniquely identified 6.3% (38/607) of all FDA-Approved evidence level variants. Finally, very important differences in the detection of clinically actionable variants associated with therapeutic sensitivity were found across variant call sets, with MuTect2, VarScan2 and Consensus2 detecting 21.7% (1163/5354) more variants on average than MuSE, Consensus3 and SomaticSniper.

Furthermore, 869 variants were found to have an association with disease prognosis (Supplementary Fig. S7 and Supplementary File 9) and 71% (617/869) were detected by all variant callers. About 52.2% (454/869) were associated with a favorable prognosis and 47.8% (415/869) with an unfavorable one.

Finally, we looked for clinically actionable variants associated with MSI. This phenotype, MSI, is a hypermutation pattern that occurs at genomic microsatellites caused by impaired DNA mismatch repair. Mismatch repair deficiency that leads to MSI has been described more frequently in colorectal (COAD and READ), endometrial (UCEC) and gastric (STAD) adenocarcinomas (Bonneville *et al.*, 2017; Cortes-Ciriano *et al.*, 2017). Furthermore, it is known that colorectal patients with DNA mismatch repair deficiency have

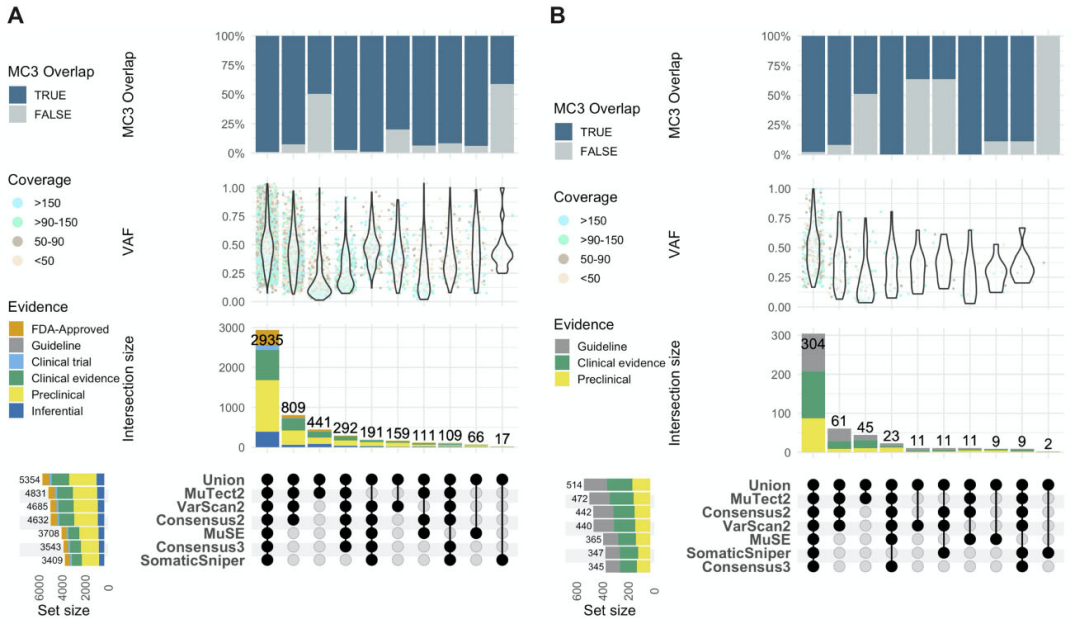


Fig. 5. Clinically actionable somatic mutations associated to therapeutic sensitivity and resistance. This UpSetR plot shows the number of clinically actionable somatic mutations associated to therapeutic sensitivity (**A**) and therapeutic resistance (**B**) detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Single points indicate those variants uniquely identified by one variant call set. Linked points indicate those variants identified by multiple variant call sets. These clinically actionable somatic variants are classified according to different evidence levels. Bar-plot indicates intersection size and colors indicate the association evidence level. Violin plots represent VAF distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Top bar-plot indicates the ratio of variants presents in the PanCancerAtlas MC3 project. Bottom left plot indicates variant call set size. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown

been shown to be more susceptible to immunotherapies, such as programmed cell death (PD-1) immune blockade. Thus, accurate identification of variants associated with MSI is of therapeutic importance.

We found a total of 1276 variants associated with MSI (Supplementary Fig. S8A and Supplementary File 10). In this case, the effect of variant calling strategy is even more significant than for the rest of CAVs, as only 19.5% of all variants (249/1276) where detected by all variant calling strategies. To further assess these important findings, we compared the performance of the different variant calling strategies to identify patients harboring at least one variant associated to MSI. To this end, we selected the four cancer types where MSI has been described more frequently (UCEC, COAD, STAD and READ) and created a reference set of MSI-High (MSI-H) samples described in the literature (Bonneville *et al.*, 2017; Cortes-Ciriano *et al.*, 2017). As expected from previous results, MuTect2, VarScan2 and Consensus2 uniquely identified 69.7% (191/274) of patients with MSI associated variants that were indeed classified as MSI-H samples in the literature (Bonneville *et al.*, 2017; Cortes-Ciriano *et al.*, 2017) (Supplementary Fig. S8B). Only 20% (55/274) of MSI-H patients were detected to bear at least one MSI associated variant with all variant calling approaches. Finally, it is worth mentioning the 49 patients detected by all variant callers that were not classified as MSI-H samples. This is likely due to the fact that we only consider those samples bearing at least one MSI associated variant, which is different from the MSI-H status. For the purpose of the analysis, we considered that MSI-H samples were expected to bear at least one MSI associated variant but not the other way around.

4 Discussion

The analysis of sequencing data from cancer genomes is critical, among others, to understand cancer etiology, identify the events driving the transformation of healthy cells into cancerous ones or guide the treatment of cancer patients (Alexandrov *et al.*, 2013; Bailey *et al.*, 2018; Huang *et al.*, 2018; Hyman *et al.*, 2017; Nik-Zainal *et al.*, 2012). Each of these analyses relies on the proper identification of true somatic variants in the cancer genome, which can be done with many different computational tools. However, we currently do not understand how variant calling approaches impact the final results of cancer sequencing data.

Here, we have quantified the impact of changing variant calling tools or strategies in three different secondary analyses across 33 different cancer types. We have shown that variant calling decisions have no impact on mutational signatures results but, importantly, may lead to significant differences in the identification of cancer driver genes and clinically actionable variants.

While we found no magic recipe, the single recommendation that we believe can be applied in all circumstances is to use, at least, more than one variant calling tool and test the results of any secondary analysis in the different variant call sets. This would give researchers a sense of how much their results might vary depending on the variant calling and whether additional efforts into running other variant calling tools are necessary or not. A useful rule of thumb is to run as many variant callers as possible using the mutations from the Union of all variant calling tools. Taking the mutations from the Consensus of two or more variant callers is the second-best alternative when running multiple variant callers. In the case of running only one variant caller, MuTect2 would be the preferred option in general, albeit we also hope that the detailed results that we provide for the different cancer types in Figure 2D help researchers in deciding which variant caller to use.

Regarding cancer driver genes, while the performance of each variant calling tool or strategy can vary depending on the cancer type, the overall results suggest that one will get the best results using the mutations from the Union of all variant calling tools. The result of the Union variant call set was a surprise, because we initially expected that the likely high number of FP somatic mutations in

the Union call set would lower the predictive power of the cancer driver detection tools in IntOGen, but this was not the case. We believe that this likely reflects the robustness of IntOGen to the presence of FP in the somatic mutation set. Another unexpected finding was that one of the most common approaches to combine somatic variant call sets, Consensus3 (Bailey *et al.*, 2018), had some of the worse overall results when detecting cancer driver genes. On the other hand, Consensus2 showed very robust results overall, being the second-best strategy when considering recall as the metric of interest. Thus, very restrictive methods, such as Consensus3, seemed to badly penalized IntOGen cancer driver genes detection tools. Nevertheless, considering the specific cancer type is important, such is the case of hematologic and lymphatic malignancies like DLBC and THYM, where SomaticSniper proved to be the best caller.

Importantly, we have also found differences in the detection of somatic missense and nonsense mutations in cancer driver genes. In some cases, a specific cancer driver gene mutation status (i.e. PTEN in UCEC) could differ in more than 20% of patients depending on the variant call set used. This result suggests that it is important to use, at least, more than one variant calling tool to analyze cancer genomes. Otherwise, a significant number of mutations in cancer driver genes can be missed. Specially considering that Consensus2 was the strategy that detected more missense and nonsense mutations in cancer driver genes.

Mutational signatures analysis is pretty robust to variant calling decisions. We found no differences in the quantification of mutational signatures across the five cancer types analyzed.

However, if the goal of the analysis of the somatic genome is to find clinically actionable mutations, we need to be aware that there are considerable differences depending on the somatic mutation calling used. Only half (57.5%) of all clinically actionable variants were detected by all variant calling strategies. On average, MuTect2, VarScan2 and Consensus2 detect 20% more clinically actionable variants than MuSE, Consensus3 and SomaticSniper. This trend remains when looking at variants associated to therapeutic sensitivity. Importantly, we found greater differences when detecting of MSI associated variants, with MuTect2, VarScan2 and Consensus2 uniquely identifying 70% of MSI-H samples. Accurately identifying these variants is of therapeutic importance considering their relevance for immunotherapy treatments.

Finally, one of the main sources of variation between variant calling strategies is the identification of subclonal mutations. Here, we included VAF information adjusted by cancer DNA fraction and ploidy, observing that MuTect2 has high sensitivity to identify subclonal somatic variants. However, intra-tumor heterogeneity would be another important factor to consider (Dentro *et al.*, 2021) since the analysis of heterogeneous cancers (i.e. PRAD) would yield more variable results compared to those of homogeneous cancers (i.e. SKCM) (Supplementary Fig. S4).

We acknowledge several limitations in our study. For example, we are not considering results for copy number and structural variants in the mutation call sets. We also have not explored the impact of other important variables, such as sequencing coverage. It is possible that with deeper coverages, such as those provided by targeted sequencing of gene panels, the differences we observed here for the variant callers are smaller.

Overall, we hope this study will help researchers understand how variant calling decisions might impact their results. It is important to account for the clinical implications that variant calling decisions have on different downstream analyses, especially in such important aspects of cancer genomics like driver genes and the identification of actionable variants. Moreover, we hope that this study will help guide variant calling design while considering the needs and goals of the different research projects.

Acknowledgements

We would like to thank the patients that donated the samples for The Cancer Genome Atlas, without them this work would not be possible. We would also like to thank Abel González-Pérez, Collin Tokheim, Brendan Reardon and Eliezer M. Van Allen for their valuable discussions and insights.

Funding

This work was supported by the BSC-Lenovo Master Collaboration Agreement (2015) and the IBM-BSC Joint Study Agreement (JSA) on Precision Medicine under the IBM-BSC Deep Learning Center Agreement (to C.A.G.-P.). F.M.-J. was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [Grant Agreement No. 682398]. A.V. received support from Institució Catalana de Recerca i Estudis Avançats (ICREA). E.P.-P. received support by a La Caixa Junior Leader Fellowship [LCF/BQ/PI18/11630003] from Fundación La Caixa and a Ramon y Cajal fellowship from the Spanish Ministry of Science [RYC2019-026415-I]. The Barcelona Supercomputing Center and IRB Barcelona are recipients of a Severo Ochoa Centre of Excellence Award from Spanish Ministry of Science, Innovation and Universities (MICINN; Government of Spain). The Josep Carreras Leukaemia Research Institute and IRB Barcelona are supported by CERCA (Generalitat de Catalunya). E.P-P. is supported by the Spanish Science Ministry (PID2019-107043R1-I00).

Conflict of Interest: The authors declare that they do not have any conflict of interest.

References

Abeshouse,A. *et al.* (2017) Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, **171**, 950–965.e28.

Alexandrov,L.B. *et al.*; Australian Pancreatic Cancer Genome Initiative. (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

Alexandrov,L.B. *et al.*; PCAWG Consortium. (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.

Alioto,T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.

Anzar,I. *et al.* (2019) NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics*, **12**, 63.

Aran,D. *et al.* (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.

Arnedo-Pac,C. *et al.* (2019) OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics*, **35**, 4788–4790.

Bailey,M.H. *et al.*; Cancer Genome Atlas Research Network. (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.e18.

Bonneville,R. *et al.* (2017) Landscape of microsatellite instability across 39 cancer types. *JCO Precis. Oncol.*, **1**. 1–15.

Cai,L. *et al.* (2016) In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.*, **6**, 36540–36549.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

Ciriello,G. *et al.*; TCGA Research Network. (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**, 506–519.

Cortes-Ciriano,I. *et al.* (2017) A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.*, **8**, 15180.

Dentro,S.C. *et al.*; PCAWG Evolution and Heterogeneity Working Group and the PCAWG Consortium. (2021) Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, **184**, 2239–2254.e39.

Dietlein,F. *et al.* (2020) Identification of cancer driver genes based on nucleotide context. *Nat. Genet.*, **52**, 208–218.

Ellrott,K. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.e7.

Fan,Y. *et al.* (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178. https://doi.org/10.1186/s13059-016-1029-6.

Forbes,S.A. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.

Gonzalez-Perez,A. *et al.* (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.

Grossman,R.L. *et al.* (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.

Hoadley,K.A. *et al.* (2018) Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304.e6.

Huang,K. *et al.* (2018) Pathogenic germline variants in 10,389 adult cancers. *Cell*, 173355–173370.e14.

Hyman,D.M. *et al.* (2017) Implementing genome-driven oncology. *Cell*, **168**, 584–599.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Larson,D.E. *et al.* (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.

Liu,J. *et al.* (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, **173**, 400–416.e11.

Martincorena,I. *et al.* (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, 1029–1041.e21.

Martínez-Jiménez,F. *et al.* (2020a) Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer*, **1**, 122–135.

Martínez-Jiménez,F. *et al.* (2020b) A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*, **20**, 555–572.

Mularoni,L. *et al.* (2016) OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, **17**, 128.

Nik-Zainal,S. *et al.*; Breast Cancer Working Group of the International Cancer Genome Consortium. (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.

Reardon,B. *et al.* (2021) Integrating molecular profiles into clinical frameworks through the molecular oncology almanac to prospectively guide precision oncology. *Nat. Cancer*, **2**, 1102–1112.

Robertson,A.G. *et al.* (2017) Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, **171**, 540–556.e25.

Robertson,A.G. *et al.*; TCGA Research Network. (2017) Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell.*, **32**, 204–220.e15.

Rosenthal,R. *et al.* (2016) deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, **17**, 31.

Sandmann,S. *et al.* (2017) Evaluating variant calling tools for Non-Matched Next-Generation sequencing data. *Sci. Rep.*, **7**, 43169.

Sondka,Z. *et al.* (2018) The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.

Tokheim,C. *et al.* (2016) Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.*, **76**, 3719–3731.

Wang,Q. *et al.* (2013) Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.*, **5**, 91.

Weghorn,D. and Sunyaev,S. (2017) Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.*, **49**, 1785–1788.

Wood,D.E. *et al.* (2018) A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.*, **10**, eaar7939.

Xiao,W. *et al.* (2021) Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.*, **39**, 1141–1150.

Xu,C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.*, **16**, 15–24.

Zhao,H. *et al.* (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.

Supplementary materials

Supplementary materials

Supplementary methods

Here, the GDC DNA-Seq pipeline data processing steps are further explained:

1. Pre-alignment: BAM files are split by read groups and transformed to FASTQ files. Reads failing to pass Illumina chastity test were removed.
2. Alignment: read groups are aligned to the human reference genome GRCh38.d1.vd1 including decoy viral sequences using two BWA algorithms (BWA-MEM if mean read length $\geq$ 70 bp, otherwise BWA-aln was used). Read groups are aligned separately and the ones belonging to a single aliquot are merged using Picard Tools, SortSam and MergeSamFiles. Duplicate reads (PCR artifacts) are flagged.
3. Co-cleaning: uses both tumor and normal matched BAMs to further improve alignment quality. To this end, the base quality score recalibration (BQSR) step is performed to adjust base quality scores based on detectable and systematic errors.
4. Somatic variant calling: variant calling is performed using different separate pipelines: MuSE, MuTect2, SomaticSniper and VarScan2 (Pindel calls were not used in this study). Variant calls are reported separately by each pipeline in a VCF file. Details of each pipeline including command line parameters can be found in the corresponding web referenced above. However, some particularities are worth mentioning:
    a. MuTect2 pipeline employs a "Panel of Normals" generated using TCGA blood normal genomes to filter out additional germline variants.
    b. VarScan2 pipeline uses the SAMtools mpileup utility to filter out reads with a mapping quality < 1.
    c. A False Positive Filter is additionally used to label low quality variants in VarScan2 and SomaticSniper outputs.
5. Annotation: raw VCF files are annotated with the Variant Effect Predictor (VEP) v84. Variants in the VCF files are also matched to known variants from external mutation databases such as GENCODE, dbSNP or ClinVar among others.
6. Aggregation: one MAF file is generated per variant calling pipeline for each TCGA project containing all available cases within this project. Open-access MAF files (known as Somatic MAFs) are modified for public release by removing columns and variants that could contain germline mutation information. The criteria used to remove variants can be found here (https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/). The low-quality variant filtering and germline masking steps are important to consider:
    a. Variants with MuTation_Status != "Somatic" or GDC_FILTER = "Gapfiller", "ContEst", "multiallelic", "nonselectedaliquot", "BCR_Duplicate" or "BadSeq" are removed.
    b. Remaining variants with GDC_Valid_Somatic = True are included in the Somatic MAF.
    c. Remaining variants with FILTER != "panel of normal" or PASS are removed.
    d. Remaining variants with MC3_Overlap = True are included in the Somatic MAF.
    e. Remaining variants with GDC_FILTER = "ndp", "NonExonic", "bitgt", "gdc_pon" are removed.
    f. Remaining variants with SOMATIC != null are included in the Somatic MAF.
    g. Remaining variants with dbSNP_RS = 'novel' or null are included in the Soamtic MAF.
    h. Remaining variants are removed.

Supplementary materials

Supplementary files:

Supplementary file 1: Cancer driver genes detected by intOGen. This file contains the results after running intOGen with all the variant call sets in the complete set of 33 TCGA projects. Additional information about the presence of each cancer driver gene per cancer type in intOGen and PanCancerAtlas MC3 project reference sets has been included. Cancer driver genes detected by intOGen are represented by 1 and those non-detected are represented by 0. In the case of intOGen and MC3 reference sets, 1 stands for those cancer driver genes present in the reference set and 0 for those not present in the reference set.

Supplementary file 2: Performance metrics of the different variant calling strategies when detecting cancer driver genes with intOGen. This file contains all the performance metrics scores of the seven variant calling strategies when detecting cancer driver genes with intOGen across the entire set of 33 TCGA cancer types.

Supplementary file 3: Cancer driver genes detected by intOGen with all possible two-caller intersection call sets. This file contains the results after running intOGen with all possible two-caller intersection call sets in 5 TCGA cancer types: adrenocortical carcinoma, bladder urothelial carcinoma, breast invasive carcinoma, prostate adenocarcinoma and uterine corpus endometrial carcinoma. Additional information about the presence of each cancer driver gene per cancer type in intOGen and PanCancerAtlas MC3 project reference sets has been included. Cancer driver genes detected by intOGen are represented by 1 and those non-detected are represented by 0. In the case of intOGen and MC3 reference sets, 1 stands for those cancer driver genes present in the reference set and 0 for those not present in the reference set.

Supplementary file 4: Performance metrics of all possible two-caller intersection variant calling strategies when detecting cancer driver genes with intOGen. This file contains all the performance metrics scores of all possible two-caller intersection variant calling strategies when detecting cancer driver genes with intOGen across five TCGA cancer types: adrenocortical carcinoma, bladder urothelial carcinoma, breast invasive carcinoma, prostate adenocarcinoma and uterine corpus endometrial carcinoma.

Supplementary file 5: Mutational signature analysis results. This file contains the results of deconstructSigs analysis. The percentage of mutations contributed by each mutational signature to individual tumor genomes across five TCGA cancer types is shown: adrenocortical carcinoma, bladder urothelial carcinoma, breast invasive carcinoma, prostate adenocarcinoma and uterine corpus endometrial carcinoma. Tumors that had few mutations (less than 50) assignment were excluded.

Supplementary file 6: Clinically actionable variants detected by Molecular Oncology Almanac. This file includes the total number of clinically actionable somatic mutations detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Clinically actionable variants detected by the different variant calling strategies are represented by 1 and those non-detected are represented by 0. In the case of 'MC3_Overlap' column, 1 stands for those variants present in the PanCancerAtlas MC3 project reference call set and 0 for those not present. For the 'GDC_Validation_Status' column, 1 stands for those variants that have been validated by a Next-Generation Sequencing orthogonal technology and 0 for those that have not been validated. A description for Molecular Oncology Almanac outputs is available at https://github.com/vanallenlab/moalmanac/blob/main/docs/description-of-outputs.md. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Supplementary file 7: Clinically actionable variants associated to therapeutic sensitivity detected by Molecular Oncology Almanac. This file includes the total number of clinically actionable somatic mutations associated to therapeutic sensitivity detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Clinically actionable variants detected by the different variant calling strategies are represented by 1 and those non-detected are represented by 0. In the case of 'MC3_Overlap' column, 1 stands for those variants present in the PanCancerAtlas MC3 project reference call set

and 0 for those not present. For the 'GDC_Validation_Status' column, 1 stands for those variants that have been validated by a Next-Generation Sequencing orthogonal technology and 0 for those that have not been validated. A description for Molecular Oncology Almanac outputs is available at https://github.com/vanallenlab/moalmanac/blob/main/docs/description-of-outputs.md. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Supplementary file 8: Clinically actionable variants associated to therapeutic resistance detected by Molecular Oncology Almanac. This file includes the total number of clinically actionable somatic mutations associated to therapeutic resistance detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Clinically actionable variants detected by the different variant calling strategies are represented by 1 and those non-detected are represented by 0. In the case of 'MC3_Overlap' column, 1 stands for those variants present in the PanCancerAtlas MC3 project reference call set and 0 for those not present. For the 'GDC_Validation_Status' column, 1 stands for those variants that have been validated by a Next-Generation Sequencing orthogonal technology and 0 for those that have not been validated. A description for Molecular Oncology Almanac outputs is available at https://github.com/vanallenlab/moalmanac/blob/main/docs/description-of-outputs.md. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Supplementary file 9: Clinically actionable variants associated to disease prognosis detected by Molecular Oncology Almanac. This file includes the total number of clinically actionable somatic mutations associated to disease prognosis detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. A description for Molecular Oncology Almanac outputs is available at https://github.com/vanallenlab/moalmanac/blob/main/docs/description-of-outputs.md. Clinically actionable variants detected by the different variant calling strategies are represented by 1 and those non-detected are represented by 0. In the case of 'MC3_Overlap' column, 1 stands for those variants present in the PanCancerAtlas MC3 project reference call set and 0 for those not present. For the 'GDC_Validation_Status' column, 1 stands for those variants that have been validated by a Next-Generation Sequencing orthogonal technology and 0 for those that have not been validated. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Supplementary file 10: Clinically actionable variants associated to microsatellite instability (MSI) detected by Molecular Oncology Almanac. This file includes the total number of clinically actionable mutations associated to microsatellite instability (MSI) detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Clinically actionable variants detected by the different variant calling strategies are represented by 1 and those non-detected are represented by 0. In the case of 'MC3_Overlap' column, 1 stands for those variants present in the PanCancerAtlas MC3 project reference call set and 0 for those not present. For the 'GDC_Validation_Status' column, 1 stands for those variants that have been validated by a Next-Generation Sequencing orthogonal technology and 0 for those that have not been validated. A description for Molecular Oncology Almanac outputs is available at https://github.com/vanallenlab/moalmanac/blob/main/docs/description-of-outputs.md.

Supplementary files can be accessed on the publication site at:
https://doi.org/10.1093/bioinformatics/btac306

Supplementary materials

Supplementary figures:

Figure S1: Impact of sample size on the correlation between cancer driver genes and sample size. The number of cancer driver genes detected by intOGen with different subsampled BRCA call sets positively correlates with sample size. Three different subsampling experiments were performed for the 25%, 50% and 75% of BRCA cohort cases respectively. Three additional iterations for selecting different samples in each subsampled cohort were conducted. Shaded area indicates 95% bootstrapped confidence interval.

Figure S2: Performance of all possible two-caller intersection variant calling strategies when detecting cancer driver genes with intOGen. Boxplots represent the different performance metrics scores when detecting cancer driver genes with intOGen for 5 TCGA cancer types: adrenocortical carcinoma, bladder urothelial carcinoma, breast invasive carcinoma, prostate adenocarcinoma and uterine corpus endometrial carcinoma. Boxplots are sorted by mean metric score. Metric scores when benchmarking against intOGen (top panel) and PanCancerAtlas MC3 project (bottom panel) reference sets of known cancer driver genes are shown.

Figure S3: Best variant calling strategy per cancer type. Alluvial plot indicating best performing variant calling strategy according to (**A**) precision and (**B**) recall for each cancer type when benchmarking against intOGen (left panels) and PanCancerAtlas project -MC3- (right panels) reference sets of known cancer driver genes. Y-axis indicate number of cancers in each group.

Figure S4: Cancer driver genes detected by intOGen. This UpSetR plot show the number and distribution of cancer driver genes detected by intOGen with the different variant call sets in the complete set of TCGA projects. Bar-plot indicates intersection size and colors indicate cancer type. Single points indicate those cancer driver genes uniquely identified by intOGen with one variant call set. Linked points indicate those cancer driver genes identified by intOGen with multiple variant call sets. Bottom left plot indicates the number of cancer driver genes detected by intOGen with each variant call set. We classified all the 33 TCGA cancer types into different groups (Hoadley *et al*., 2018): central nervous system malignancies contain glioblastoma multiforme (GBM) and brain lower-grade glioma (LGG); endocrine cancers include thyroid carcinoma (THCA) and adrenocortical carcinoma (ACC); gastrointestinal tumors include esophageal carcinoma (ESCA), stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), liver hepatocellular carcinoma (LIHC), cholangiocarcinoma (CHOL) and pancreatic adenocarcinoma (PAAD); gynecologic tumors comprise ovarian (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) and breast invasive carcinoma (BRCA); hematologic and lymphatic cancers include acute myeloid leukemia (LAML), lymphoid neoplasm diffuse large B cell lymphoma (DLBC) and thymoma (THYM); thoracic tumors contain lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and mesothelioma (MESO); urologic cancers contain bladder urothelial carcinoma (BLCA), prostate adenocarcinoma (PRAD), testicular germ cell tumors (TGCT), kidney renal cell carcinoma (KIRC), kidney chromophobe (KICH) and kidney renal papillary cell carcinoma (KIRP); finally the remaining cancer types were classified as "other" including head and neck squamous cell carcinoma (HNSC), pheochromocytoma and paraganglioma (PCPG), skin cutaneous melanoma (SKCM) and uveal melanoma (UVM) and soft tissue cancers like sarcoma (SARC) and uterine carcinosarcoma (UCS).

Figure S5: Samples with missense and nonsense mutations in cancer driver genes. This plot shows the percentage of samples for each of the 33 TCGA cancer types bearing at least one missense and/or nonsense mutation in the four most overrepresented cancer driver genes: TP53, PTEN, KRAS and PIK3CA.

Figure S6: Clinically actionable somatic mutations. This UpSetR plot show the total number of clinically actionable somatic mutations detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Single points indicate those variants uniquely identified by one variant call set. Linked points indicate those variants identified by multiple variant call sets. Clinically actionable somatic variants are classified according to different levels of clinical actionability or biological relevance depending on how closely they match an alteration-action relationship, as given by catalogued assertions. Bar-plot

indicates intersection. Violin plots represent variant allele frequency (VAF) distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Top bar-plot indicates the ratio of variants present in the PanCancerAtlas MC3 project. Bottom left plot indicates variant call set size. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Figure S7: Clinically actionable somatic mutations associated to disease prognosis. This UpSetR plot show the number of clinically actionable somatic mutations associated to disease prognosis detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Single points indicate those variants uniquely identified by one variant call set. Linked points indicate those variants identified by multiple variant call sets. These clinically actionable somatic variants are classified according to different evidence levels. The bottom bar-plot indicates intersection size and colors indicate the association evidence level. The next bar-plot indicates the ratio of variants associated to a favorable or unfavorable prognosis. Violin plots represent variant allele frequency (VAF) distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Top bar-plot indicates the ratio of variants present in the PanCancerAtlas MC3 project. Bottom left plot indicates variant call set size. Only those clinically actionable somatic variants in which the disease for which the association has been reported coincides with the cancer type of the tumor under analysis are shown.

Figure S8: Clinically actionable mutations associated to microsatellite instability (MSI). (**A**) This UpSetR plot show the number of clinically actionable mutations associated to MSI detected by the Molecular Oncology Almanac with the different variant calling strategies in the complete set of TCGA projects. Single points indicate those variants uniquely identified by one variant call set. Linked points indicate those variants identified by multiple variant call sets. These clinically actionable variants are classified according to their presence in different databases. The clinically actionable variants present in the Molecular Oncology Database are classified according to different levels of clinical actionability or biological relevance depending on how closely they match an alteration-action relationship, as given by catalogued assertions. The bottom bar-plot indicates intersection size. Violin plots represent variant allele frequency (VAF) distribution adjusted by cancer DNA fraction and ploidy; colors indicate total coverage (read depth) across loci. Top bar-plot indicates the ratio of variants present in the PanCancerAtlas MC3 project. Bottom left plot indicates variant call set size. (**B**) This UpSetR plot show the number of samples bearing at least one clinically actionable mutation associated to MSI detected by the Molecular Oncology Almanac with the different variant calling strategies in the four cancer types where MSI has been described more frequently. A reference set of MSI-High (MSI-H) samples described in the literature (Bonneville et al., 2017; Cortes- Ciriano et al., 2017) was included. Single points indicate those samples uniquely identified by one variant call set. Linked points indicate those samples identified by multiple variant call sets. Bottom left plot indicates number of samples identified with each call set. COAD, colon adenocarcinoma; READ, rectum adenocarcinoma; STAD, stomach adenocarcinoma; UCEC, uterine corpus endometrial carcinoma.

Figure S1

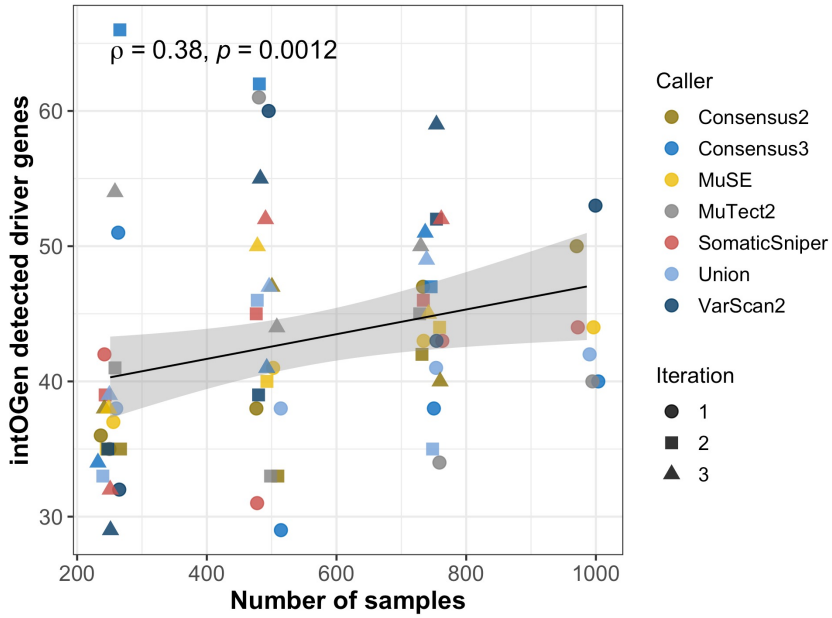


Figure S2

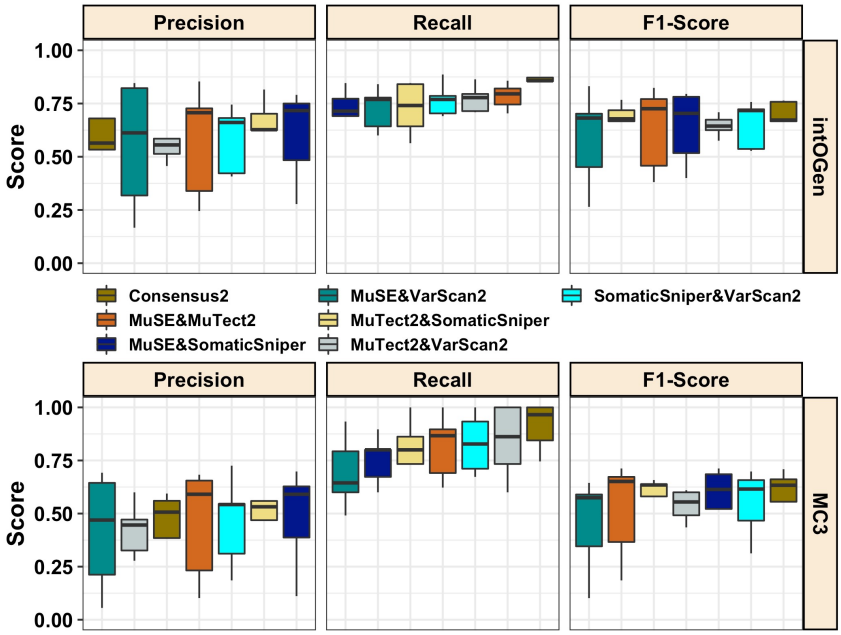
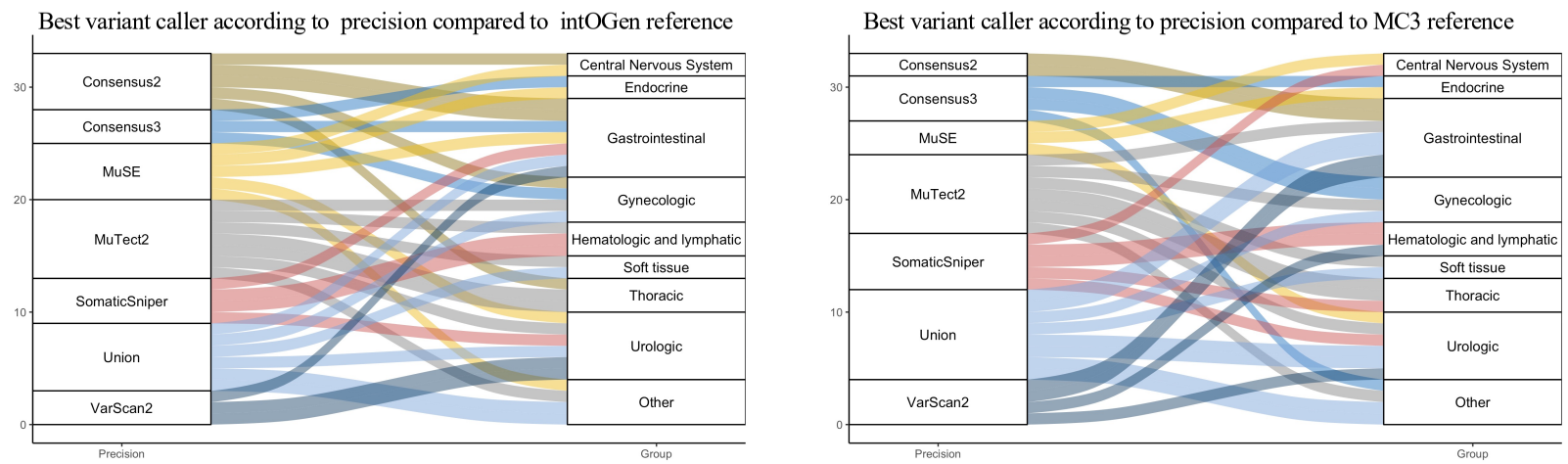


Figure S3

A


Best variant caller according to precision compared to intOGen reference
Consensus2
Consensus3
MuSE
MuTect2
SomaticSniper
Union
VarScan2
Central Nervous System
Endocrine
Gastrointestinal
Gynecologic
Hematologic and lymphatic
Soft tissue
Thoracic
Urologic
Other
30
20
10
0
Precision
Group
Best variant caller according to precision compared to MC3 reference
Consensus2
Consensus3
MuSE
MuTect2
SomaticSniper
Union
VarScan2
Central Nervous System
Endocrine
Gastrointestinal
Gynecologic
Hematologic and lymphatic
Soft tissue
Thoracic
Urologic
Other
30
20
10
0
Precision
Group

B

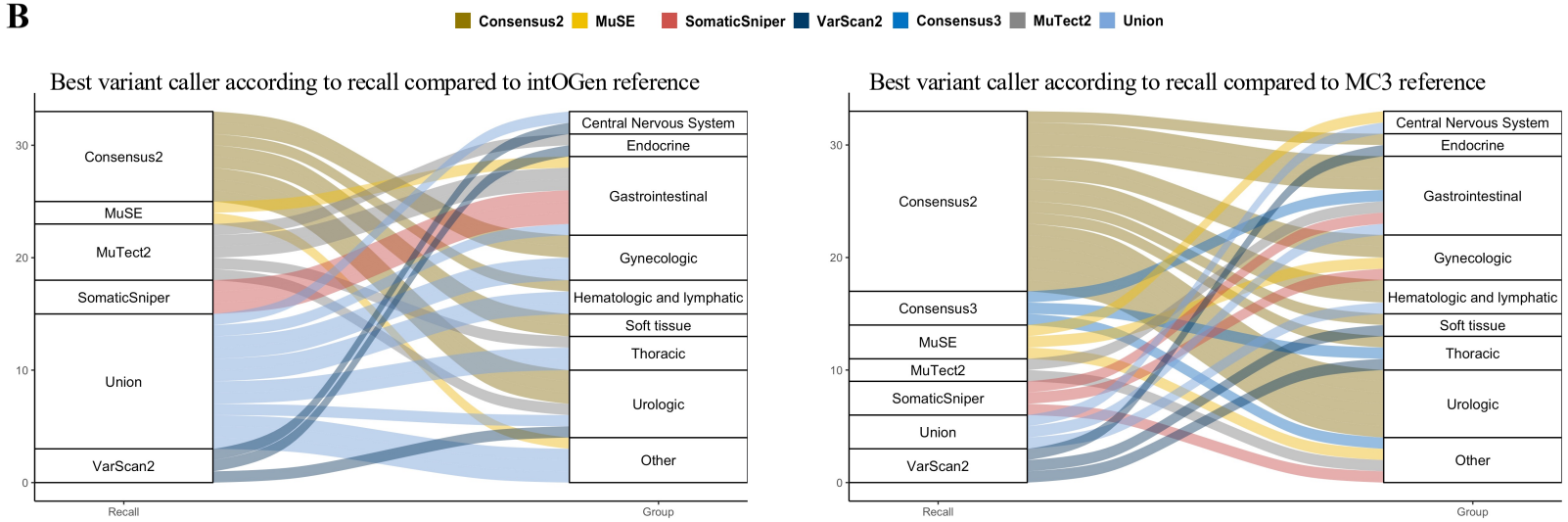

Consensus2
MuSE
SomaticSniper
VarScan2
Consensus3
MuTect2
Union
Best variant caller according to recall compared to intOGen reference
Consensus2
MuSE
MuTect2
SomaticSniper
Union
VarScan2
Central Nervous System
Endocrine
Gastrointestinal
Gynecologic
Hematologic and lymphatic
Soft tissue
Thoracic
Urologic
Other
30
20
10
0
Recall
Group
Best variant caller according to recall compared to MC3 reference
Consensus2
Consensus3
MuSE
MuTect2
SomaticSniper
Union
VarScan2
Central Nervous System
Endocrine
Gastrointestinal
Gynecologic
Hematologic and lymphatic
Soft tissue
Thoracic
Urologic
Other
30
20
10
0
Recall
Group

Figure S4

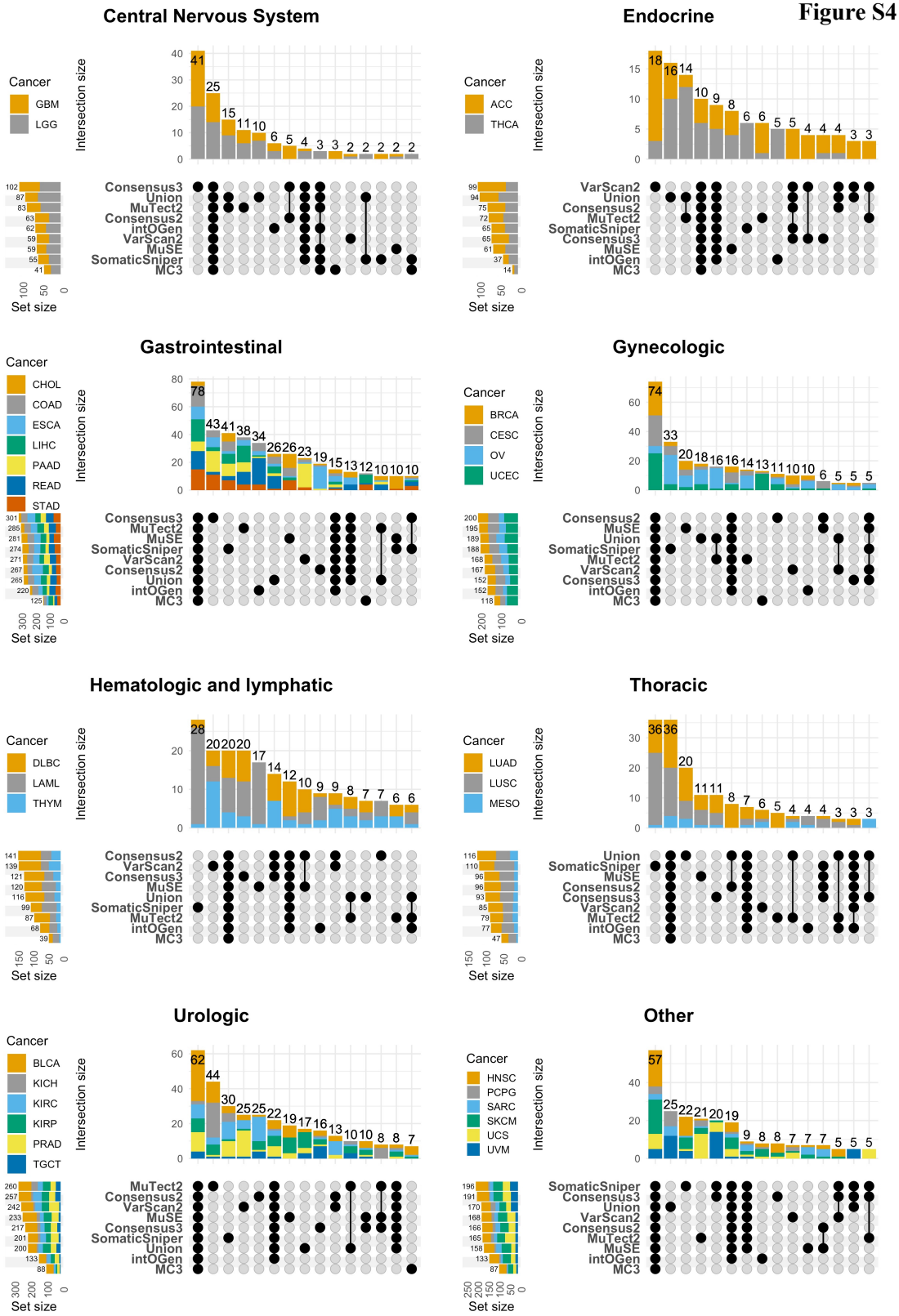

Central Nervous System
Cancer
GBM
LGG
Intersection size
41 25 15 11 10 6 5 4 3 3 2 2 2 2 2
Consensus3
Union
MuTect2
Consensus2
intOGen
VarScan2
MuSE
SomaticSniper
MC3
Set size
Endocrine
Cancer
ACC
THCA
Intersection size
18 16 14 10 9 8 6 6 5 5 4 4 4 3 3
VarScan2
Union
Consensus2
MuTect2
SomaticSniper
Consensus3
MuSE
intOGen
MC3
Set size
Gastrointestinal
Cancer
CHOL
COAD
ESCA
LIHC
PAAD
READ
STAD
Intersection size
78 43 41 38 34 26 26 23 19 15 13 12 10 10 10
Consensus3
MuTect2
MuSE
SomaticSniper
VarScan2
Consensus2
Union
intOGen
MC3
Set size
Gynecologic
Cancer
BRCA
CESC
OV
UCEC
Intersection size
74 33 20 18 16 16 14 13 11 10 10 6 5 5 5
Consensus2
MuSE
Union
SomaticSniper
MuTect2
VarScan2
Consensus3
intOGen
MC3
Set size
Hematologic and lymphatic
Cancer
DLBC
LAML
THYM
Intersection size
28 20 20 20 17 14 12 10 9 9 8 7 7 6 6
Consensus2
VarScan2
Consensus3
MuSE
Union
SomaticSniper
MuTect2
intOGen
MC3
Set size
Thoracic
Cancer
LUAD
LUSC
MESO
Intersection size
36 36 20 11 11 8 7 6 5 4 4 4 3 3 3
Union
SomaticSniper
MuSE
Consensus2
Consensus3
VarScan2
MuTect2
intOGen
MC3
Set size
Urologic
Cancer
BLCA
KICH
KIRC
KIRP
PRAD
TGCT
Intersection size
62 44 30 25 25 22 19 17 16 13 10 10 8 8 7
MuTect2
Consensus2
VarScan2
MuSE
Consensus3
SomaticSniper
Union
intOGen
MC3
Set size
Other
Cancer
HNSC
PCPG
SARC
SKCM
UCS
UVM
Intersection size
57 25 22 21 20 19 9 8 8 7 7 7 5 5 5
SomaticSniper
Consensus3
Union
VarScan2
Consensus2
MuTect2
MuSE
intOGen
MC3
Set size

Figure S5

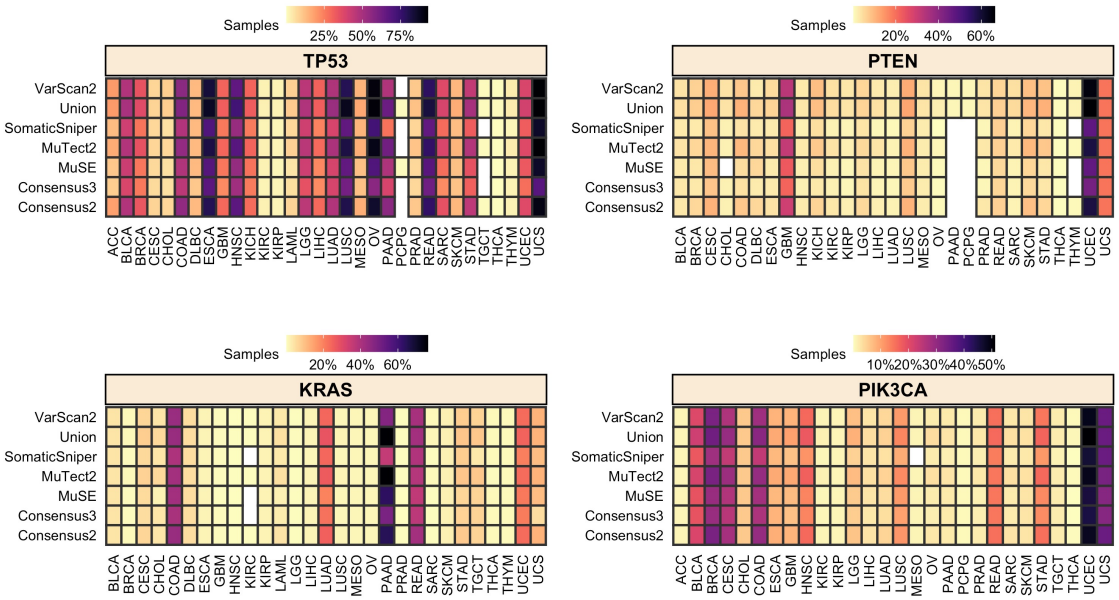


Figure S6

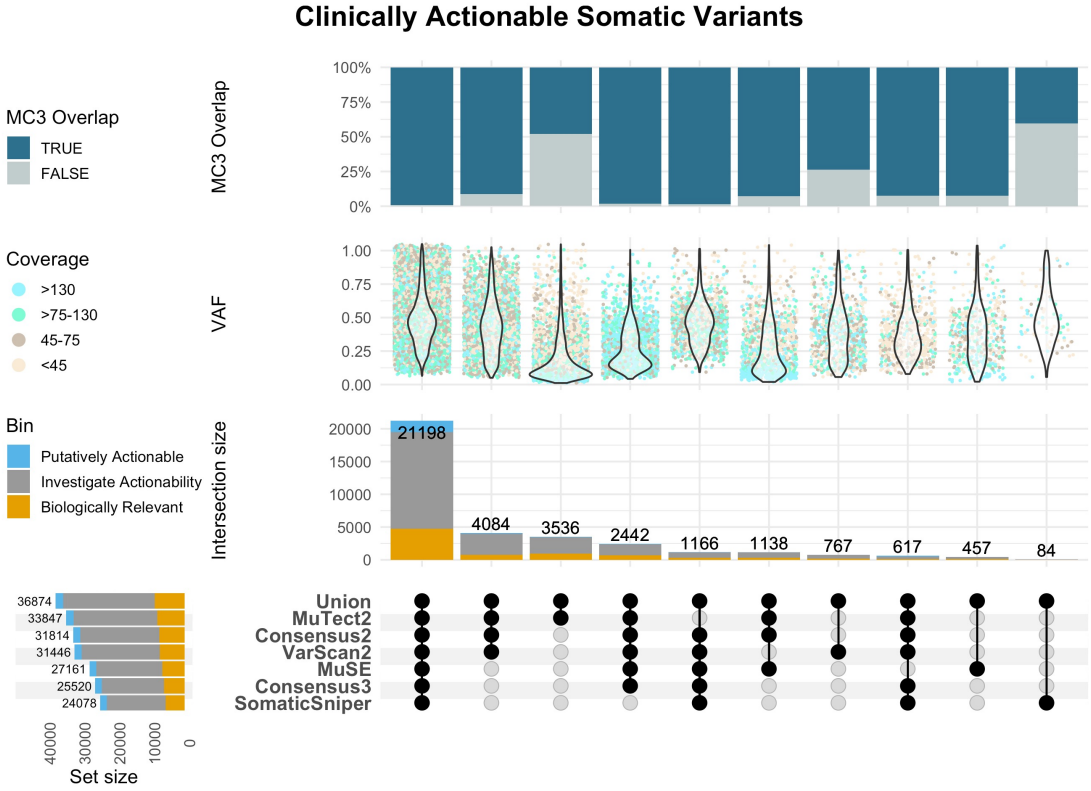
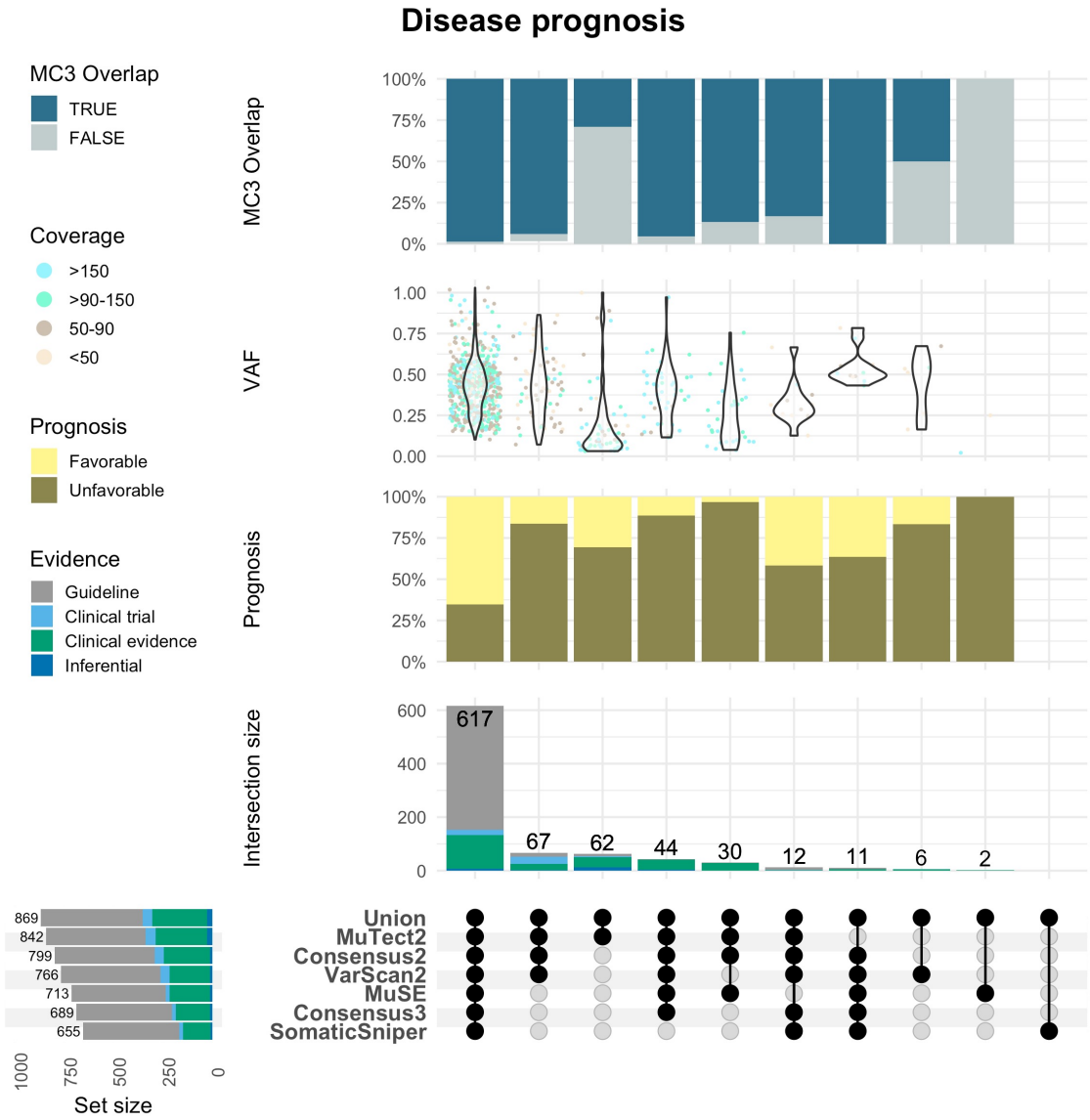
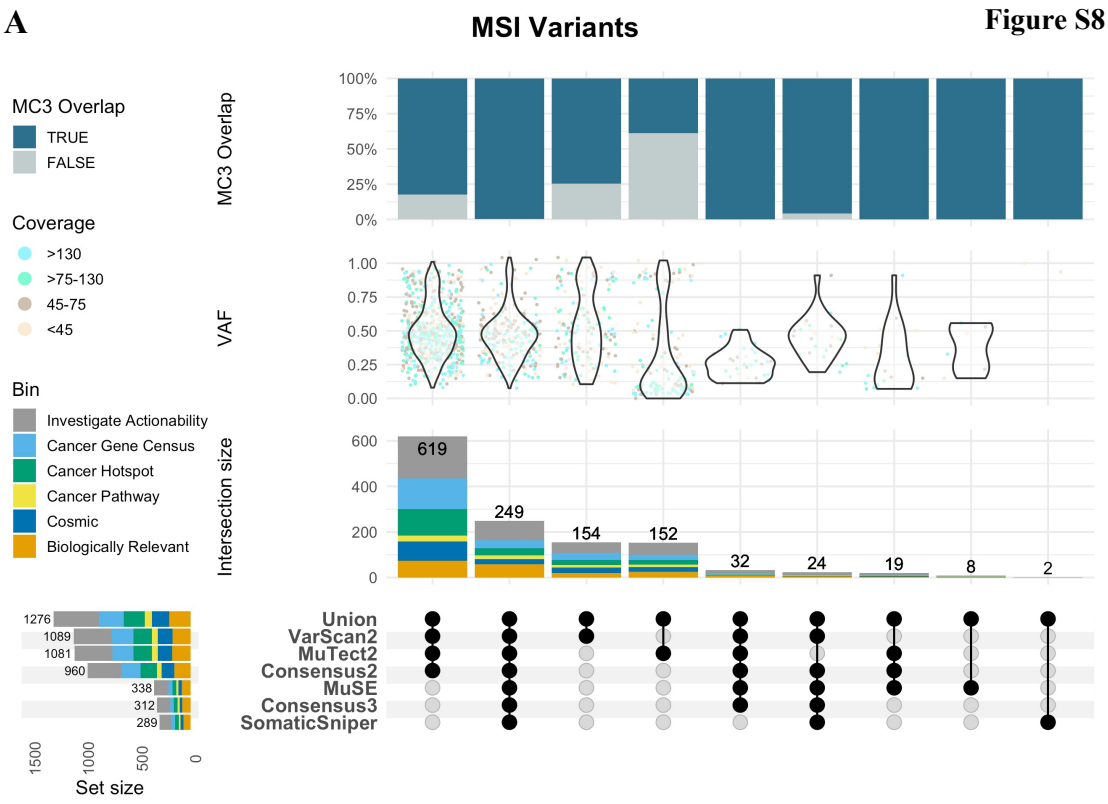
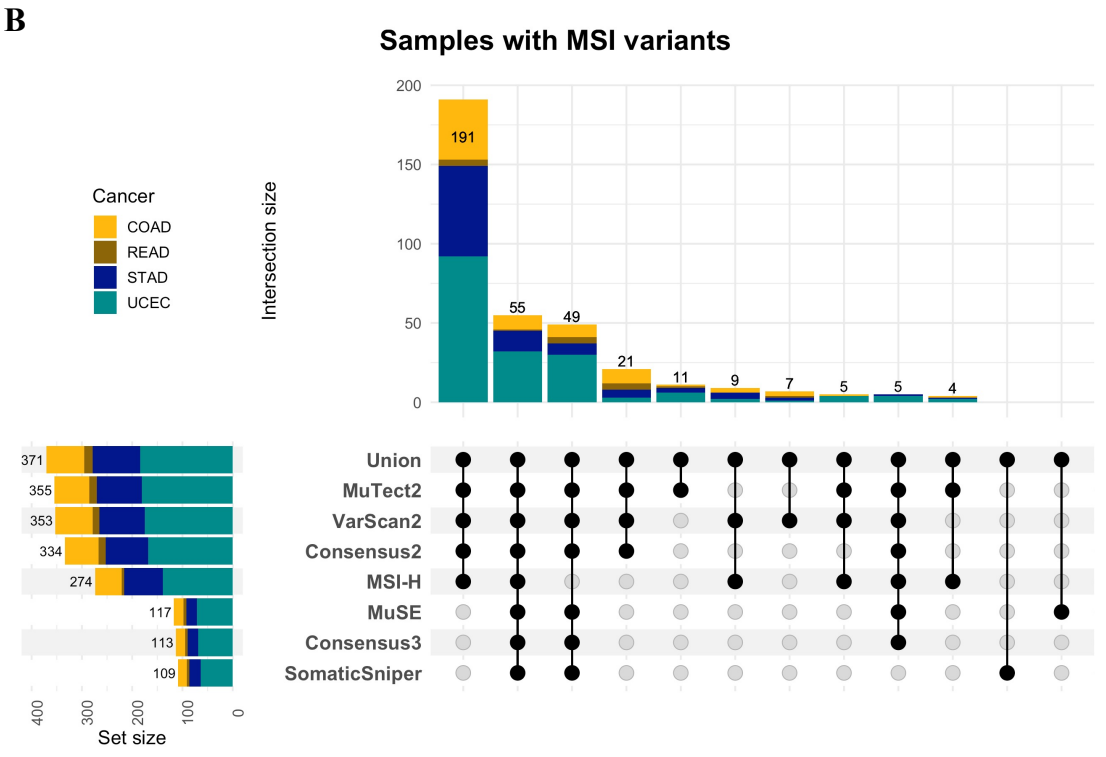


Figure S7


Disease prognosis
MC3 Overlap
TRUE
FALSE
Coverage
>150
>90-150
50-90
<50
Prognosis
Favorable
Unfavorable
Evidence
Guideline
Clinical trial
Clinical evidence
Inferential
MC3 Overlap
VAF
Prognosis
Intersection size
100%
75%
50%
25%
0%
1.00
0.75
0.50
0.25
0.00
600
400
200
0
617
67
62
44
30
12
11
6
2
Union
MuTect2
Consensus2
VarScan2
MuSE
Consensus3
SomaticSniper
869
842
799
766
713
689
655
1000
750
500
250
0
Set size


A
MSI Variants
MC3 Overlap
TRUE
FALSE
Coverage
>130
>75-130
45-75
<45
Bin
Investigate Actionability
Cancer Gene Census
Cancer Hotspot
Cancer Pathway
Cosmic
Biologically Relevant
MC3 Overlap
100%
75%
50%
25%
0%
VAF
1.00
0.75
0.50
0.25
0.00
Intersection size
600
400
200
0
619
249
154
152
32
24
19
8
2
1276
1089
1081
960
338
312
289
1500
1000
500
0
Set size
Union
VarScan2
MuTect2
Consensus2
MuSE
Consensus3
SomaticSniper


B
Samples with MSI variants
Cancer
COAD
READ
STAD
UCEC
Intersection size
200
150
100
50
0
191
55
49
21
11
9
7
5
5
4
371
355
353
334
274
117
113
109
400
300
200
100
0
Set size
Union
MuTect2
VarScan2
Consensus2
MSI-H
MuSE
Consensus3
SomaticSniper

Chapter II | Epigenetic profiling and response to CD19 chimeric antigen receptor T-cell therapy in B-cell malignancies

OXFORD

JNCI J Natl Cancer Inst (2022) 114(3): djab194

doi: 10.1093/jnci/djab194
First published online September 28, 2021
Article

Epigenetic Profiling and Response to CD19 Chimeric Antigen Receptor T-Cell Therapy in B-Cell Malignancies

Carlos A. Garcia-Prieto, MD,[1,2,‡] Lorea Villanueva, PhD,[1,‡] Alberto Bueno-Costa, MSc,[1] Veronica Davalos, PhD,[1] Europa Azucena González-Navarro, PhD,[3] Manel Juan, MD, PhD,[3,4] Álvaro Urbano-Ispizua, MD,[1,4,5] Julio Delgado, MD,[4,6] Valentín Ortiz-Maldonado, MD,[4] Francesca del Bufalo, MD,[7] Franco Locatelli, MD,[7,8] Concetta Quintarelli, MD,[7,9] Matilde Sinibaldi, PhD,[7] Marta Soler, MSc,[1] Manuel Castro de Moura, MSc,[1] Gerardo Ferrer, PhD,[1] Rocio G. Urdinguio, PhD,[10] Agustin F. Fernandez, PhD,[10] Mario F. Fraga, PhD,[10] Diana Bar, BSc,[11] Amilia Meir, MSc,[11] Orit Itzhaki, PhD,[12] Michal J. Besser, PhD,[12,13] Abraham Avigdor, MD,[13,14] Elad Jacoby, MD,[11,13] Manel Esteller, MD, PhD,[1,6,15,16,*]

[1]Cancer and Leukemia Epigenetics and Biology Program (PEBCL), Josep Carreras Leukaemia Research Institute (IJC), Badalona, Spain; [2]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain; [3]Department of Immunology, Hospital Clinic, Barcelona, Spain; [4]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; [5]Department of Hematology, University of Barcelona (UB), Barcelona, Spain; [6]Centro de Investigación Biomédica en Red de Cancer (CIBERONC), Madrid, Spain; [7]Department of Paediatric Haematology and Oncology, Cell and Gene Therapy, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy; [8]Department of Pediatrics, Sapienza University of Rome, Rome, Italy; [9]Department of Clinical Medicine and Surgery, University of Naples Federico II, Naples, Italy; [10]Nanomaterials and Nanotechnology Research Center (CINNCSIC), Health Research Institute of Asturias (ISPA), Institute of Oncology of Asturias (IUOPA), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Department of Organisms and Systems Biology (BOS), University of Oviedo, Oviedo, Spain; [11]Division of Pediatric Hematology and Oncology, The Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel; [12]Ella Lemelbaum Institute for Immuno Oncology, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel; [13]Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel; [14]Institute of Hematology, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel; [15]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain; and [16]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Spain

‡These authors contributed equally.

*Correspondence to: Manel Esteller, MD, PhD, Josep Carreras Leukaemia Research Institute (IJC), Carretera de Can Ruti, Camí de les Escoles s/n, 08916 Badalona, Barcelona, Catalonia, Spain (e-mail: mesteller@carrerasresearch.org).

Abstract

Background: Chimeric antigen receptor (CAR) T cells directed against CD19 (CART19) are effective in B-cell malignancies, but little is known about the molecular factors predicting clinical outcome of CART19 therapy. The increasingly recognized relevance of epigenetic changes in cancer immunology prompted us to determine the impact of the DNA methylation profiles of CART19 cells on the clinical course. **Methods:** We recruited 114 patients with B-cell malignancies, comprising 77 patients with acute lymphoblastic leukemia and 37 patients with non-Hodgkin lymphoma who were treated with CART19 cells. Using a comprehensive DNA methylation microarray, we determined the epigenomic changes that occur in the patient T cells upon transduction of the CAR vector. The effects of the identified DNA methylation sites on clinical response, cytokine release syndrome, immune effector cell-associated neurotoxicity syndrome, event-free survival, and overall survival were assessed. All statistical tests were 2-sided. **Results:** We identified 984 genomic sites with differential DNA methylation between CAR-untransduced and CAR-transduced T cells before infusion into the patient. Eighteen of these distinct epigenetic loci were associated with complete response (CR), adjusting by multiple testing. Using the sites linked to CR, an epigenetic signature, referred to hereafter as the EPICART signature, was established in the initial discovery cohort (n = 79), which was associated with CR (Fisher exact test, $P < .001$) and enhanced event-free survival (hazard ratio [HR] = 0.36; 95% confidence interval [CI] = 0.19 to 0.70; $P = .002$; log-rank $P = .003$) and overall survival (HR = 0.45; 95% CI = 0.20 to 0.99; $P = .047$; log-rank $P = .04$;). Most important, the EPICART profile maintained its clinical course predictive value in the validation cohort (n = 35), where it was associated with CR (Fisher exact test, $P < .001$) and enhanced overall survival (HR = 0.31; 95% CI = 0.11 to 0.84; $P = .02$; log-rank $P = .02$). **Conclusions:** We show that the DNA methylation landscape of patient CART19 cells influences the efficacy of the cellular immunotherapy treatment in patients with B-cell malignancy.

Received: February 18, 2021; **Revised:** July 11, 2021; **Accepted:** September 22, 2021

Chimeric antigen receptor (CAR) T-cell therapy has proved to be effective in patients for whom few therapeutic options otherwise remain, such as those with relapsed/refractory B-cell acute lymphoblastic leukemia (ALL) and B-cell lymphomas (1-4). These results have led to clinical approval of commercially available treatments (1). Despite the great hopes that CAR T cells directed against CD19 (CART19) cells has raised, treatment failure is not uncommon. The discovery of predictive biomarkers of clinical outcome to CART19 therapy would be highly relevant for risk stratification and the selection of alternative therapies. The lack of initial clinical response or the occurrence of relapse could have several causes related to the CART construct, preparation of infused cells, delivery of transduced cells, and biological features of the targeted B-cells, but only a few defects associated with CART19 inefficacy have been identified, the most studied being tumor antigen escape by loss of the CD19 protein (5). Other candidate molecular biomarkers for predicting CART19 clinical response in preinfused cells include CAR genomic integration sites (6-8) and cytokine expression profiles (9).

Herein, we have addressed whether the epigenetic status of the autologous CAR-transduced T-cells could also affect the clinical course of CART19 therapy. DNA methylation is altered in cancer (10,11), affecting the immune system and immunotherapy efficacy (12). In this regard, DNA methylation signatures are associated with clinical response to programmed cell death protein 1 checkpoint blockade (13) and the DNA methylation status of the vector for transgenic T-cell receptor adoptive cell therapy relates to changes in tumor burden (14). For these reasons, we decided to assess the effects of the DNA methylation landscape of preinfused CART19 cells on the clinical outcome of patients with B-cell malignancies.

Methods

Study Design

Patients were eligible to enter the study if they had an relapsed/refractory B-cell malignancy for which CART19 therapy was recommended. Patient CD19-engineered T cells from 114 cases, comprising 77 patients with ALL and 37 patients with non-Hodgkin lymphoma (NHL), were obtained from 3 academic clinical trials: NCT03144583 (15), NCT02772198 (16,17), and NCT03373071 (18). Written informed consent was obtained, and the Sheba Medical Center institutional review board and the Israeli Ministry of Health, the Research Ethics Committee (Celm) of the Hospital Clinic, and the institutional review board of Bambino Gesù Children Hospital, respectively, provided study approval. The clinical characteristics of the studied 114 patients are summarized in Table 1. The type of CART19 therapy used in each trial is described in Supplementary Methods (available online). High-molecular-weight DNA was extracted from all samples before CART19 infusion into patients.

DNA Methylation Procedure and Analysis

The DNA methylation status of the CART19 cells from each patient was established using the Infinium MethylationEPIC Kit (Illumina, Inc, San Diego, CA) (19). DNA methylation data are available in the Gene Expression Omnibus repository (GSE179414, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179414). An epigenetic signature, referred to hereafter as the EPICART DNA methylation signature (EPICART signature) was obtained using a trained, supervised classification model based on ridge-regularized logistic regression to predict clinical response. The classification model was optimized by tuning parameters (best performance with $\alpha = 0$ from ridge regression and regularization parameter $\lambda = 0.03$), with 10-fold cross-validation, repeated 3 times. Our model performance was assessed using the receiver operating characteristic curve of the resamples (area under the curve mean = 0.83; 95% confidence interval [CI] = 0.75 to 0.91). Flow cytometry analysis was used for validation. DNA methylation status of specific CpG sites was validated by pyrosequencing and bisulfite genomic sequencing of multiple clones. Quantitative reverse transcription–polymerase chain reaction (qRT-PCR) and Western blot were used to assess gene expression (Supplementary Methods; Supplementary Table 1, available online).

Clinical Statistical Analysis

Assay results were compared with patient outcomes in a double-blind manner. The statistical significance of the differences between distributions of the groups was estimated with Fisher exact test. The Mann-Whitney-Wilcoxon test was used to test the statistical significance of the differences between distributions of methylation or expression values among groups. The correlation between methylation and gene expression was estimated using Spearman test. Overrepresentation of T-cell population phenotypes in EPICART-positive and EPICART-negative CART19 cells was estimated using the Student *t* test. Event-free survival (EFS) was defined as the time from the start of CART19 treatment until the first occurrence of progression, relapse, or death. Overall survival (OS) was defined as the time from the start of CART19 treatment until death. The Kaplan-Meier method was also used to estimate the EFS and OS, the differences between the groups being calculated with the log-rank test. Hazard ratios from univariate Cox regressions were used to determine the association between clinicopathological features and survival. A *P* value less than 0.05 was considered statistically significant. All statistical tests were 2-sided unless otherwise stated.

Results

The Epigenomic Landscape of CART19 Cells

To discover an epigenomic profile associated with patients with a B-cell malignancy who would gain clinical benefit from CART19 treatment, we first studied the DNA methylation landscape of untransduced and transduced preinfusion T cells for the CD19 CAR retrovirus in 43 patients from the NCT02772198 clinical trial (Figure 1, A). This set of cases included 30 NHL (28 adult and 2 pediatric patients) and 13 ALL (8 pediatric and 5 adult patients). In this initial set, we interrogated the methylation status of approximately 850 000 CpG sites (19). In the 43 patients with a B-cell malignancy, DNA methylation levels differed between CART19 untransduced and transduced cells at 984 CpG sites (Supplementary Table 2, available online). Among these differential CpG sites, 52.7% (519 of 984) were

Table 1. Clinicopathological characteristics of the patients with a B-cell malignancy treated with chimeric antigen receptor T cells directed against CD19 cells[a]

Characteristic	Entire cohort (n = 114)	Discovery cohort (n = 79)	Validation cohort (n = 35)
Sex, No. (%)			
Male	68 (59.6)	41 (51.9)	27 (77.1)
Female	46 (40.4)	38 (48.1)	8 (22.9)
Median age (range), y	24 (3-70)	22 (3-70)	27 (4-70)
Age, No. (%), y			
<18	42 (36.8)	32 (40.5)	10 (28.6)
18-29	27 (23.7)	16 (20.3)	11 (31.4)
30-59	34 (29.8)	26 (32.9)	8 (22.9)
≥60	11 (9.6)	5 (6.3)	6 (17.1)
Diagnosis, No. (%)			
B-ALL	77 (67.5)	53 (67.1)	24 (68.6)
B-NHL	37 (32.5)	26 (32.9)	11 (31.4)
DLBCL	20 (17.5)	13 (16.5)	7 (20.0)
PMBCL	11 (9.6)	9 (11.4)	2 (5.7)
Follicular lymphoma	4 (3.5)	3 (3.8)	1 (2.9)
Burkitt lymphoma	1 (0.9)	0 (0.0)	1 (2.9)
Mantle cell lymphoma	1 (0.9)	1 (1.3)	0 (0.0)
Response, No. (%)			
CR	74 (64.9)	50 (63.3)	24 (68.6)
PR	16 (14)	11 (13.9)	5 (14.3)
Stable disease	9 (7.9)	6 (7.6)	3 (8.6)
Disease progression	15 (13.2)	12 (15.2)	3 (8.6)
CRS, No. (%)			
Grade 0	41 (36.0)	28 (35.4)	13 (37.1)
Grade 1	46 (40.4)	33 (41.8)	13 (37.1)
Grade 2	13 (11.4)	10 (12.7)	3 (8.6)
Grade 3	8 (7.0)	4 (5.1)	4 (11.4)
Grade 4	4 (3.5)	2 (2.5)	2 (5.7)
Grade 5	2 (1.8)	2 (2.5)	0 (0.0)
ICANS, No. (%)			
Grade 0	87 (76.3)	59 (74.7)	28 (80.0)
Grade 1	11 (9.6)	8 (10.1)	3 (8.6)
Grade 2	5 (4.4)	4 (5.1)	1 (2.9)
Grade 3	6 (5.3)	4 (5.1)	2 (5.7)
Grade 4	5 (4.4)	4 (5.1)	1 (2.9)
Grade 5	0 (0.0)	0 (0.0)	0 (0.0)
Origin of the CAR T cells			
NCT02772198	43 (37.7)	30 (38.0)	13 (37.1)
NCT03144583	45 (39.5)	31 (39.2)	14 (40.0)
NCT03373071	26 (22.8)	18 (22.8)	8 (22.9)

[a]B-ALL = B-cell acute lymphoblastic leukemia; B-NHL = B-cell non-Hodgkin lymphoma; CAR = chimeric antigen receptor; CR = complete response; CRS = cytokine release syndrome; DLBCL = diffuse large B-cell lymphoma; ICANS = immune effector cell-associated neurotoxicity syndrome; PMBCL = primary mediastinal B-cell lymphoma; PR = partial response.

hypermethylation events at the CART19 transduced cells vs the untransduced cells, whereas 47.3% (465 of 984) were hypomethylation changes. The CpG methylation content of these 984 sites was not distinct between CD4 and CD8 T cells (CD4 methylation β value 95% CI = 0.57 to 0.61; CD8 methylation β value 95% CI = 0.57 to 0.61, Mann-Whitney-Wilcoxon test $P = .73$). The genomic distribution of these CpG sites is illustrated in Figure 1, B. They were associated with known genes in 75.1% (739 of 984) of cases and, of these, were located within a defined regulatory region in 45.9% (339 of 739) of cases. Gene set enrichment analysis using gene ontology collections showed that the most overrepresented biological processes and Kyoto Encyclopedia of Genes and Genomes and Reactome pathways were the "T-cell receptor signaling pathway," "Pathways in cancer," and "Separation of sister chromatids," respectively (Figure 1, C). Using only CpG sites for regulatory regions, the most overrepresented categories were "T-cell receptor signaling pathway" and "Transcriptional regulation by RUNX3" (Supplementary Figure 1, A, available online), whereas using only gene body sites, the most overrepresented categories were "Homophilic cell adhesion via plasma membrane adhesion molecules" and "Separation of sister chromatids" (Supplementary Figure 1, B, available online).

T cells transduced with CD19 CAR retroviruses could themselves be vulnerable to DNA methylation silencing (20). Thus, we examined whether a distinct DNA methylation status of the retrovirus in the transduced T-cell could also influence clinical outcome. Pyrosequencing analyses of the retroviral vector showed an unmethylated status of the retroviral vector in the CART19 cells (Supplementary Figure 1, C, available online).

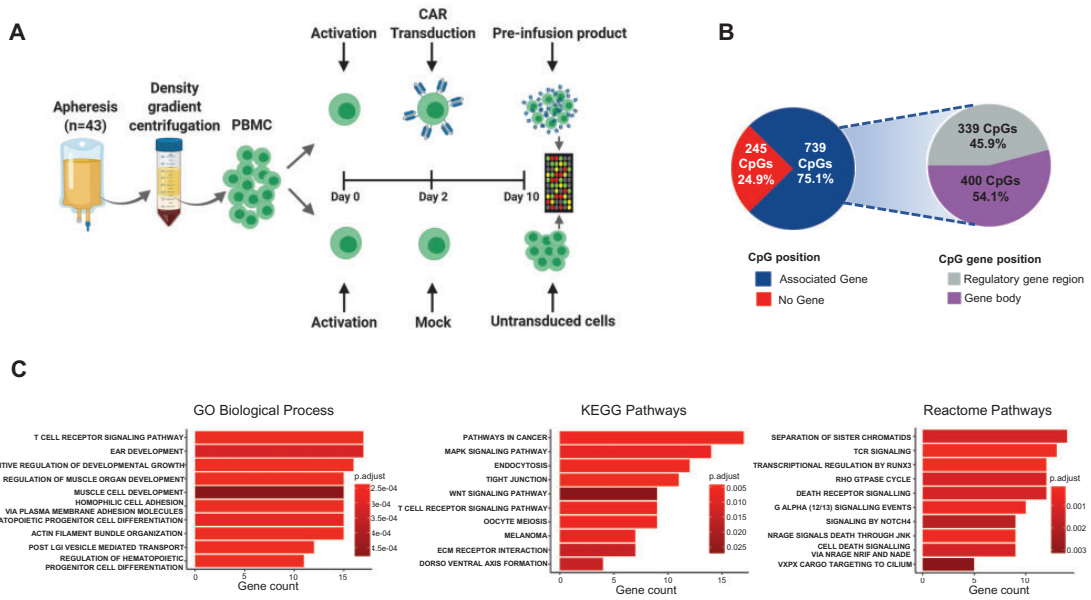


Figure 1. Characterization of epigenetic changes in patient T cells upon transduction of the chimeric antigen receptor (CAR) vector. **A)** Experimental design developed to detect DNA methylation changes in patient T cells upon CAR transduction. **B)** Distribution of the 984 CpG sites identified in the human genome. **C)** Gene ontology (GO) analysis of genes with CpGs that changed upon CAR transduction (overrepresentation analysis with false discovery rate adjusted $P < .05$). KEGG = Kyoto Encyclopedia of Genes and Genomes; PBMC = Peripherla Blood Mononuclear Cell.

Impact of CART19 Epigenetics in Clinical Outcome: The EPICART Signature

Fisher exact test with correction for multiple hypothesis testing using the false discovery rate (FDR) was applied to identify any association between the DNA methylation status of the 984 CpG sites identified in CART19-transduced cells and the clinical outcomes in 114 patients with a B-cell malignancy treated with this type of cell therapy (Table 1). For the contingency tables, clinical response was categorized as complete response (CR) vs non-CR (partial response + stable disease + disease progression). For the adverse effects, we followed the guidelines of the American Society for Transplantation and Cellular Therapy (21): Cytokine release syndrome (CRS) was divided into grade 0 vs grades 1 through 5, and immune effector cell-associated neurotoxicity syndrome (ICANS) was split into grade 0 vs grades 1 through 5. These cases were divided into a discovery cohort of 79 patients and a validation cohort of 35 patients (Table 1). The 2 cohorts did not show statistically significant differences related to age (pediatric vs adult; Fisher exact test, $P = .29$), origin of the sample (NCT03144583, NCT02772198, and NCT03373071; Fisher exact test, $P = 1$), type of B-cell malignancy (ALL vs NHL; Fisher exact test, $P = 1$), clinical response (CR vs partial response, stable disease, or disease progression; Fisher exact test, $P = .67$), or the appearance of CRS (0 vs 1-5; Fisher exact test, $P = 1$) or ICANS (0 vs 1-5; Fisher exact test, $P = .64$). DNA from the CART19-transduced cells infused in each patient was hybridized to the described DNA methylation microarray (19).

In our discovery cohort ($n = 79$), we found 54 CpG sites (5.5% of the 984 sites) at the initial screening by Fisher exact test for which the DNA methylation levels were statistically significantly associated with clinical variables. The DNA methylation status of 45, 8, and 5 CpG sites was associated, respectively, with CR (Supplementary Table 3, available online), CRS (Supplementary Table 4, available online), and ICANS (Supplementary Table 5, available online). We then applied to all the identified CpG sites with potential clinical value derived from the Fisher exact test the FDR statistical approach used in multiple-hypothesis testing to correct for multiple comparisons. We found that although the epigenetic loci linked to CRS and ICANS failed this test, 40.0% (18 of 45) of the CpG sites associated with CR passed the FDR for multiple testing (Supplementary Table 6, available online).

When we had established that a set of 18 epigenomic loci adjusted by multiple testing could discriminate a CR result following CART19 treatment (Supplementary Table 6, available online), we examined whether these sites could also predict EFS and OS in our discovery cohort ($n = 79$). In this regard, the presence of a CR was associated with enhanced EFS and improved OS (Figure 2, A). When we selected the 18 methylation sites associated with CR (Supplementary Table 6, available online) to train a supervised classification model based on ridge-regularized logistic regression, we obtained an EPICART signature. The use of the EPICART signature in the supervised hierarchical clustering for the discovery cohort of CART cases classified patients as those exhibiting CR or non-CR (Fisher exact test, $P < .001$) (Supplementary Figure 2, available online). Most important, the EPICART signature was associated with EFS (Figure 2, B) and OS (Figure 2, B).

Taking advantage of the dissected DNA methylation patterns of the different T-cell populations from the International Human Epigenome Consortium (22), we undertook a molecular

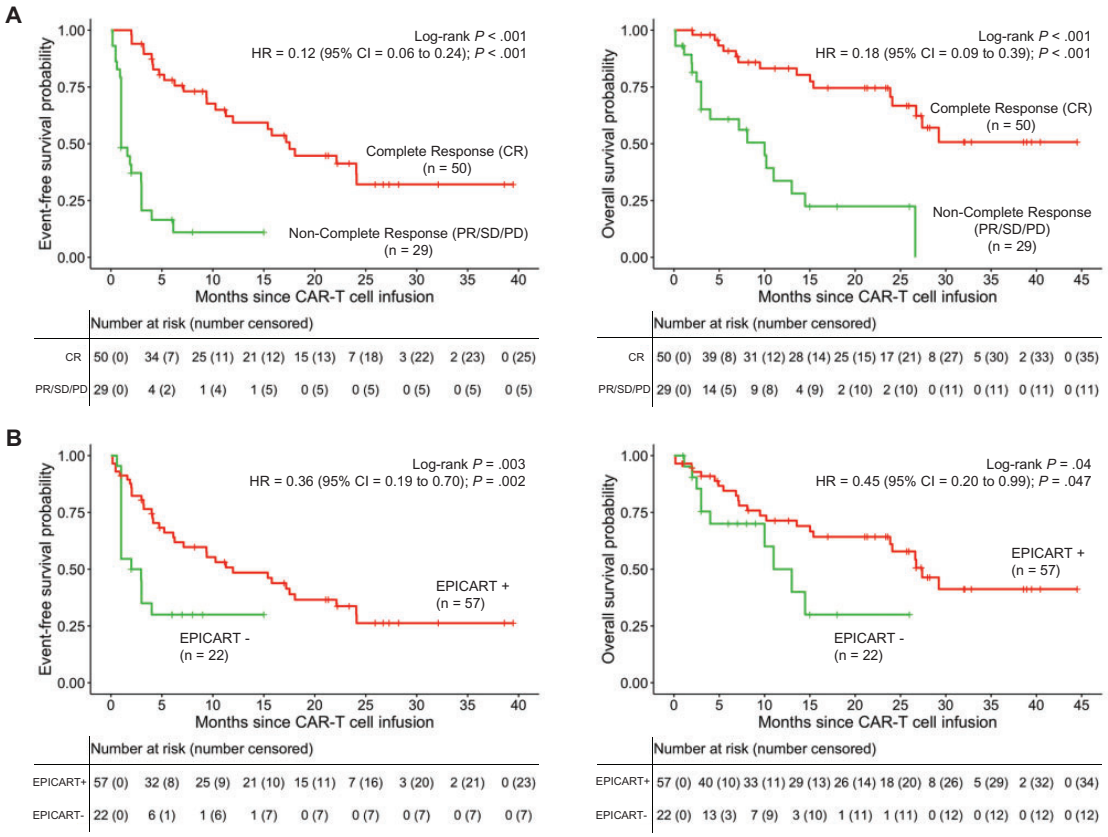


Figure 2. Complete response (CR) and DNA methylation signature (EPICART) associated with event-free survival (EFS) and overall survival (OS) in the discovery cohort of patients with a B-cell malignancy treated with chimeric antigen receptor T cells directed against CD19 (CART19) therapy. **A)** Kaplan-Meier analysis of EFS (left) and OS (right) in 79 patients with a B-cell malignancy according to the presence of CR or its absence (partial response [PR] + stable disease [SD] + progression of the disease [PD]). **B)** Kaplan-Meier analysis of EFS (left) and OS (right) in the same patients with a B-cell malignancy according to the presence of an EPICART signature in the preinfused CART19 cells, defined by the methylation status of the 18 CpG sites associated with CR (EPICART-positive [+] signature). For all cases, the P value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR), with a 95% confidence interval (CI). A P value less than .05 was considered statistically significant. The number of events is also shown. All statistical tests were 2-sided.

dissection of the T-cell classes in our EPICART signature. We found that the EPICART-positive signature identified CART19 cells enriched in CD4 and CD8 naive-like or early memory phenotype T cells (Fisher exact test, $P = .03$). Conversely, EPICART-negative CART19 cells were enriched in more committed and differentiated lineages, such as effector memory CD4 and CD8 T cells, and terminally differentiated effector memory CD8 T cells (Fisher exact test, $P < .001$). The described population phenotypes assigned by computational projection were validated by flow cytometry analyses in 43 cases (38 ALL and 5 NHL) of the discovery cohort, where these data were available. The use of the markers CD3, CD45RA, and CCR7 to define the population status of naive T cells (TNs: CD3+CD45RA+CCR7+), central memory T cells (TCMs: CD3+CD45RA-CCR7+), effector memory T cells (CD3+CD45RA−CCR7−), and effector T cells (TEMRAs: CD3+CD45RA+CCR7−) confirmed that EPICART-positive CART19 cells were enriched in TNs/central TCMs (EPICART-positive cells, 95% CI = 48.39% to 66.17%; EPICART-negative cells, 95% CI = 20.13% to 56.19%; Student *t* test, $P = .04$), whereas in EPICART-negative cells effector memory T-cell/TEMRA populations were overrepresented (EPICART-positive cells, 95% CI = 28.31% to 45.63%; EPICART-negative cells, 95% CI = 39.86% to 75.24%; Student *t* test, $P = .03$) (Supplementary Methods, available online). Examples of flow cytometry analyses are shown in Supplementary Figure 3, A (available online). Importantly, we observed that those patients with a B-cell malignancy receiving CARTs enriched with TN + TCM showed improved EFS and OS compared with those given adoptive cell therapy enriched with effector memory T cell + TEMRA (Supplementary Figure 3, B, available online). These results are consistent with the adoptive cell therapy concept that TNs or early TCMs can outperform TEMRAs because of the limited niche homing, survival, and self-renewal capacity of the effector cells relative to the less committed and more immature T cells (23-27).

Related to any obvious impact on gene expression for the 18 CpG sites that defined the EPICART signature, RNA or protein for the CART19 cells was not available; thus, we data-mined 100 blood cell lines analyzed for DNA methylation and expression (28). We observed that hypermethylation of those CpG sites

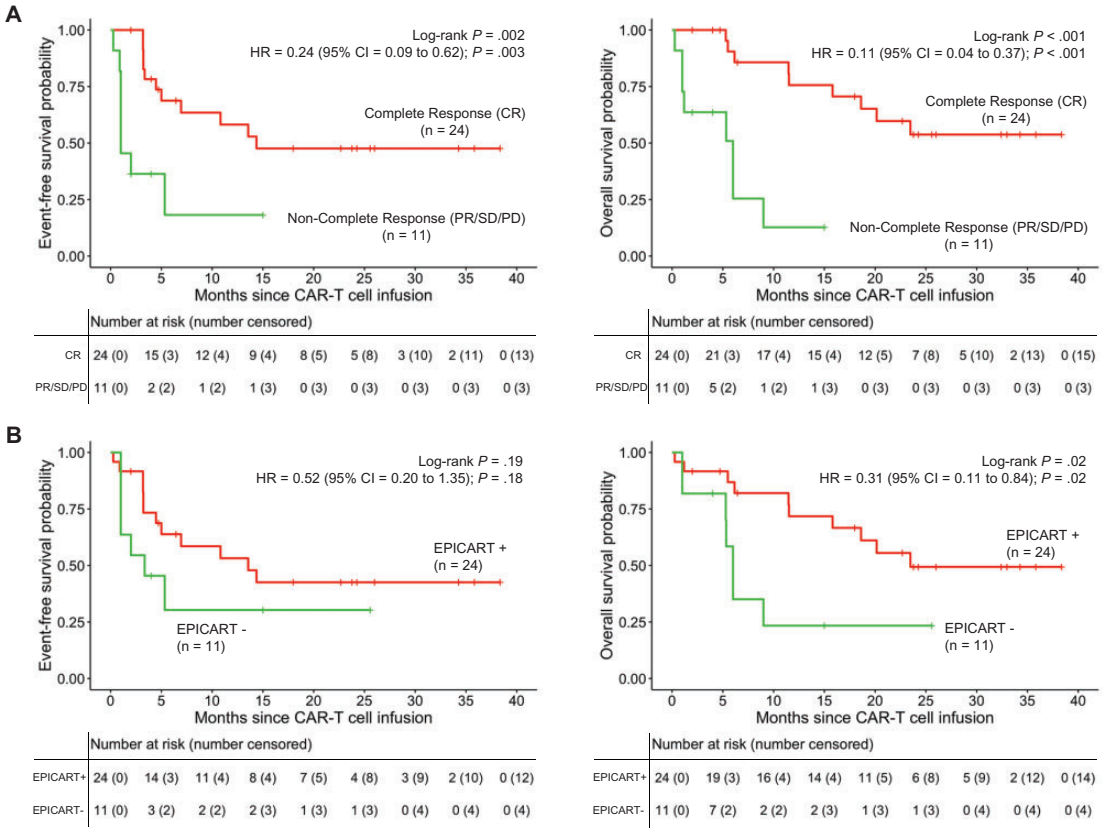


Figure 3. Complete response (CR) and DNA methylation (EPICART) signature associated with event-free survival (EFS) and overall survival (OS) in the validation cohort of patients with a B-cell malignancy treated with chimeric antigen receptor T cells directed against CD19 (CART19) therapy. **A)** Kaplan-Meier analysis of EFS (left) and OS (right) in 35 patients with a B-cell malignancy according to the presence of CR or its absence (partial response [PR] + stable disease [SD] + progression of the disease [PD]). **B)** Kaplan-Meier analysis of EFS (left) and OS (right) in the same patients with a B-cell malignancy according to the presence EPICART signature in the preinfused CART19 cells, defined by the methylation status of the 18 CpG sites associated with CR (EPICART-positive [+] signature). For all cases, the *P* value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR) with a 95% confidence interval (CI). A *P* value less than .05 was considered statistically significant. The number of events is also shown. All statistical tests were 2-sided.

located in the gene bodies was associated with transcript upregulation (methylated CpGs z score, 95% CI = 0.30 to 0.49; unmethylated CpGs z score, 95% CI = −0.28 to −0.02; Mann-Whitney-Wilcoxon test, $P < .001$) (Supplementary Figure 4, A, available online). Illustrative examples are shown for the hypermethylated CpG sites in the gene bodies of *INPP5A* and *ECHDC1* (Supplementary Table 6, available online) (Spearman test in 100 blood cell lines, $\rho > 0.3$; $P < .001$) (Supplementary Figure 4, B, available online). The presence of gene body hypermethylation accompanied by gene upregulation has been reported (29). Importantly, using T-cell–derived lines from these analyses, we validated that *INPP5A* and *ECHDC1* gene-body hypermethylation was associated with elevated expression, whereas gene-body hypomethylation was associated with gene downregulation (Supplementary Figure 4, C, available online). Concordantly, the use of the DNA methylation inhibitor 5-Aza-2'-deoxycytidine in the hypermethylated cell lines downregulated INPP5A and ECHDC1 expression (Supplementary Figure 4, D, available online). Furthermore, we experimentally validated by pyrosequencing and bisulfite genomic sequencing of multiple clones the DNA methylation status of these CpG sites in EPICART-positive and negative patients (Supplementary Figure 4, E, available online). Further data mining of the T-cell–derived lines showed that hypermethylation of 5′-end CpG sites was statistically significantly associated with transcript downregulation (Supplementary Figure 4, F, available online). An illustrative example is the 5′-UTR CpG hypermethylation of *FOXN3*, a candidate tumor suppressor for T-cell ALL (30) (Supplementary Figure 4, G, available online).

EPICART Validation and Single Loci Associated With Clinical Course

Having characterized the EPICART signature as being a predictor of CR, EFS, and OS in the discovery cohort of B-cell malignancies treated with CART19, we asked whether the identified DNA methylation landscape could also distinguish clinical outcome in the validation cohort (Table 1). From a clinical standpoint, CR was associated with enhanced EFS and improved OS in the

Table 2. Annotation of the 6 CpGs correlated with complete response and with statistically significant improvement in event-free survival and overall survival[a]

Probe ID[b]	Chromosomal position (hg19)[c]	Associated gene[d]	CR FDR P value[e]	EFS P value[f]	OS P value[f]
cg12012941	chr1:188676237	Not described	<.001	.01	.01
cg04267686	chr6:105907265	Not described	.001	.02	.001
cg25534076	chr1:234087867	*SLC35F3*	.002	.04	.03
cg12260379	chr2:86332162	*PTCD3; POLR1A*	.01	.03	.04
cg09992216	chr11:32353565	Not described	.01	.009	.004
cg12610471	chr10:22634199	*SPAG6*	.02	.001	.003

[a]Annotation retrieved from the Infinium MethylationEPIC Array Kit (Illumina, Inc, San Diego, CA) manifest. CR = complete response; EFS = event-free survival; FDR = false discovery rate; OS = overall survival.
[b]Unique identifier from the Illumina CpG database.
[c]Chromosomal coordinates of the CpG (build hg19).
[d]Target gene name from the University of California Santa Cruz database.
[e]The FDR-adjusted P value of the CR is derived from the Fisher exact test (CR vs no response/stable disease/disease progression). All tests were 2-sided.
[f]The P value of EFS and OS is derived from the log-rank test in Kaplan-Meier curves. All tests were 2-sided.

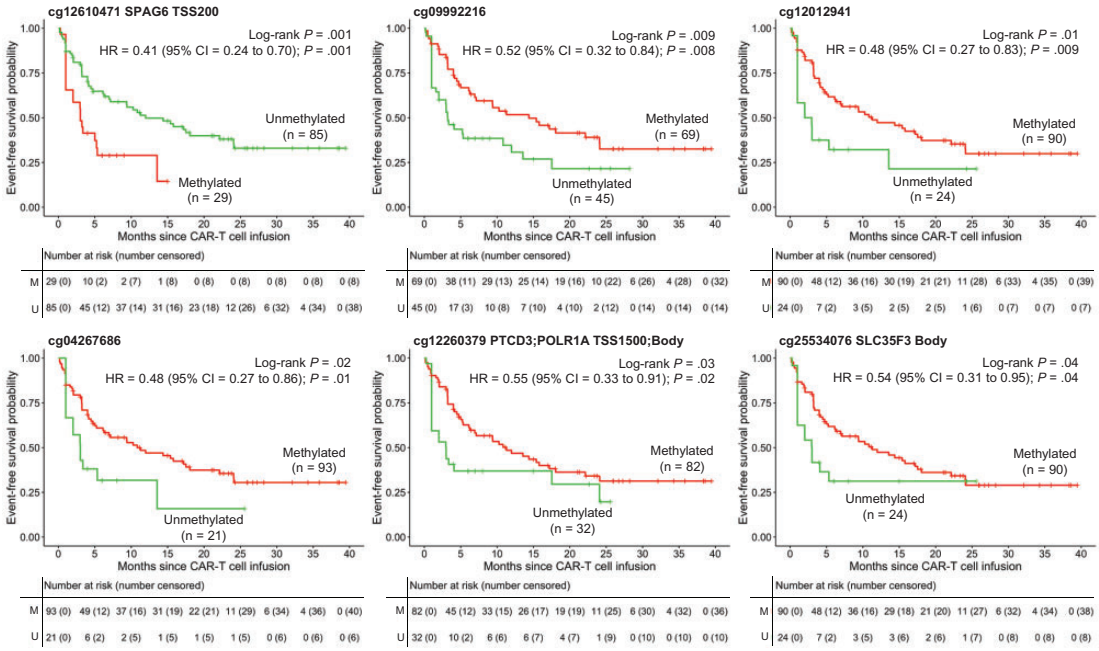


Figure 4. Kaplan-Meier estimates of event-free survival with respect to the chimeric antigen receptor T cells directed against CD19 (CART19) cell preinfusion methylation status of 6 candidate single CpG loci in patients with a B-cell malignancy treated with the adoptive cell therapy. The P value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR) with a 95% confidence interval (CI). A P value of less than .05 was considered statistically significant. The number of events is also shown. All statistical tests were 2-sided. M = methylated; U = unmethylated.

validation set (Figure 3, A). Importantly, EPICART signature predicted CR to CART cell therapy with 82.9% accuracy (95% CI = 66.4% to 93.4%; κ = 0.60), 87.5% sensitivity, and 72.7% specificity in the validation cohort. We further evaluated the model performance using the receiver operating characteristic curve, obtaining an area under the curve value of 0.80. Use of the EPICART signature in the supervised hierarchical clustering for the validation cohort of CART cases also distinguished CR or non-CR (Fisher exact test, P < .001) (Supplementary Figure 5, A, available online). Remarkably, the EPICART-positive signature was associated with improved OS in the validation cohort (hazard ratio = 0.31; 95% CI = 0.11 to 0.84; P = .02; log-rank P = .02) (Figure 3, B). We also found a nonstatistically significant trend between the EPICART-positive signature and EFS (hazard ratio = 0.52; 95% CI = 0.20 to 1.35; P = .18; log-rank P = .19) (Figure 3, B).

Finally, for the entire cohort, CR was associated with EFS and OS (Supplementary Figure 5, B, available online). The EPICART signature in the supervised hierarchical clustering for the complete set of available cases (discovery + validation, n = 114) also

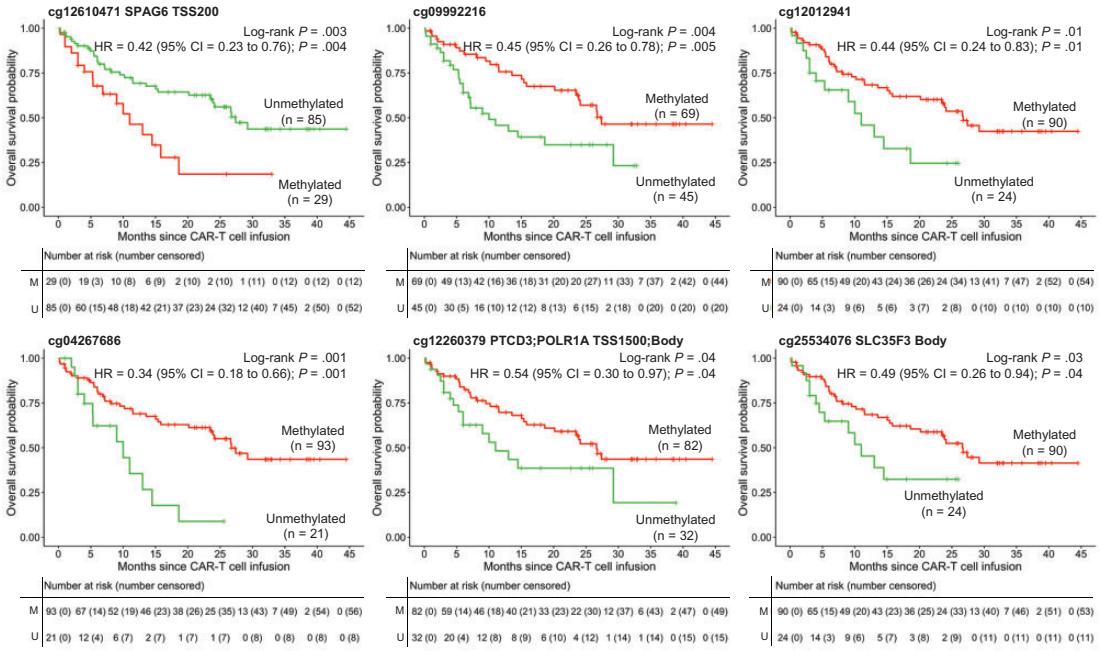


Figure 5. Kaplan-Meier estimates of overall survival relative to the chimeric antigen receptor T cells directed against CD19 (CART19) cell preinfusion methylation status of 6 candidate single CpG loci in patients with a B-cell malignancy treated with adoptive cell therapy. The *P* value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR) with a 95% confidence interval (CI). A P value of less than .05 was considered statistically significant. The number of events is also shown. All statistical tests were 2-sided. M = methylated; U = unmethylated.

classified patients as those exhibiting CR or non-CR (Fisher exact test, $P < .001$) (Supplementary Figure 5, C, available online). Importantly, in the entire cohort, EPICART-positive signature was associated with improved EFS and OS (Supplementary Figure 5, D, available online). The hazard ratios and *P* values for EFS and OS obtained from each cohort are summarized in Supplementary Table 7 (available online).

To identify a smaller set of biomarkers that could simplify the analysis, we found 6 epigenomic loci from the EPICART signature that, analyzed alone, were also associated with improved EFS and OS. These CpG sites are summarized in Table 2, and the corresponding Kaplan-Meier curves for EFS and OS are shown in Figures 4 and 5, respectively. The 4 genes associated with these 6 DNA methylation loci were *PTCD3* and *POLR1A*, involved in protein production regulation at ribosomes (31,32); *SLC35F3*, a thiamine transferase involved in T-cell infiltration (33); and *SPAG6*, which regulates cell apoptosis through tumor necrosis factor–related apoptosis-inducing ligand (TRAIL) signaling (34). *SPAG6* was further studied, given the proposed used of a TRAIL variant to overcome CART resistance (35) and the CpG location at the transcription start site. We observed in T-cell–derived lines the association between *SPAG6* hypermethylation and downregulation measured by qRT-PCR and Western blot (Supplementary Figure 6, A and B, available online). Hypermethylation-associated silencing was also found for *PTCD3*, the other candidate gene with an identified differentially methylated CpG site in its promoter region (Supplementary Figure 6, C and D, available online).

Discussion

The use of CART19 therapy has improved the clinical outcome of patients with relapsed/refractory B-cell malignancies (1-4). Despite the promising initial results, however, a not-negligible proportion of cases does not show CR or does not achieve long-term remission (1-4). This finding is relevant from the point of view of patient care because this therapy may be accompanied by some serious side effects, such as CRS and ICANS, and also from the health provider standpoint because it is an expensive therapy. Thus, it would be helpful to identify biomarkers associated with CART19 clinical outcomes. Our study shows that the epigenetic profiling in CAR19-transduced T lymphocytes provides a consistent readout associated with clinical outcomes.

Our findings indicate that the intrinsic molecular features of the preinfusion cells determine the success of the adoptive cell therapy. In this regard, global RNA expression patterns of the preinfused T-cell differs between CR and non-CR patients (6), an observation added to the impact on outcome of the CAR integration site (8). All these findings support the finding that the "fitness" of preinfused CART19 cells contributes to treatment effectiveness. In this regard, CART19 cell products that harbor particular T-cell subsets are more clinically effective (6). Differences in the conditions of the manufacturing process from commercially available treatments and the unique functional background of the transduced T cells of each patient can modify the "omics" landscape of preinfused cells, directly affecting their activity. Importantly, it has recently been reported that epigenetic remodeling can restore functionality in

exhausted CART cells (36), further supporting the impact of these changes.

Our results strengthen the notion that the molecular profiles of the cells used in adoptive cell therapy is of great value for determining treatment success. This approach has also been proposed for immune checkpoint inhibitors (13,14). Thus, biomarkers of the efficacy of adoptive cell therapy, similar to those cited here (5-9,37), and the DNA methylation markers discovered in our study almost certainly await discovery. Two examples highlight the potential of studies in this area. One is the occurrence of T-cell receptor epigenetic inactivation associated with reduced tumor responsiveness in patients with melanoma and sarcoma infused with autologous T cells transduced with a retrovirus (14). Importantly, US Food and Drug Administration–approved CART19 treatments with axicabtagene ciloleucel and brexucabtagene autoleucel use retroviruses. A second pertinent study used single-cell RNA sequencing, a technique recently applied for CAR T cells (38), to show that mural cells, which surround the endothelium maintaining blood-brain barrier integrity, express the CD19 antigen (39), which may explain the neurotoxicity observed in CART19 therapies (40).

Overall, we report that the DNA methylation landscape of preinfusion CART19 cells can predict which patients with a B-cell malignancy will gain a clinical benefit. Importantly for its proposed clinical use, the best of the candidate sites identified within our epigenomic signature could be assessed using single PCR-based assays. In this regard, although larger, prospective clinical studies are required to determine the final value of the DNA methylation loci identified here, assessing the epigenetic profile of the CAR19-transduced, preinfused T cells could help solve the unmet medical need to identify patients who would benefit the most from CAR T-cell therapy.

Funding

Supported by CERCA Programme/Generalitat de Catalunya, Health Department PERIS #SLT/002/16/00374, AGAUR-project #2017SGR1080; MCI/AEI/ERDF project #RTI2018-094049-B-I00; ERC EPIPHARM; Cellex Foundation; “la Caixa” Foundation (LCF/PR/GN18/51140001 and LCF/PR/GN18/50310007), RF-2016–02364388, Accelerator Award—Cancer Research UK/AIRC—INCAR Associazione Italiana Ricerca per la Ricerca sul Cancro (AIRC) Project 5 × 1000 no. 9962, AIRC IG 2018 id. 21724, AIRC MFAG id. 21769 and id. 20450; MIUR (Grant PRIN 2017); and RCR-2019–23669115.

Notes

Role of the funder: The funder had no role in the design of the study; the collection, analysis, and interpretation of the data; or the writing and the submission of the manuscript.

Disclosures: ME is a consultant to Ferrer and Quimatryx. The other authors have no disclosures.

Author contributions: Conceptualization: CAGP, LV, and ME. Data curation: CAGP, LV, ABC, VD, EAGN, MJ, AUI, JD, VOM, FdB, FL, CQ, MSi, MSo, MCdM, GF, RGU, AFF, MFF, DB, AM, MB, AA, EJ, and ME. Formal analysis: CAGP, LV, ABC, VD, MCdM, and ME. Funding acquisition: ME. Software: CAGP, MCdM. Supervision: ME. Investigation: All authors. Writing—original draft: All authors. Writing—review and editing: All authors.

Data Availability

All relevant data are shown in the main manuscript and the Supplementary Materials. DNA methylation data are available in the Gene Expression Omnibus repository (GSE179414, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179414).

References

1. Elsallab M, Levine BL, Wayne AS, Abou-El-Enein M . CAR T-cell product performance in haematological malignancies before and after marketing authorisation. *Lancet Oncol*. 2020;21(2):e104–e116.
2. Singh AK, McGuirk JP. CAR T cells: continuation in a revolution of immunotherapy. *Lancet Oncol*. 2020;21(3):e168–e178.
3. Kersten MJ, Spanjaart AM, Thieblemont C. CD19-directed CAR T-cell therapy in B-cell NHL. *Curr Opin Oncol*. 2020;32(5):408–417.
4. Malard F, Mohty M. Acute lymphoblastic leukaemia. *Lancet*. 2020;395(10230):1146–1162.
5. Majzner RG, Mackall CL. Tumor antigen escape from CAR T-cell therapy. *Cancer Discov*. 2018;8(10):1219–1226.
6. Fraietta JA, Lacey SF, Orlando EJ, et al. Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia. *Nat Med*. 2018;24(5):563–571.
7. Fraietta JA, Nobles CL, Sammons MA, et al. Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. *Nature*. 2018;558(7709):307–312.
8. Nobles CL, Sherrill-Mix S, Everett JK, et al. CD19-targeting CAR T cell immunotherapy outcomes correlate with genomic modification by vector integration. *J Clin Invest*. 2020;130(2):673–685.
9. Rossi J, Paczkowski P, Shen YW, et al. Preinfusion polyfunctional anti-CD19 chimeric antigen receptor T cells are associated with clinical outcomes in NHL. *Blood*. 2018;132(8):804–814.
10. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. *Nat Rev Genet*. 2019;20(2):109–127.
11. Moran S, Martínez-Cardús A, Sayols S, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol*. 2016;17(10):1386–1395.
12. Villanueva L, Álvarez-Errico D, Esteller M. The contribution of epigenetics to cancer immunotherapy. *Trends Immunol*. 2020;41(8):676–691.
13. Duruisseaux M, Martínez-Cardús A, Calleja-Cervantes ME, et al. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: a multicentre, retrospective analysis. *Lancet Respir Med*. 2018;6(10):771–781.
14. Nowicki TS, Farrell C, Morselli M, et al. Epigenetic suppression of transgenic T-cell receptor expression via gamma-retroviral vector methylation in adoptive cell transfer therapy. *Cancer Discov*. 2020;10(11):1645–1653.
15. Ortíz-Maldonado V, Rives S, Castellà M, et al. CART19-BE-01: a multicenter trial of ARI-0001 cell therapy in patients with CD19+ relapsed/refractory malignancies. *Mol Ther*. 2021;29(2):636–644.
16. Jacoby E, Bielorai B, Avigdor A, et al. Locally produced CD19 CAR T cells leading to clinical remissions in medullary and extramedullary relapsed acute lymphoblastic leukemia. *Am J Hematol*. 2018;93(12):1485–1492.
17. Itzhaki O, Jacoby E, Nissani A, et al. Head-to-head comparison of in-house produced CD19 CAR-T cell in ALL and NHL patients. *J Immunother Cancer*. 2020;8(1):e000148.
18. Quintarelli C, Guercio M, Manni S, et al. Strategy to prevent epitope masking in CAR.CD19+ B-cell leukemia blasts. *J. Immunother. Cancer*. 2021;9(6):e001514.
19. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*. 2016;8(3):389–399.
20. Yao S, Sukonnik T, Kean T, Bharadwaj RR, Pasceri P, Ellis J. Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications. *Mol Ther*. 2004;10(1):27–36.
21. Lee DW, Santomasso BD, Locke FL, et al. ASTCT consensus grading for cytokine release syndrome and neurologic toxicity associated with immune effector cells. *Biol Blood Marrow Transplant*. 2019;25(4):625–638.
22. Stunnenberg HG, Hirst M; International Human Epigenome Consortium. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell*. 2016;167(5):1145–1149.
23. Xu Y, Zhang M, Ramos CA, et al. Closely related T-memory stem cells correlate with in vivo expansion of CAR.CD19-T cells and are preserved by IL-7 and IL-15. *Blood*. 2014;123(24):3750–3759.
24. Singh N, Perazzelli J, Grupp SA, Barrett DM. Early memory phenotypes drive T cell proliferation in patients with pediatric malignancies. *Sci Transl Med*. 2016;8(320):320ra3.
25. Sadelain M, Rivière I, Riddell S. Therapeutic T cell engineering. *Nature*. 2017;545(7655):423–431.
26. Gattinoni L, Klebanoff CA, Restifo NP. Paths to stemness: building the ultimate antitumour T cell. *Nat Rev Cancer*. 2012;12(10):671–684.
27. Deng Q, Han G, Puebla-Osorio N, et al. Characteristics of anti-CD19 CAR T cell infusion products associated with efficacy and toxicity in patients with large B cell lymphomas. *Nat Med*. 2020;26(12):1878–1887.

28. Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. 2016;166(3):740–754.
29. Murtha M, Esteller M. Extraordinary cancer epigenomics: thinking outside the classical coding and promoter box. *Trends Cancer*. 2016;2(10):572–584.
30. Nagel S, Pommerenke C, Meyer C, Kaufmann M, MacLeod RAF, Drexler HG. Identification of a tumor suppressor network in T-cell leukemia. *Leuk Lymphoma*. 2017;58(9):2196–2207.
31. D'Andrea A, Gritti I, Nicoli P, et al. The mitochondrial translation machinery as a therapeutic target in Myc-driven lymphomas. *Oncotarget*. 2016;7(45):72415–72430.
32. Donati G, Brighenti E, Vici M, et al. Selective inhibition of rRNA transcription downregulates E2F-1: a new p53-independent mechanism linking cell growth to cell proliferation. *J Cell Sci*. 2011;124(pt 17):3017–3028.
33. Ji Z, Fan Z, Zhang Y, et al. Thiamine deficiency promotes T cell infiltration in experimental autoimmune encephalomyelitis: the involvement of CCL2. *J Immunol*. 2014;193(5):2157–2167.
34. Li X, Yang B, Wang L, Chen L, Luo X, Liu L. SPAG6 regulates cell apoptosis through the TRAIL signal pathway in myelodysplastic syndromes. *Oncol Rep*. 2017;37(5):2839–2846.
35. Holthof LC, Stikvoort A, van der Horst HJ, et al. Bone marrow mesenchymal stromal cell-mediated resistance in multiple myeloma against NK cells can be overcome by introduction of CD38-CAR or TRAIL-variant. *Hemasphere*. 2021;5(5):e561.
36. Weber EW, Parker KR, Sotillo E, et al. Transient rest restores functionality in exhausted CAR-T cells through epigenetic remodeling. *Science*. 2021; 372(6537):eaba1786.
37. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*. 2017;168(4):707–723.
38. Wang W, Fasolino M, Cattau B, et al. Joint profiling of chromatin accessibility and CAR-T integration site analysis at population and single-cell levels. *Proc Natl Acad Sci U S A*. 2020;117(10):5442–5452.
39. Parker KR, Migliorini D, Perkey E, et al. Single-cell analyses identify brain mural cells expressing CD19 as potential off-tumor targets for CAR-T immunotherapies. *Cell*. 2020;183(1):126–142.e17.
40. Hunter BD, Jacobson CA. CAR T-cell associated neurotoxicity: mechanisms, clinicopathologic correlates, and future directions. *J Natl Cancer Inst*. 2019; 111(7):646–654.

Supplementary materials

SUPPLEMENTARY MATERIALS

Supplementary Methods

Supplementary Table 1. List of primers for pyrosequencing (Pyro), bisulfite sequencing (Bseq) and qRT-PCR (qPCR) used in the study.

Supplementary Table 2. Annotation of the 984 differentially methylated CpG sites. (Available for separate download as an .xls file).

Supplementary Table 3. Annotation of the 45 CpGs associated to Complete Response.

Supplementary Table 4. Annotation of the 8 CpGs associated to Cytokine Release Syndrome (CRS).

Supplementary Table 5. Annotation of the 5 CpGs associated to Immune Effector Cell-Associated Neurotoxicity Syndrome (ICANS).

Supplementary Table 6. Annotation of the 18 CpGs associated to Complete Response with FDR adjusted *P* values <.05.

Supplementary Table 7. Data derived from Kaplan-Meir analyses of event-free survival and overall survival associated to complete response and DNA methylation signature (EPICART) in discovery, validation and entire cohort of patients with B-cell malignancy treated with CART19 therapy.

Supplementary Figure 1. Gene Ontology (GO) analysis of genes with CpGs that changed upon CAR transduction and CAR retroviral vector DNA methylation analysis.

Supplementary Figure 2. Use of the EPICART signature in the supervised hierarchical clustering for the discovery cohort.

Supplementary Figure 3. Flow cytometry plots showing CD45RA and CCR7 expression and impact cell populations in clinical outcome.

Supplementary Figure 4. CpG methylation and gene expression data, including how hypermethylation of the CpG sites at the gene bodies of *INPP5A* and *ECHDC1* is associated with transcript and protein upregulation.

Supplementary Figure 5. Use of the EPICART signature in the supervised hierarchical clustering for the validation and entire cohorts, and impact on EFS and OS for the entire cohort.

Supplementary Figure 6. Hypermethylation of the identified CpG sites located at the 5'-end regulatory region of *SPAG6* and *PTCD3* is associated with gene downregulation in T-cell derived lines.

Supplementary Methods

CART19 cells

Patient material was obtained as part of the previously reported clinical trials NCT02772198, NCT03144583 and NCT03373071 evaluating three different academic CD19-specific Chimeric Antigen Receptor (CAR) T cells.

NCT02772198 (1,2) was approved by the Sheba Medical Center IRB and the Israeli Ministry of Health. CAR-T-cells were produced as previously described (2). Briefly, peripheral-blood mononuclear cells (PBMCs) were isolated from a fresh leukapheresis product and activated in T-cell medium. On day 2 of culture, activated cells were transduced with the CD19 CAR retrovirus, which was kindly provided by Dr. Steven Rosenberg. This construct comprises the variable regions of anti-CD19 monoclonal antibody FMC63 fused to the CD28 costimulatory domain and to the CD3 zeta chain, which were cloned into a mouse stem-cell virus gamma-retroviral (MSGV) backbone. CAR-T-cells were then further expanded in IL-2 containing T-cell medium until day 9–10. High molecular weight DNA was extracted from paired T-cell samples consisting of CART19 untransduced and transduced T-cells before infusion into patients.

NCT03144583 samples were provided by Hospital Clinic of Barcelona as part of the mentioned clinical trial developed in adult and pediatric patients with relapse and refractory CD19+ B-leukemia and lymphoma. Autologous CAR-T CD19 cells were produced as previously described (3). Succinctly, T cells were selected by CliniMACs Prodigy ® system (Miltenyi Biotec) from apheresis products and culture in IL-7 and IL-15 containing media. 24 hours upon activation cells were transduced with a lentivirus expressing the anti-hCD19 A3B1 monoclonal antibody conjugated with the costimulatory regions 4-1BB and CD3zeta chain. After expansion during 7-10 days, cells were cryopreserved prior to infusion.

NCT03373071 samples were obtained from patients enrolled in the described clinical trial conducted by IRCCS Ospedale Pediatrico Bambino Gesù in Rome, Italy. Pediatric patients with relapsed or refractory CD19+ Acute Lymphoblastic Leukemia (ALL) were treated with a second generation retroviral vector encompassing the variable regions

derived from the anti-CD19 monoclonal antibody FMC63 with a CD34 16aa peptide trackable marker in the hinge region fused to the 4-1BB costimulatory domain and the CD3 zeta chain and linked to the iC9 suicide gene (4). For CAR CD19 T-cells generation, autologous peripheral blood-derived mononuclear cells were activated using anti-CD3/anti-CD28 antibodies, transduced and subsequently expanded during 14 days in IL-7 and IL-15 containing media and at that point cells were cryopreserved until infusion.

DNA methylation and gene expression assays

DNA methylation status was determined using the Illumina MethylationEPIC BeadChip 850K microarray. Briefly, 600 ng of DNA of the studied samples was used to hybridize the BeadChip and scanned using HiScan SQ system (Illumina). Raw signal intensity data were initially QC'd and preprocessed from resulting idat files in R statistical environment (v4.0.3) using minfi Bioconductor package (v1.36.0). A number of quality control steps were applied to minimize errors and remove erratic probe signals, such as failed probes (detection P value > 0.01), cross-reacting probes, and probes that overlapped SNPs within ±1 bp of CpG sites, followed by background correction and dye-based normalization using ssNoob algorithm (single-sample normal-exponential out-of-band). Z-scores were used for gene expression data. The z-score indicates the number of standard deviations away from the mean of expression in a particular gene. All downstream analyses were performed under R statistical environment (v4.0.3). CpG methylation status for selected sites was validated by pyrosequencing and bisulfite genomic sequencing of multiple clones as previously described (5). Real-time quantitative PCR was developed as previously described (5). Primer sequences are shown in the Table below. Western-blot assays were performed as previously described (5). The antibodies used were INPP5A (Invitrogen Thermofisher, Ref PA5-28158), ECHDC1 (Invitrogen Thermofisher, Ref PA5-43232) and SPAG6 (Abcam, Ref ab155653). For DNA demethylating treatments, the T-cell derived lines H9 and KARPAS-45 were incubated with 1 μM 5-aza-2'-deoxycytidine (AZA; Sigma) during 72h.

Flow cytometry

The following antibodies were used: CD3 (VioBlue; Miltenyi Biotech or Pacific blue and PE; BioLegend), CD45RA (APC-Vio770; Miltenyi Biotec or Brilliant Violet; BioLegend) and CCR7 (PerCP-Vio770; Miltenyi Biotec or PerCP; BioLegend, San Diego, CA). CAR T-cells were washed and re-suspended in cell staining buffer (BioLegend, San Diego, CA). Cells were incubated for 30 minutes with the antibodies on ice, washed in buffer, and measured using FACS cytometer MACSQuant (Miltenyi Biotec). Samples were analyzed using FlowJo software (FlowJo LLC, Ashland, OR). Cell populations were defined as follows: Naïve T-cells (TN: CD3+CD45RA+CCR7+), central memory T-cells (TCM: CD3+CD45RA-CCR7+), effector memory T-cells (TEM: CD3+CD45RA-CCR7-) and effector T-cells (TEMRA: CD3+CD45RA+CCR7-). Significance of variation between the different T-cell phenotypes in CART19 samples was evaluated using a two-tailed Student's t-test. Equality of means was tested comparing the mean percentage of TN+TCM vs TEM+TEMRA T-cell populations in EPICART-positive and EPICART-negative CART19 cells. For the correlation of CART19 T-cell phenotypes with EFS and OS, we classified patients according to the predominant ratio of TN+TCM vs TEM+TEMRA T-cell populations. Thus, those CART19 samples with a larger proportion of naïve and central memory T-cells were classified within the TN+TCM subgroup, whereas those with a prevailing percentage of effector memory and effector T-cells were included within the TEM+TEMRA subgroup.

References

1. Jacoby E, Bielorai B, Avigdor A, et al. Locally produced CD19 CAR T cells leading to clinical remissions in medullary and extramedullary relapsed acute lymphoblastic leukemia. *Am. J. Hematol.* 2018;93(12):1485-1492.
2. Itzhaki O, Jacoby E, Nissani A, et al. Head-to-head comparison of in-house produced CD19 CAR-T cell in ALL and NHL patients. *J. Immunother. Cancer.* 2020;8(1):e000148.
3. Castella M, Caballero-Baños M, Ortiz-Maldonado V, et al. Point-Of-Care CAR T-Cell Production (ARI-0001) Using a Closed Semi-automatic Bioreactor: Experience From an Academic Phase I Clinical Trial. *Front. Immunol.* 2020;11:482.
4. Quintarelli C, Guercio M, Manni S, et al. Strategy to prevent epitope masking in CAR.CD19+ B-cell leukemia blasts. *J. Immunother. Cancer.* 2021;9(6):e001514.
5. Esteve-Puig R, Climent F, Piñeyro D, et al. Epigenetic loss of m1A RNA demethylase ALKBH3 in Hodgkin lymphoma targets collagen, conferring poor clinical outcome. *Blood.* 2021;137(7):994-999.

Supplementary Table 1. List of primers for pyrosequencing (Pyro), bisulfite sequencing (Bseq) and qRT-PCR (qPCR) used in the study[a]

Primer ID[a]	Sequence (5' - 3')[a]
INPP5A_cg25268100_Pyro_Fw	[Btn]TTTGGGTTTGAAGGTAGTGGG
INPP5A_cg25268100_Pyro_Rv	ATAAACCCCTCCCTCCTAA
INPP5A_cg25268100_Pyro_Seq	CCCCTCCCTCCTAAA
SPAG6_cg12610471_Pyro_Fw	TTTAGATAATTTTAGGGTTGTAATTT
SPAG6_cg12610471_Pyro_Rv	[Btn]AATATCCCTACACTAC
SPAG6_cg12610471_Pyro_Seq	GTTTTGTAAGGAGTTT
ECHDC1_cg25571136_Pyro_Fw	[Btn]GTTATGGGATTTTTATGAATAGGATGATTA
ECHDC1_cg25571136_Pyro_Rv	CCAACCCAACTTAAAATCTTCTTTTTTATA
ECHDC1_cg25571136_Pyro_Seq	ACTTCTTAAAACATACAATCAA
INPP5A_cg25268100_Bseq_F	AGGGAGAAGTGTATTGTTTGG
INPP5A_cg25268100_Bseq_R	TATAAACATAACCCACCTCCC
SPAG6_cg12610471_Bseq_F	AAGTTTAGATAATTTTAGGGTTGT
SPAG6_cg12610471_Bseq_R	AACTACTAAAACTCTCAA
ECHDC1_cg25571136_Bseq_F	TGGATTAGATTGATAGTAGTGAGT
ECHDC1_cg25571136_Bseq_R	ACCAAATCACCTACATTTAAA
INPP5A_qPCR_Fw	AGAACTATTGTCGAGTGATGCGA
INPP5A_qPCR_Rv	GCTTCCTAGTGCCGTGAAGT
SPAG6_qPCR_Fw	GTGCGACATTCTTCCACAGC
SPAG6_qPCR_Rv	TCCAGTGCTCCACAATCGAC
ECHDC1_qPCR_Fw	GTTCAAGGTTGGGCATTGGG
ECHDC1_qPCR_Rv	GCCACCCCAGCTTGGTATTA
PTCD3_qPCR_Fw	CTCCGCAGCAGGCTTGG
PTCD3_qPCR_Rv	ACCTTTGAGAGGGTTGCACT
GUSB_qPCR_Fw	TGGTTGGAGAGCTCATTTGGA
GUSB_qPCR_Rv	GCACTCTCGTCGGTGACTGTT
CAR vector_5'LTR_Pyro_Fw	GGGTGTTTTAAGGATTTGAAATGATTTTG
CAR vector _5'LTR_Pyro_Rv	[Btn]AAAAACCCTCCCAAAAATCAAC
CAR vector _5'LTR_Pyro_Seq	TGATTTTGTGTTTTATTTGAATTAA
CAR vector_psi_gag_Pyro_Fw	GGGTTATTTTTTGTTTTGTAGAATGG
CAR vector _psi_gag_Pyro_Rv	[Btn]AAACCAAAACTTCCCAAATCAC
CAR vector _psi_gag_Pyro_Seq	TTTTTGTTTTGTAGAATGGTTA
CAR vector_ insert_Pyro_Fw	GTTTAGTGGTAGTGGGTTTGGAATAGAT
CAR vector_insert_Pyro_Rv	[Btn]CCAAACCAAATCCTAACTCCTACAA
CAR vector_insert_Pyro_Seq	ATTTATTTTTGTTAATAGGGTA

[a]Abbreviations: [Btn] = biotin; Fw = forward; Rv = reverse; Seq = sequencing.

Supplementary Table 2. Annotation of the 984 differentially methylated CpG sites. (Available for separate download as an .xls file).

Supplementary Table 2 can be accessed on the publication site at: https://doi.org/10.1093/jnci/djab194

Supplementary Table 3. Annotation of the 45 CpGs associated to Complete Response[a]

Probe ID[b]	Chromosomal position (hg19)[c]	Associated gene[d]	Complete Response *P* value[e]
cg12012941	chr1:188676237	Not described	<.001
cg04267686	chr6:105907265	Not described	<.001
cg25534076	chr1:234087867	*SLC35F3*	<.001
cg25571136	chr6:127612751	*ECHDC1*	<.001
cg10039734	chr10:95139986	*MYOF*	<.001
cg12260379	chr2:86332162	*PTCD3;POLR1A*	<.001
cg01311063	chr2:131058184	Not described	<.001
cg12504912	chr14:90081872	*FOXN3*	<.001
cg10236435	chr12:123944014	*SNRNP35*	<.001
cg09992216	chr11:32353565	Not described	<.001
cg25268100	chr10:134457731	*INPP5A*	<.001
cg25995980	chr10:46993515	*GPRIN2*	<.001
cg12610471	chr10:22634199	*SPAG6*	<.001
cg15253304	chr6:209809	Not described	<.001
cg17511575	chr2:122144477	*CLASP1*	<.001
cg09367268	chr6:6643814	*LY86*	<.001
cg11416737	chr18:60877850	*BCL2*	<.001
cg24267358	chr19:42299379	*CEACAM3*	<.001
cg22171055	chr1:62905816	*USP1*	.001
cg04458195	chr1:220414164	*RAB3GAP2*	.001
cg03593578	chr2:45028225	Not described	.002
cg05948940	chr16:68481342	*SMPD3*	.002
cg26098972	chr12:131166906	Not described	.003
cg01029450	chr22:43253559	*ARFGAP3*	.003
cg19759671	chr4:183063459	*MGC45800*	.004
cg18739950	chr15:95870440	Not described	.005
cg14780466	chr2:20870812	*GDF7*	.006
cg02775469	chr6:33181031	Not described	.009
cg12448747	chr3:12898045	Not described	.009
cg13554177	chr6:79780164	*PHIP*	.009
cg26934960	chr16:87228921	Not described	.01
cg12197459	chr17:686450	*GLOD4;RNMTL1*	.01
cg10549986	chr2:7018153	*RSAD2*	.01
cg20017856	chr14:29990921	*MIR548AI*	.02
cg09234616	chr11:32452592	*WT1*	.02
cg01140143	chr6:33039396	*HLA-DPA1*	.02
cg06689619	chr4:99935464	*METAP1*	.02
cg18989133	chr11:17518482	*USH1C*	.02

cg25606201	chr5:180614858	Not described	.03
cg13546658	chr2:162164472	*PSMD14*	.03
cg26346210	chr8:98610507	Not described	.03
cg03216691	chr12:123466396	*ARL6IP4*	.04
cg27196695	chr10:134571377	*INPP5A*	.04
cg06354455	chr13:114054873	Not described	.046
cg25654695	chr19:2273216	*OAZ1*	.046

[a]This annotation was retrieved from the Infinium MethylationEPIC Array manifest.

[b]Probe ID = unique identifier from the Illumina CG database.

[c]Chromosomal position (hg19): chromosomal coordinates of the CpG (build hg19).

[d]Associated gene: target gene name from the UCSC database.

[e]The *P* value of the Complete Response is derived from the Fisher´s exact test (CR vs NR/SD/PD). All tests were 2-sided.

Supplementary Table 4. Annotation of the 8 CpGs associated to Cytokine Release Syndrome (CRS)[a]

Probe ID[b]	Chromosomal position (hg19)[c]	Associated gene[d]	CRS *P* value[e]
cg21847720	chr8:2075777	*MYOM2*	.01
cg01311063	chr2:131058184	Not described	.02
cg00994804	chr21:36259383	*RUNX1*	.02
cg25606201	chr5:180614858	Not described	.03
cg26669806	chr19:18899483	*COMP*	.04
cg24365464	chr1:190448126	*FAM5C*	.04
cg14538944	chr2:218340518	*DIRC3*	.04
cg22836400	chr6:10415636	*TFAP2A*	.04

[a]Annotation retrieved from the Infinium MethylationEPIC Array manifest.

[b]Probe ID: unique identifier from the Illumina CG database.

[c]Chromosomal position (hg19): chromosomal coordinates of the CpG (build hg19).

[d]Associated gene: target gene name from the UCSC database.

[e]The *P* value of the CRS is derived from the Fisher´s exact test (CRS grade 0 vs grades 1-5). All tests were 2-sided.

Supplementary Table 5. Annotation of the 5 CpGs associated to Immune Effector Cell-Associated Neurotoxicity Syndrome (ICANS)[a]

Probe ID[b]	Chromosomal position (hg19)[c]	Associated gene[d]	ICANS *P* value[e]
cg01311063	chr2:131058184	Not described	<.001
cg26195366	chr10:102242535	*WNT8B*	.01
cg22534145	chr20:23015936	*SSTR4*	.04
cg27272679	chr8:65294635	Not described	.04
cg27196695	chr10:134571377	*INPP5A*	.046

[a]Annotation retrieved from the Infinium MethylationEPIC Array manifest.

[b]Probe ID: unique identifier from the Illumina CG database.

[c]Chromosomal position (hg19): chromosomal coordinates of the CpG (build hg19).

[d]Gene name: target gene name from the UCSC database.

[e]The *P* value of the ICANS is derived from the Fisher´s exact test (ICANS grade 0 vs grades 1-5). All tests were 2-sided.

Supplementary Table 6. Annotation of the 18 CpGs associated to Complete Response with FDR adjusted *P* values < .05[a]

Probe ID[b]	Chromosomal position (hg19)[c]	Associated gene[d]	Complete Response FDR *P*[e]
cg12012941	chr1:188676237	Not described	.001
cg04267686	chr6:105907265	Not described	.001
cg25534076	chr1:234087867	*SLC35F3*	.002
cg10039734	chr10:95139986	*MYOF*	.007
cg25571136	chr6:127612751	*ECHDC1*	.007
cg01311063	chr2:131058184	Not described	.01
cg12260379	chr2:86332162	*PTCD3;POLR1A*	.01
cg12504912	chr14:90081872	*FOXN3*	.01
cg10236435	chr12:123944014	*SNRNP35*	.01
cg09992216	chr11:32353565	Not described	.01
cg25268100	chr10:134457731	*INPP5A*	.01
cg25995980	chr10:46993515	*GPRIN2*	.01
cg12610471	chr10:22634199	*SPAG6*	.02
cg15253304	chr6:209809	Not described	.02
cg17511575	chr2:122144477	*CLASP1*	.02
cg09367268	chr6:6643814	*LY86*	.03
cg11416737	chr18:60877850	*BCL2*	.04
cg24267358	chr19:42299379	*CEACAM3*	.04

[a]Annotation retrieved from the Infinium MethylationEPIC Array manifest.

[b]Probe ID: unique identifier from the Illumina CG database.

[c]Chromosomal position (hg19): chromosomal coordinates of the CpG (build hg19).

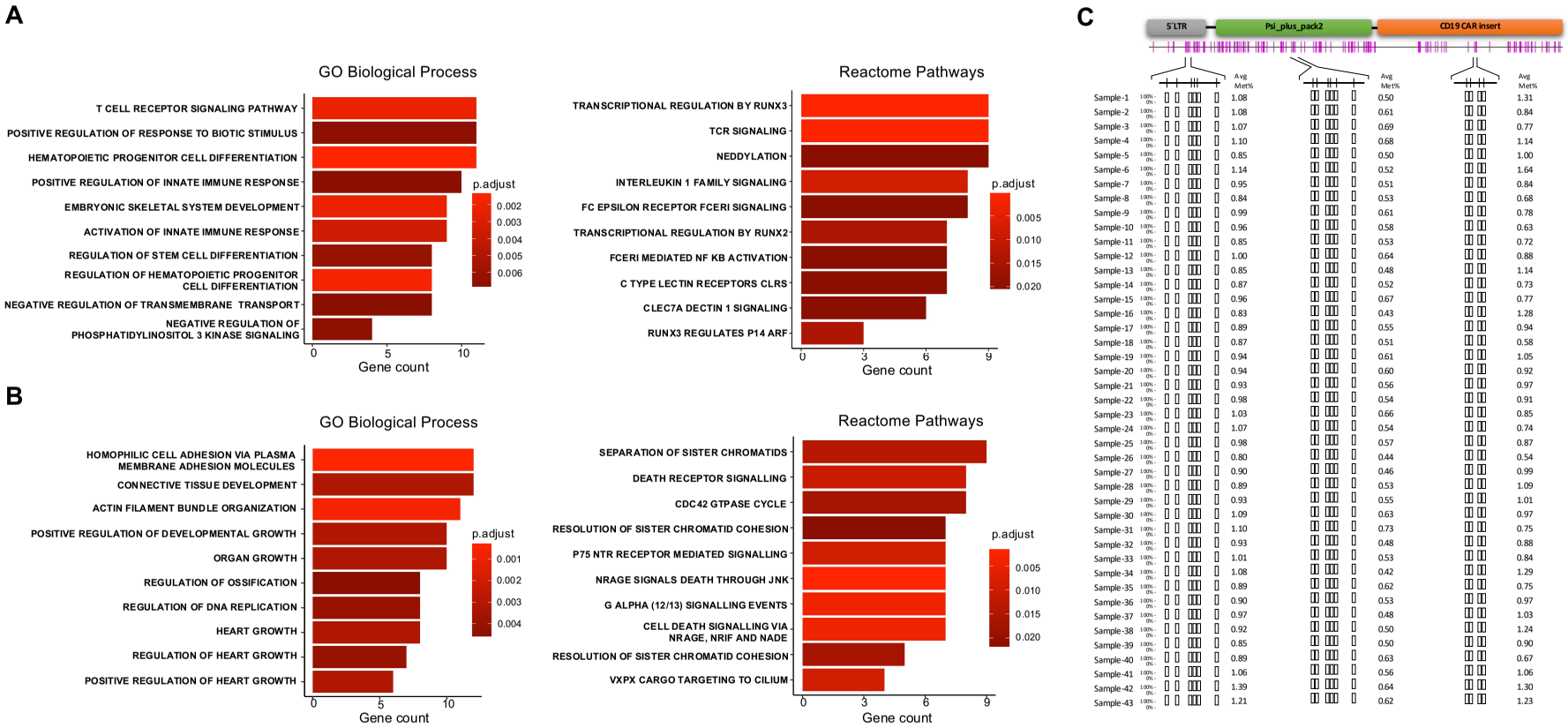
[d]Associated gene: target gene name from the UCSC database.

[e]The False Discovery Rate (FDR) adjusted *P* value of the Complete Response is derived from the Fisher´s exact test (CR vs NR/SD/PD). All tests were 2-sided.

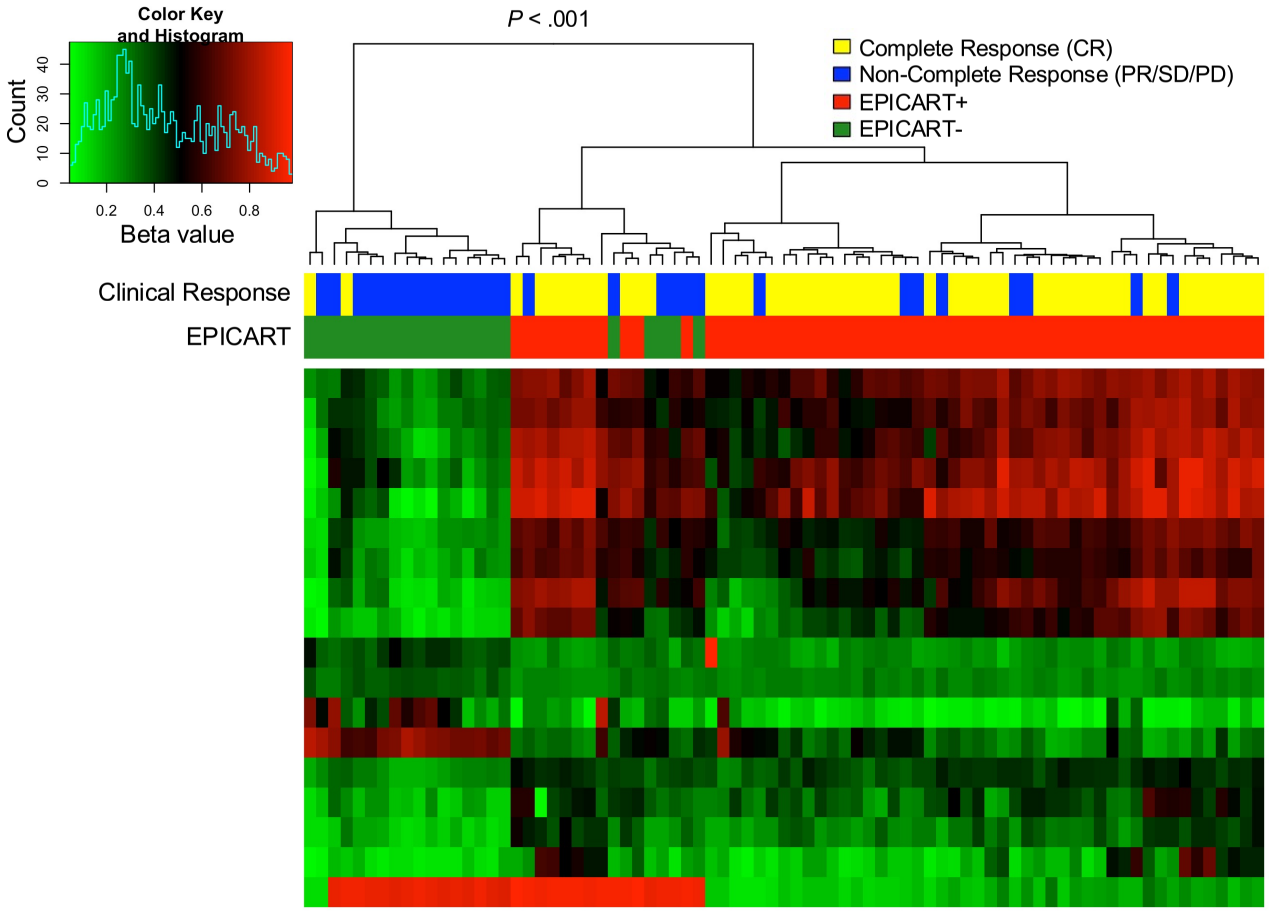
Supplementary Table 7. Data derived from Kaplan-Meir analyses of event-free survival and overall survival associated to complete response and DNA methylation signature (EPICART) in discovery, validation, and entire cohort of patients with B-cell malignancy treated with CART19 therapy.[a]

Cohort comparisons	Event-free survival			Overall survival		
	Log-rank *P* value	HR (95% CI)	HR *P* value	Log-rank *P* value	HR (95% CI)	HR *P* value
CR vs non-CR (PR/SD/PD)						
Discovery	<.001	0.12 (0.06 to 0.24)	<.001	<.001	0.18 (0.09 to 0.39)	<.001
Validation	.002	0.24 (0.09 to 0.62)	.003	<.001	0.11 (0.04 to 0.37)	<.001
Entire	<.001	0.15 (0.09 to 0.26)	<.001	<.001	0.18 (0.10 to 0.32)	<.001
EPICART+ vs EPICART-						
Discovery	.003	0.36 (0.19 to 0.70)	.002	.04	0.45 (0.20 to 0.99)	.047
Validation	.19	0.52 (0.20 to 1.35)	0.18	.02	0.31 (0.11 to 0.84)	.02
Entire	.003	0.43 (0.26 to 0.74)	.002	.003	0.39 (0.21 to 0.74)	.003

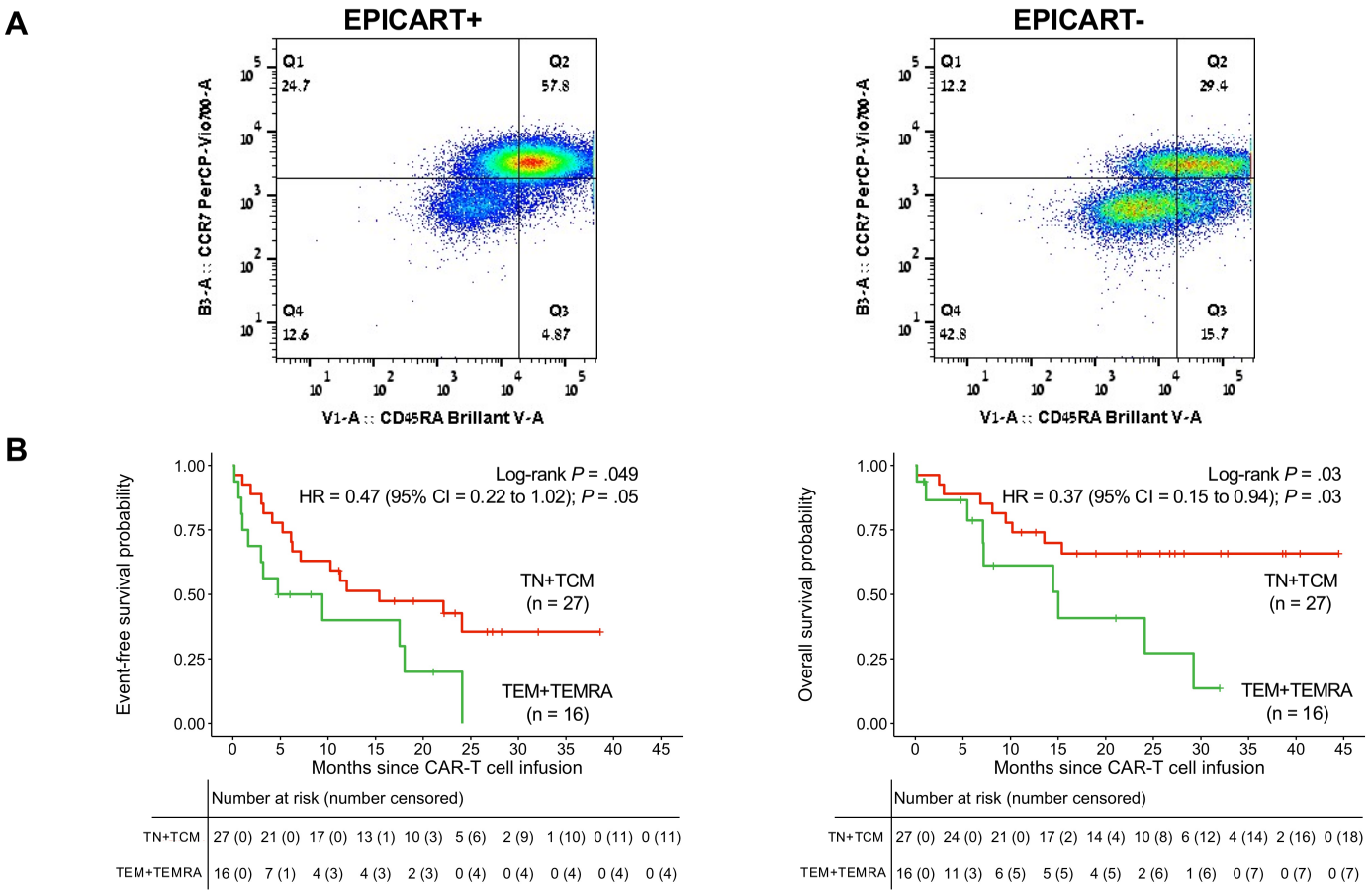
[a]All statistical tests were 2-sided. HR = hazard ratio; CI = confidence interval; CR = complete response; PR = partial response; SD = stable disease; PD = progression of the disease.



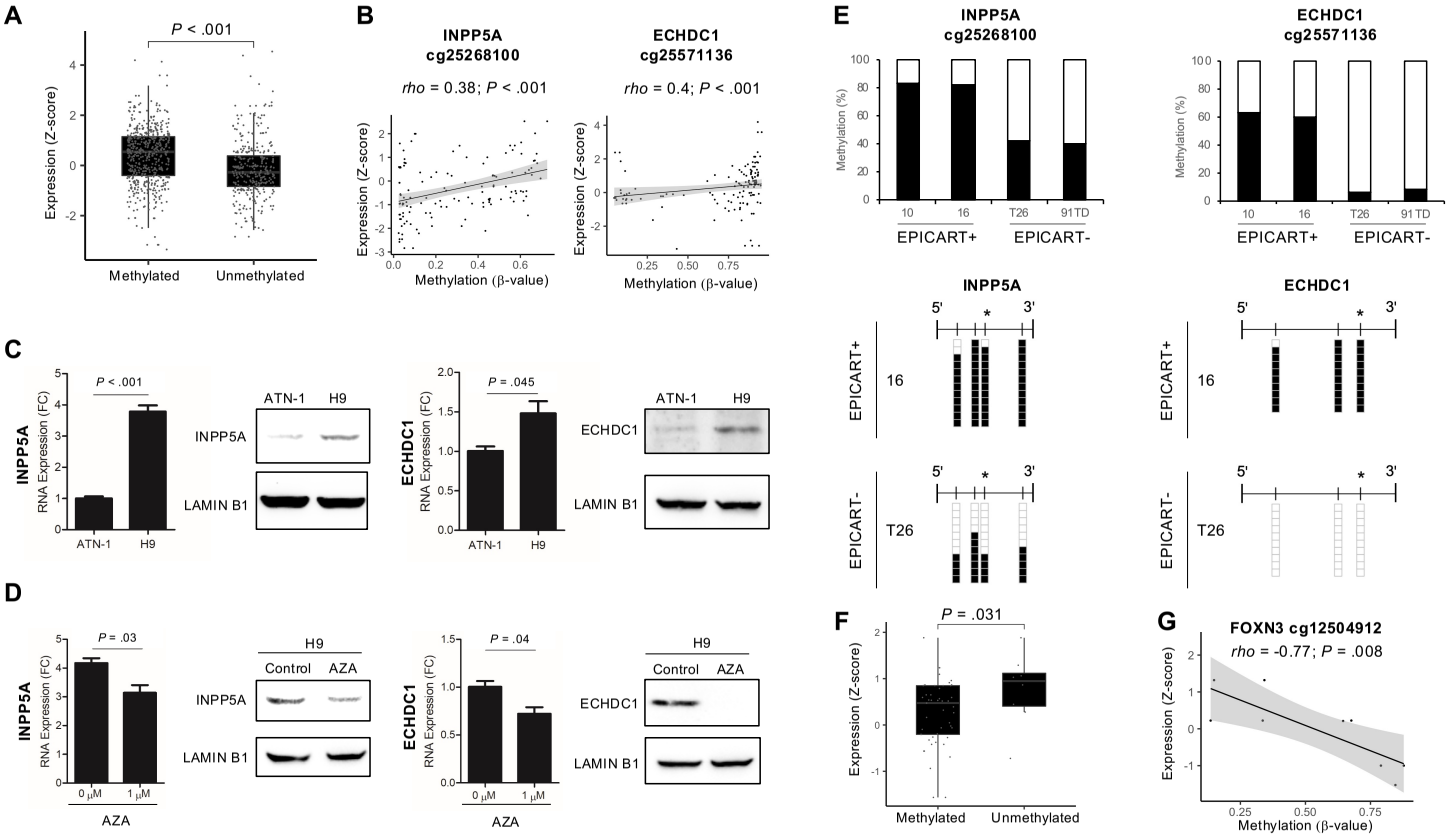
Supplementary Figure 1. Gene Ontology (GO) analysis of genes with CpGs that changed upon CAR transduction and CAR retroviral vector DNA methylation analysis. **A.** GO for genes associated with CpGs located within regulatory regions (TSS 1500, TSS 200, 5'UTR, 1st exon, 3'UTR). Over-representation analysis with one-sided Fisher's exact test FDR adjusted *P*<.05. TSS = transcription start site; UTR = untranslated region. **B.** GO for genes associated with CpGs located within the gene body. Over-representation analysis with one-sided Fisher's exact test FDR adjusted *P*<.05 **C.** Diagram representing CD19 CAR vector regions analyzed by bisulfite pyrosequencing in the NCT02772198 trial. The location of each CpG site is represented by vertical lines. Bars show the percentage of DNA methylation from 0 to 100, where black indicates methylated CpG sites, and white unmethylated CpG sites. Percentage of average methylation for each region is shown on the right. *P* values <.05 were considered to be statistically significant.



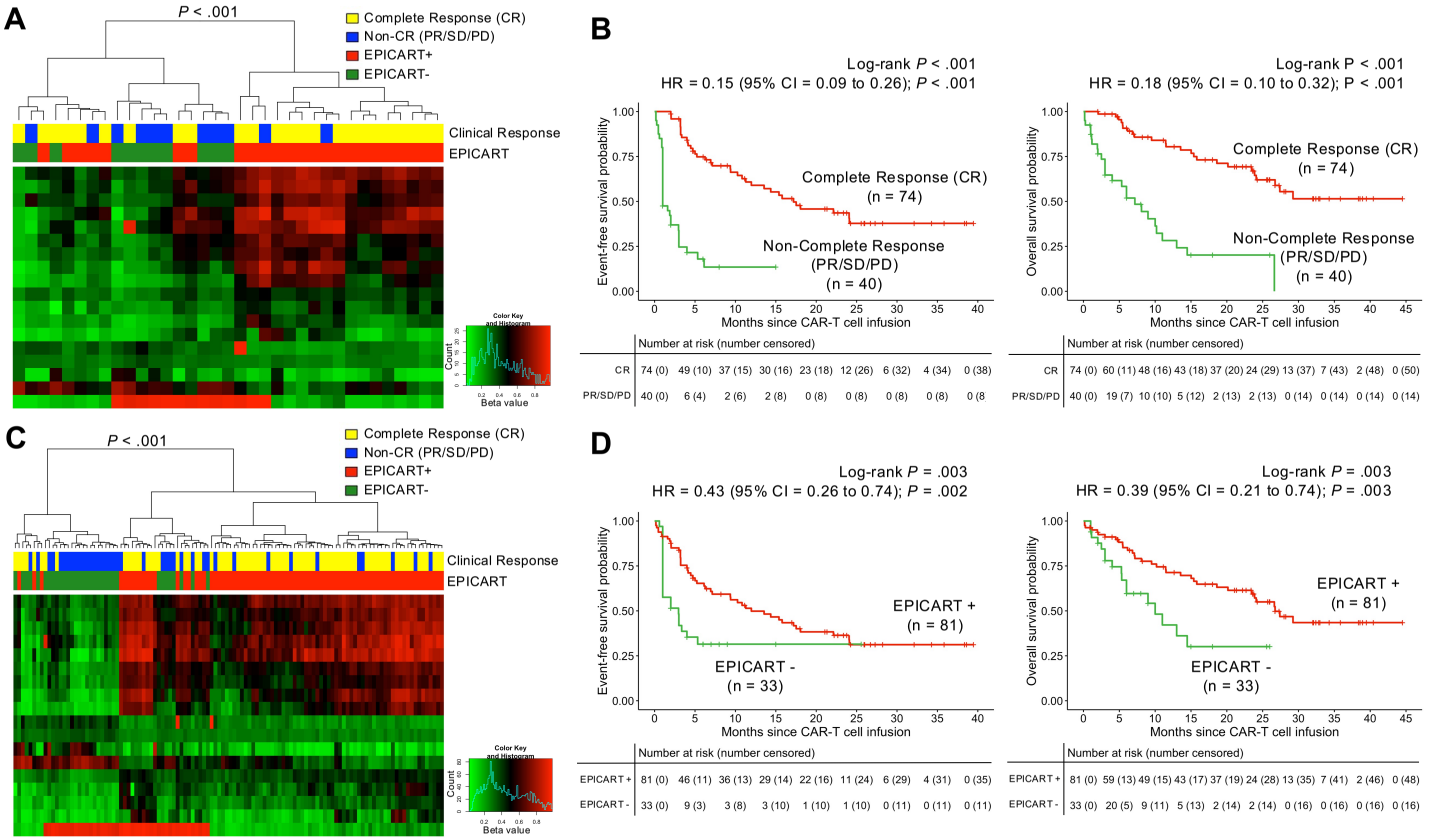
Supplementary Figure 2. Use of the EPICART signature in the supervised hierarchical clustering for the discovery cohort (n=79) of CAR-T cases confirmed the existence of two branches that classified patients as those exhibiting CR or non-CR (2-sided Fisher's exact test, $P <$ .001). Hierarchical clustering was performed using Euclidean distances and the *Ward's* minimum variance agglomeration method. CR: complete response; PR: partial response; SD: stable disease; PD: progression of the disease. *P* value < .05 was considered to be statistically significant.



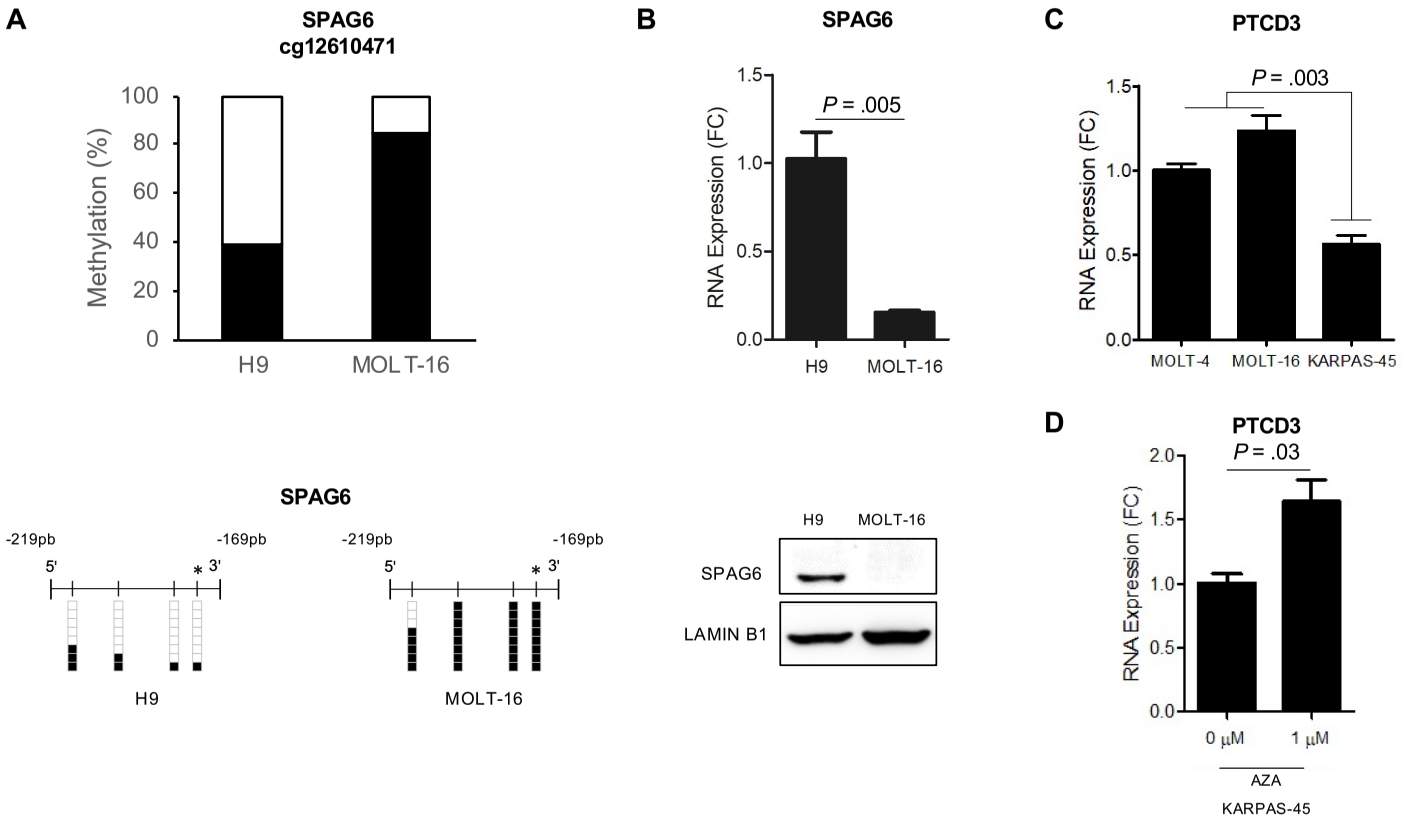
Supplementary Figure 3. Flow cytometry analyses and cell populations. **A.** Flow cytometry plots showing CD45RA and CCR7 expression with quadrants indicating naïve T cells (Q2: CD45RA+CCR7+), central memory T cells (Q1: CD45RA-CCR7+), effector memory T cells (Q4: CD45RA-CCR7-) and effector T cells (Q3: CD45RA+CCR7-). Illustrative EPICART positive and negative CART19 samples from two patients are shown in the left and right, respectively. **B.** Kaplan–Meier analysis of event-free survival (EFS; left) and overall survival (OS; right) in the cohort of 43 B-cell malignancy patients with flow cytometry data available according to the expression of CD45RA and CCR7 in the pre-infused CART19 cells, defined by the predominant ratio of naïve (TN+TCM) vs effector (TEM+TEMRA) T cells . TN = naïve T cells; TCM = central memory T cells; TEM = effector memory T cells; TEMRA = effector T cells. The number of events is also shown. *P* value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR) with a 95% confidence interval (95% CI). *P* values < .05 were considered to be statistically significant. All statistical tests were 2-sided.



Supplementary Figure 4. CpG methylation and gene expression data. **A.** Gene body CpG hypermethylation is significantly associated with transcript upregulation in the panel of blood cell lines. One-sided Mann-Whitney-Wilcoxon test was performed to compare RNA expression between methylated and unmethylated cell lines. **B**. One-sided Spearman´s test between methylation of the identified CpGs located at the gene bodies of *INPP5A* and *ECHDC1* vs. expression in blood derived cell lines. **C**. RNA and protein expression of *INPP5A* (left) and *ECHDC1* (right) in the T-cell derived ATN-1 (hypomethylated) and H9 (hypermethylated) cell lines, analyzed by qRT-PCR (triplicates, one-sided Student´s t-test) and western blot, respectively. **D**. *INPP5A* and *ECHDC1* RNA expression by qRT-PCR (triplicates, one-sided Student´s t-test) and protein expression by western blot upon treatment with the DNA demethylating agent 5-aza-2'-deoxycytidine (AZA) in the hypermethylated H9 cell line. In western blots, Lamin B1 was used as loading control. **E.** CpGs differentially methylated between EPICART+ and EPICART- cells, illustrated by the *INPP5A* and *ECHDC1* genes, that were validated by pyrosequencing (upper panel) and bisulfite genomic sequencing (lower panel). In pyrosequencing graphs, bars show the percentage of DNA methylation from 0 to 100. In bisulfite genomic sequencing graphs, single clones are shown for each sample; presence of a methylated or unmethylated cytosine is indicated by a black or white square, respectively; and CpGs interrogated in EPIC array are indicated by asterisk. **F.** 5'-end CpG hypermethylation and transcript downregulation in T-cell lines. One-sided Mann-Whitney-Wilcoxon test was used to compare RNA expression between methylated and unmethylated cell lines. **G.** Hypermethylation of the 5'-UTR CpG site of *FOXN3* is associated with transcript downregulation. One-sided Spearman's test was performed. *P* values <.05 were considered to be statistically significant.



Supplementary Figure 5. EPICART signature in validation and entire cohorts. **A.** Use of the EPICART signature in the supervised hierarchical clustering for the validation cohort (n=35) of CAR-T cases confirmed the existence of two branches that classified patients as those exhibiting complete response (CR) or non-CR (2-sided Fisher's exact test *P*<.001). **B.** Kaplan–Meier analysis of event-free survival (EFS; left) and overall survival (OS; right) in the entire cohort of 114 B-cell malignancy patients according to the presence of complete response or its absence (PR + SD + PD). **C.** Use of the EPICART signature in the supervised hierarchical clustering for the entire cohort (n=114) of CAR-T cases confirmed the existence of two branches that classified patients as those exhibiting CR or non-CR (2-sided Fisher's exact test *P*<.001). **D.** Kaplan–Meier analysis of EFS (left) and OS (right) in the entire cohort of B-cell malignancy patients according to the presence of the EPICART signature in the pre-infused CART19 cells, defined by the methylation status of the 18 CpG sites associated with CR (EPICART-positive [+] signature). CR: complete response; PR: partial response; SD: stable disease; PD: progression of the disease. In A and C, hierarchical clustering was performed using Euclidean distances and the *Ward's* minimum variance agglomeration method. In B and D, *P* value was calculated using the log-rank function. Univariate Cox regression analysis is represented as the hazard ratio (HR) with a 95% confidence interval (95% CI). The number of events is also shown. *P* values <.05 were considered to be statistically significant. All statistical tests were 2-sided.



Supplementary Figure 6. Hypermethylation of the identified CpG sites located at the 5'-end regulatory regions of *SPAG6* and *PTCD3* is associated with gene downregulation in T-cell derived lines. **A**. Methylation of the *SPAG6* promoter-associated CpG cg12610471 in H9 and MOLT-16 cell lines. Upper panel: Pyrosequencing, bars show the percentage of DNA methylation from 0 to 100. Lower panel: Bisulfite genomic sequencing, single clones are shown for each sample; presence of a methylated or unmethylated cytosine is indicated by a black or white square, respectively; and the CpGs interrogated in EPIC array are indicated by asterisk. **B**. *SPAG6* expression in H9 and MOLT-16 cell lines. Upper panel: RNA expression by qRT-PCR (triplicates, one-sided Student´s t-test). Lower panel: protein expression by western blot (Lamin B1 was used as loading control). **C**. One-sided Student´s t-test between *PTCD3* RNA expression determined by qRT-PCR (triplicates) in unmethylated (MOLT-4 and MOLT-16) vs methylated (KARPAS-45) cell lines. **D**. Upregulation of *PTCD3* RNA expression upon the use of the DNA demethylating agent 5-aza-2'-deoxycytidine (AZA) in the methylated cell line KARPAS-45. One-sided Student´s t-test was performed. *P* values <.05 were considered to be statistically significant.

Chapter III | Epigenetic profiling linked to multisystem inflammatory syndrome in children (MIS-C): A multicenter, retrospective study

Epigenetic profiling linked to multisystem inflammatory syndrome in children (MIS-C): A multicenter, retrospective study

Veronica Davalos,[a,1] Carlos A. García-Prieto,[a,b,1] Gerardo Ferrer,[a,c] Sergio Aguilera-Albesa,[d] Juan Valencia-Ramos,[e] Agustí Rodríguez-Palmero,[f,g,h] Montserrat Ruiz,[f,g] Laura Planas-Serra,[f,g] Iolanda Jordan,[i] Iosune Alegría,[d] Patricia Flores-Pérez,[j] Verónica Cantarín,[j] Victoria Fumadó,[k] Maria Teresa Viadero,[l] Carlos Rodrigo,[h] Maria Méndez-Hernández,[h] Eduardo López-Granados,[g,m] Roger Colobran,[n] Jacques G. Rivière,[o] Pere Soler-Palacín,[o] Aurora Pujol,[f,g,p,] and Manel Esteller [a,c,p,q,**]*

[a]Josep Carreras Leukaemia Research Institute (IJC), Badalona, Barcelona, Catalonia, Spain
[b]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Catalonia, Spain
[c]Centro de Investigación Biomédica en Red de Cancer (CIBERONC), Spain
[d]Navarra Health Service Hospital, Pamplona, Spain
[e]University Hospital of Burgos, Burgos, Spain
[f]Neurometabolic Diseases Laboratory, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Catalonia, Spain
[g]Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Spain
[h]Germans Trias i Pujol Research Institute (IGTP), Universitat Autònoma de Barcelona (UAB), Badalona, Barcelona, Spain
[i]Pediatric Critical Care Unit, Hospital Universitari Sant Joan de Deu, Barcelona, Catalonia, Spain
[j]Pediatrics Department, Hospital Universitario Niño Jesús, Madrid, Spain
[k]Unitat de Malalties Infeccioses i Importades, Servei de Pediatría, Infectious and Imported Diseases, Pediatric Unit, Hospital Universitari Sant Joan de Deú, Barcelona, Catalonia, Spain
[l]Servicio de Pediatría del Hospital Universitario Marqués de Valdecilla, Santander, Spain
[m]Department of Immunology, La Paz University Hospital, Madrid, Spain; La Paz Institute of Biomedical Research, Madrid, Spain
[n]Immunology Division, Department of Clinical and Molecular Genetics, Hospital Universitari Vall d'Hebron (HUVH), Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain
[o]Pediatric Infectious Diseases and Immunodeficiencies Unit, Hospital Universitari Vall d'Hebron (HUVH), Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain
[p]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[q]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Catalonia, Spain

Summary

Background Most children and adolescents infected with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) remain asymptomatic or develop a mild coronavirus disease 2019 (COVID-19) that usually does not require medical intervention. However, a small proportion of pediatric patients develop a severe clinical condition, multisystem inflammatory syndrome in children (MIS-C). The involvement of epigenetics in the control of the immune response and viral activity prompted us to carry out an epigenomic study to uncover target loci regulated by DNA methylation that could be altered upon the appearance of MIS-C.

Methods Peripheral blood samples were recruited from 43 confirmed MIS-C patients. 69 non-COVID-19 pediatric samples and 15 COVID-19 pediatric samples without MIS-C were used as controls. The cases in the two groups were mixed and divided into discovery (MIS-C = 29 and non-MIS-C = 56) and validation (MIS-C = 14 and non-MIS-C = 28) cohorts, and balanced for age, gender and ethnic background. We interrogated 850,000 CpG sites of the human genome for DNA methylation variants.

Findings The DNA methylation content of 33 CpG loci was linked with the presence of MIS-C. Of these sites, 18 (54.5%) were located in described genes. The top candidate gene was the immune T-cell mediator ZEB2; and others

*Corresponding author at: Neurometabolic Diseases Laboratory, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, 08908 Barcelona, Catalonia, Spain.
**Corresponding author at: Josep Carreras Leukaemia Research Institute (IJC), Carretera de Can Ruti, Camí de les Escoles s/n, 08916 Badalona, Barcelona, Catalonia, Spain.
E-mail addresses: apujol@idibell.cat (A. Pujol), mesteller@carrerasresearch.org (M. Esteller).

[1] These authors contributed equally.

highly ranked candidates included the regulator of natural killer cell functional competence SH2D1B; VWA8, which contains a domain of the Von Willebrand factor A involved in the pediatric hemostasis disease; and human leukocyte antigen complex member HLA-DRB1; in addition to pro-inflammatory genes such as CUL2 and AIM2. The identified loci were used to construct a DNA methylation profile (EPIMISC) that was associated with MIS-C in both cohorts. The EPIMISC signature was also overrepresented in Kawasaki disease patients, a childhood pathology with a possible viral trigger, that shares many of the clinical features of MIS-C.

Interpretation We have characterized DNA methylation loci that are associated with MIS-C diagnosis. The identified genes are likely contributors to the characteristic exaggerated host inflammatory response observed in these patients. The described epigenetic signature could also provide new targets for more specific therapies for the disorder.

Funding Unstoppable campaign of Josep Carreras Leukaemia Foundation, Fundació La Marató de TV3, Cellex Foundation and CERCA Programme/Generalitat de Catalunya.

Keywords: Multisystem inflammatory syndrome in children; COVID-19; Kawasaki disease; Epigenetics; DNA methylation

Research in context

Evidence before this study

Most members of the pediatric population infected with the SARS-CoV-2 virus, which is responsible for the COVID-19 pandemic, escape severe disease. However, in a few cases, a rare and serious health condition, known as multisystem inflammatory syndrome in children (MIS-C), may occur. The clinical spectrum of MIS-C can affect multiple organ systems, often requiring admission to intensive care unit. Risk factors for the disease are not well defined, and the hyperinflammatory condition resembles another rare disorder known as Kawasaki disease. To our knowledge, this is the first epigenomic study of MIS-C after acute SARS-CoV-2 infection. Our search of PubMed on January 20th, 2022, limited to articles in English, but not by date, using the terms "MIS-C", "epigenomics", "DNA methylation", and "marker", identified no studies addressing this topic.

Added value of this study

Our results indicate the existence of distinct DNA methylation loci that distinguish MIS-C patients from COVID-19 pediatric patients without MIS-C, and from healthy children and adolescents without SARS-CoV-2 infection. The epigenetic sites found were mostly located within genes associated with immune response and pro-inflammatory pathways. Taking advantage of these DNA methylation markers, we produced an epigenomic profile that exhibited great accuracy in predicting MIS-C diagnosis. We have named this profile the EPIMISC.

Implications of all the available evidence

Our research has revealed new biomarkers linked to MIS-C onset that provide new information about the pathophysiological mechanisms of the disorder, and highlight its close similarity to Kawasaki disease. The genes identified could also be candidate targets for more precise treatments of the disease. Most importantly, the assessment of the DNA methylation levels of these loci can be swiftly added to the measurement of other biochemical and clinical parameters to improve early MIS-C diagnosis.

Introduction

In late 2019, an unexpected increase in the number of pneumonia cases in China led to the identification of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2),[1] and the subsequent worldwide spread of the derived disease, termed COVID-19. At the time of the writing (January 24th, 2022), more than 350 million confirmed cases and more than 5,6 million deaths have been reported worldwide (https://coronavirus.jhu.edu/map.html). High mortality of COVID-19 patients with serious respiratory failure linked to acute respiratory distress syndrome (ARDS) and interstitial pneumonia have been associated with male sex, old age and concomitant medical conditions, such as diabetes, obesity, hypertension, and cardiovascular pathology.[2] In comparison to the presentation in adults, most children and adolescents with SARS-CoV-2 infection are fully asymptomatic or have very mild clinical manifestations.[3] The severity of the disease in the pediatric population also depends on their underlying conditions, and children may manifest ARDS and pneumonia as do adults.[3] However, a more specific complication appeared in April 2020,[4] a new and rare syndrome, termed

multisystem inflammatory syndrome in children (MIS-C). This is also known as pediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2 infection (PIMS-TS). MIS-C arises days to weeks after the initial infection.[4–10] Unlike severe adult COVID-19 patients, who are characterized by respiratory failure, MIS-C patients show a broad spectrum of additional clinical features (e.g., rash, fever, abdominal and/or chest pain, conjunctival hyperemia, etc.) as a result of multiple organ involvement (e.g., the cardiovascular, gastrointestinal, mucocutaneous, or hematological systems, amongst others). Although it is a rare disease, MIS-C is a serious health condition that require admission to intensive care unit in around 60% of cases, and ultimately lead to death to a not negligible 2% of cases.[5] The exact pathways that give rise to the clinical manifestations of MIS-C, and the factors predisposing to development of the disease are largely unknown.[4–10]

In the context of adult COVID-19, in addition to the aforementioned concomitant medical conditions,[2] genetic studies suggest that several genetic loci are associated with the severity of the disease (summarized in Supplementary Materials). We have also recently shown that epigenetic variation, particularly DNA methylation, which is altered in many human diseases,[11] is also associated with adult COVID-19 severity.[12] DNA and RNA viral activity are controlled by DNA methylation changes,[11] but more importantly, this epigenetic mark is key to proper immune system activity and could predict the efficacy of immune-related therapies.[13] To investigate if epigenetic changes are involved in MIS-C, we undertook a comprehensive epigenomic study to identify candidate DNA methylation loci linked to the disease that distinguish these patients from standard COVID-19 pediatric patients and from SARS-CoV-2-uninfected children and adolescent subjects of the pre-COVID-19 era.

Methods

Study design and participants

Whole blood samples and clinical data from 43 patients with MIS-C were previously collected between April 16th, 2020 and August 17th, 2021 from seven Hospitals in Spain. MIS-C was diagnosed based on the case definition provided by the World Health Organization. Briefly, children with clinical and biochemical evidence of inflammation in at least two systems, without other cause, and with evidence of SARS-CoV2 infection or close contact. The complete description is available in https://www.who.int/news-room/commentaries/detail/multisystem-inflammatory-syndrome-in-children-and-adolescents-with-covid-19. Clinicopathological characteristics of the MIS-C patients studied are summarized in Table 1. Whole blood samples were also obtained from 15 pediatric COVID-19 cases with no evidence of MIS-C, and from 69 healthy children and adolescents collected during the pre-COVID-19 era (before December 2019), in the setting of routine surgical procedures such as circumcision, orchiopexy, inguinal or umbilical hernia repair, adenoidectomy, tonsillectomy or tympanic membrane incision; or from unaffected sibling controls collected in previous studies. The data from these samples are summarized in Tables S1 and S2, respectively. The protocol of this retrospective study was approved by the ethics review boards of the participating institutions. Written informed consent was obtained from all participants. The study protocol is described in the Supplementary Methods.

DNA methylation data and computational analyses

The DNA methylation status of the studied samples was obtained using the Infinium MethylationEPIC Array (~850,000 CpG sites) (Supplementary Methods). The MIS-C epigenetic signature, referred to hereafter as EPIMISC, was obtained by first identifying the probes differentially methylated between MIS-C cases and healthy control donors, filtering out in a second step those probes found to be differentially methylated between pediatric COVID-19 cases and healthy controls (Figure S1). This approach enabled us to effectively discover the differentially methylated probes between MIS-C and non-MIS-C cases. This involved deriving a linear model adjusted by the age covariate with the limma R package (v3.46.0), using the methylation values of the discovery dataset. A significance threshold for CpGs with a False Discovery Rate (FDR) adjusted *P*-value <0.05 and an absolute mean methylation beta difference between groups of >0.15 was established. The linear model was adjusted by the age covariate after performing a principal component analysis (PCA) that identified disease status and age as the greatest sources of variation in our dataset (Supplementary Methods). The significantly differential DNA methylation sites (Table S3) were used to train a supervised classification model based on a ridge-regularized logistic regression to predict MIS-C diagnosis using the glmnet R package (v4.1-1). The classification model was optimized by tuning parameters (best performance with alpha = 0 from ridge regression, and regularization parameter lambda = 0.1) after resampling with 10-fold cross-validation carried out three times using the caret package in R (v6.0-86). Once the model and tuning parameters values have been defined after resampling, our model performance was assessed using the receiver operating characteristic (ROC) and calibration curves. Further details are provided in Supplementary Methods.

Role of funders

Funded by the Josep Carreras Leukaemia Foundation, the Cellex Foundation and the CERCA Programme of

Characteristics	MIS-C cohorts		
	Discovery cohort (*N* = 29)	Validation cohort (*N* = 14)	Entire cohort (*N* = 43)
Gender - Frequency (%)			
Female	9 (31.0%)	8 (57.1%)	17 (39.5%)
Male	20 (69.0%)	6 (42.9%)	26 (60.5%)
Age (years) - Median [range]	8.0 [0.5—17]	6.5 [1—11]	7.0 [0.5—17]
Age group- Frequency (%)			
≤2 yr	5 (17.2%)	1 (7.1%)	6 (14.0%)
3—5 yr	4 (13.8%)	5 (35.7%)	9 (20.9%)
6—9 yr	7 (24.1%)	5 (35.7%)	12 (27.9%)
10—13 yr	8 (27.6%)	3 (21.4%)	11 (25.6%)
14—18 yr	5 (17.2%)	0 (0.0%)	5 (11.6%)
Ethnicity - Frequency (%)			
West-Eurasia	20 (69.0%)	11 (78.6%)	31 (72.1%)
Central-South America	6 (20.7%)	2 (14.3%)	8 (18.6%)
African	3 (10.3%)	1 (7.1%)	4 (9.3%)
Underlying conditions - Frequency (%)			
Previously healthy	26 (89.7%)	14 (100%)	40 (93.0%)
Asthma	3 (10.3%)	0 (0%)	3 (7.0%)
SARS-CoV-2 status- Frequency (%)			
IgG and/or PCR positive	26 (89.7%)	13 (92.9%)	39 (90.7%)
Near contact with SARS-CoV-2 positive*	3 (10.3%)	1 (7.1%)	4 (9.3%)
Detection of additional virus - Frequency (%)†	5 (17.2%)	0 (0.0%)	5 (11.6%)
Organ system involvement - Frequency (%)			
Two systems	4 (13.8%)	2 (14.3%)	6 (14.0%)
Three systems	5 (17.2%)	6 (42.9%)	11 (25.6%)
Four or more systems	20 (69.0%)	6 (42.9%)	26 (60.5%)
Gastrointestinal involvement # - Frequency (%)	26 (89.7%)	11 (78.6%)	37 (86.0%)
Respiratory involvement # - Frequency (%)	21 (72.4%)	6 (42.9%)	27 (62.8%)
Cardiovascular involvement # - Frequency (%)	22 (75.9%)	11 (78.6%)	33 (76.7%)
Mucocutaneous involvement # - Frequency (%)	19 (65.5%)	10 (71.4%)	29 (67.4%)
Hematologic involvement # - Frequency (%)	21 (72.4%)	8 (57.1%)	29 (67.4%)
Neurologic involvement # - Frequency (%)	6 (20.7%)	2 (14.3%)	8 (18.6%)
Renal involvement # - Frequency (%)	3 (10.3%)	0 (0.0%)	3 (7.0%)
Musculoskeletal involvement # - Frequency (%)	1 (3.4%)	1 (7.1%)	2 (4.7%)
Highest level of care - Frequency (%)			
Home	1 (3.4%)	0 (0.0%)	1 (2.3%)
Ward	8 (27.6%)	9 (64.3%)	17 (39.5%)
Intensive care unit	20 (69.0%)	5 (35.7%)	25 (58.1%)
Oxygen supplementation - Frequency (%)			
None	9 (31.0%)	10 (71.4%)	19 (44.2%)
Nasal cannula	7 (24.1%)	1 (7.1%)	8 (18.6%)
Non-Invasive Ventilation or High Flow Oxygen	10 (34.5%)	2 (14.3%)	12 (27.9%)
Mechanical Ventilation	2 (6.9%)	1 (7.1%)	3 (7.0%)
Extracorporeal membrane oxygenation	1 (3.4%)	0 (0%)	1 (2.3%)

***Table 1*: Clinicopathological characteristics of the studied MIS-C patients.**

* For cases of unknown SARS-CoV-2 status.

† Additional viruses: Parainfluenza virus type 4 (HPIV-4), Rhinovirus/Enterovirus (HRV/ENT) and Adenovirus.

Following the definitions used for organ involvement in Feldstein et al., *N Engl J Med*, 2020.

the Generalitat de Catalunya. Additional support was provided by the Fundació La Marató de TV3 (202131-32-33), MCIU/AEI/FEDER (RTI2018-094049-B-I00) and AGAUR (2017SGR1080). The sponsors of the study had no role in the study design, data collection, data analysis, data interpretation, or the writing of the manuscript. The authors collected the data, and had full access to all of the data in the study. They also took the final decision and had responsibility for submitting the study results for publication.

Characteristics	Cohorts		
	Discovery cohort (*N* = 85)	Validation cohort (*N* = 42)	Entire cohort (*N* = 127)
Cases - Frequency (%)			
MIS-C	29 (34.1%)	14 (33.3%)	43 (33.9%)
Non-MIS-C	56 (65.9%)	28 (66.7%)	84 (66.1%)
Gender - Frequency (%)			
Female	36 (42.4%)	22 (52.4%)	58 (45.7%)
Male	49 (57.6%)	20 (47.6%)	69 (54.3%)
Age (years) - Median [range]	9.0 [0–17]	7.5 [0–16]	8.0 [0 - 17]
Age group- Frequency (%)			
≤2 yr	11 (12.9%)	3 (7.1%)	14 (11.0%)
3–5 yr	11 (12.9%)	13 (31.0%)	24 (18.9%)
6–9 yr	22 (25.9%)	12 (28.6%)	34 (26.8%)
10–13 yr	23 (27.1%)	7 (16.7%)	30 (23.6%)
14–18 yr	18 (21.2%)	7 (16.7%)	25 (19.7%)
Ethnicity - Frequency (%)			
West-Eurasia	61 (71.8%)	25 (59.5%)	86 (67.7%)
Central-South America	7 (8.2%)	4 (9.5%)	11 (8.7%)
African	4 (4.7%)	2 (4.8%)	6 (4.7%)
Unknown	13 (15.3%)	11 (26.2%)	24 (18.9%)

***Table 2*: Clinicopathological characteristics of the discovery and validation cohorts.**

Results

Patients and epigenomic study

Between April 16th, 2020 and August 17th, 2021, we obtained whole blood samples from 43 patients diagnosed with MIS-C, using the case definition provided by the World Health Organization and summarized in Methods. Table 1 lists the clinicopathological features of the MIS-C patients studied. MIS-C-associated laboratory findings for these cases are summarized in Figure S2. Overall, the median age was 7.0 years old (Interquartile range, IQR = 7), and the majority of the children were male (26 cases; 60.5%) and from a West-Eurasian ethnic background (31 cases, 72.1%). Most patients exhibited a previous healthy status (40 cases, 93%) and IgG and/or PCR positivity for SARS-CoV-2 (39 cases, 90.7%). Most cases had affectation of four or more of their organ systems (26 cases, 60.5%), and were admitted to an intensive care unit (25 cases, 58.1%). As previously described in other MIS-C series, only a few cases presented prominent respiratory symptoms that required mechanical ventilation (3 cases, 7%), in contrast to the classic severe COVID-19 illness, which often requires active and interventional oxygen supplementation. We also collected whole blood samples from 15 pediatric COVID-19 patients with IgG-positive and/or PCR-positive status for SARS-CoV-2, but without MIS-C (Table S1). Finally, we obtained whole blood samples from 69 children collected before December 2019, when the COVID-19 disease first appeared (Table S2).

To optimize our analyses, we compared the MIS-C group (n = 43) with the non-MIS-C group (n = 84). The latter group comprised the pediatric COVID-19 cases without MIS-C (n = 15) and the pediatric controls obtained before the COVID-19 pandemic (n = 69). The 127 samples collected were divided into discovery and validation cohorts (85 and 42 cases, respectively) (Table 2). There were no significant differences between the two cohorts with respect to the frequencies of MIS-C and non-MIS-C cases (Fisher's exact test, P = 1), gender (Fisher's exact test, P = 0.345), age (Mann–Whitney–Wilcoxon test, P = 0.282) and ethnicity (Fisher's exact test, P = 0.579) (Table 2). DNA from the whole blood samples was purified for all cases and analyzed to determine DNA methylation status. The study aimed to characterize those genomic sites with a distinct DNA methylation status in MIS-C patients compared with the non-MIS-C population. The overall study design is illustrated in Figure S1.

Epigenomic analysis of MIS-C in the discovery cohort

Using the experimental and bioinformatic pipeline shown in Figure S1 and described in Supplementary Methods, the DNA methylation analysis of 85 pediatric individuals in the discovery cohort identified 33 CpG sites with a distinct methylation status between MIS-C (n = 29) and non-MIS-C (n = 56) cases (Table S3). The Volcano plot of the fully adjusted P-values from the DNA methylation loci linked to MIS-C diagnosis in the discovery cohort is shown in Figure 1. The genomic

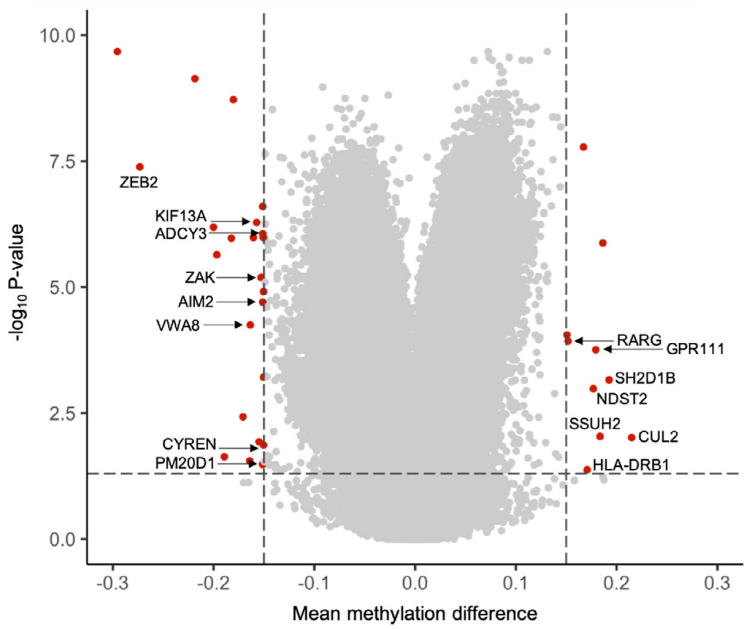


Figure 1. The volcano plot shows significant differences in the DNA methylation status of 850K CpG sites between MIS-C and non-MISC using the described experimental and bioinformatic pipeline. Y-axis shows the -$\log_{10}$ *P*-value and X-axis shows the mean methylation difference according to beta value. A total of 33 CpGs with a delta beta >0.15 and FDR adjusted *P*-value <0.05 are shown in red. For those with an associated coding sequence, the gene name is also indicated. CpG-sites that exhibited a methylation beta value difference <0.15 and/or FDR adjusted *P*-value >0.05 are shown in grey. Dashed lines indicate cut-offs for significance.

annotation of these differentially methylated 33 CpG sites is described in Table S3. Fifteen (45.45%) of the identified sites were located in regions of the genome with no currently annotated gene sequences; three (9.1%) were associated with three long non-coding RNAs (LINC00880, LOC645434, LOC100996286); and the other 15 (45.45%) CpG loci were located within 15 known protein-coding genes (Table 3).

To investigate further the activities of the 15 candidate coding genes identified by the MIS-C DNA methylation screening, we performed an enrichment analysis (Supplementary Methods). Significantly enriched Gene Ontology (GO) biological processes (hypergeometric test, FDR adjusted *P*-value < 0.05) included "regulation of inflammatory response to antigenic stimulus" and "regulation of immune response". All these enriched processes and pathways indicate that a broad exaggerated engagement of the immune response to the SARS-CoV-2 infection contributes to the characteristic hyperinflammatory clinical picture observed in these children.

Of the 15 candidate coding genes derived from the MIS-C epigenomic analysis, among the highest ranked coding genes according to the DNA methylation difference and adjusted *P*-value derived from the MIS-C epigenomic analysis (Table S3), the zinc finger E-box binding homeobox 2 (ZEB2) gene, the G protein-coupled receptor 111 (GPR111) gene, the SH2 domain containing 1B (SH2D1B) gene and the ubiquitin-protein ligase component Cullin-2 (CUL2) exhibit activities that could directly relate to MIS-C (Table 3). ZEB2 promotes terminal differentiation of effector and memory T cell populations during infections and the development of plasmacytoid dendritic cells, monocytes, B-cells, natural killer cells, and macrophages.[14] GPR111 is involved in tolerance induction, granulopoiesis and the control of cytotoxicity.[15] SH2D1B is a unique adaptor protein that enhances innate and adaptive immune responses to antigens.[16,17] In this regard, the SH2D1B signaling pathway has the potential to be co-opted to produce enhanced vaccination responses.[16] CUL2 is a mediator of inflammation and, in this regard, its pharmacological inhibition protects against hyperinflammatory responses,[18] a finding that could be relevant for those MIS-C patients that do not respond to the standard treatment.

Of the other genes with a distinct DNA methylation profile in MIS-C patients (Table 3), the cases of AIM2 (absent in melanoma 2) and PM20D1 (peptidase M20 domain-containing 1) are particularly interesting because methylation events at these loci are also characteristic of adults who develop severe COVID-19 disease.[12] The AIM2 gene is related to the hyperinflammatory manifestation of MIS-C patients since triggers

Gene symbol	Gene name	Gene function	Adjusted *P*-value
ADCY3	Adenylate cyclase 3	Catalyzes the synthesis of cyclic AMP (cAMP) from ATP. ADCY3 variants have been associated to risk/susceptibility to obesity, diabetes and chronic inflammatory diseases.	<0.001
AIM2	Absent in melanoma 2	Assembles the macromolecular inflammasome complex.	<0.001
CUL2	Cullin 2	Mediator of inflammation.	0.0096
CYREN	Cell cycle regulator of NHEJ	Cell-cycle-specific regulator of classical non-homologous end joining (NHEJ) of DNA double-strand break (DSB) repair.	0.0136
GPR111	G Protein-Coupled Receptor 111	Member of the adhesion G protein-coupled receptors (aGPCRs).	<0.001
HLA-DRB1	Major histocompatibility complex, class II, DR beta 1	Encodes a beta chain of antigen-presenting major histocompatibility complex class II (MHCII) molecule.	0.0421
KIF13A	Kinesin family member 13A	Motor protein that also mediates the trafficking of influenza A virus ribonucleoproteins, and transport of an arenavirus protein.	<0.001
NDST2	N-deacetylase and N-sulfotransferase 2	Enzyme with dual functions in processing glucosamine and heparin polymers.	0.0010
PM20D1	Peptidase M20 domain containing 1	Enzyme that regulates the production of N-fatty-acyl amino acids. Considered a metabolic disease-associated gene also linked to neurodegenerative disorders.	0.0328
RARG	Retinoic acid receptor gamma	Receptor for retinoic acid. Act as transcriptional regulator.	<0.001
SH2D1B	SH2 domain containing 1B	Adaptor protein for the signaling lymphocytic activation molecule family of receptors that enhances immune responses to antigens, including viral proteins such as HIV-Gag.	<0.001
SSUH2	Ssu-2 homolog	A putative chaperone protein.	0.0092
VWA8	Von Willebrand factor A domain-containing protein 8	Mitochondrial ATPase protein.	<0.001
ZAK	ZAK1 Homolog, Leucine Zipper And Sterile-Alpha Motif Kinase	Mitogen-activated protein kinase, also known as MAP3K20.	<0.001
ZEB2	Zinc finger E-box binding homeobox 2	ZEB2 is a DNA-binding transcriptional repressor.	<0.001

***Table 3*: Epigenetic changes in coding genes associated with MIS-C diagnosis.**

caspase-1 and unleashes pro-inflammatory cytokines such as IL-1β and IL-18,[19] which are also involved in the innate immune response to viral infections. Regarding PM20D1, recent data suggest that it contributes to autoimmune disorders and allergies,[20,21] all of which are pathologies with an important hyperinflammatory component. In this study, we have identified that a DNA methylation site of the HLA-DRB1 (major histocompatibility complex, class II, DR beta 1) gene is linked to MIS-C. Interestingly, our study of adult COVID-19 cases identified that an epigenetic mark in HLA-C (major histocompatibility complex, class I, C) was associated with the severe disease.[12] Importantly, allelic genotypes of HLA-DRB1 have been associated with the clinical severity of adult COVID-19 cases,[22–24] and $CD8^+$ T-cells from critically ill adult COVID-19 patients show upregulation of the HLA-DRB1 gene.[25]

We also investigated whether the DNA methylation status of MIS-C was distinct from that of non-MIS-C groups for genes that, according to the literature, are likely candidates for adult COVID-19. The 47 genes analyzed were the ACE2 receptor and TMPRSS2 protease, GWAS-derived genes, genes associated with inborn errors of type I IFN immunity in cases with life-threatening COVID-19, and other genes involved in immune host-cell pathways (Table S4). Only one gene, VWA8 (Von Willebrand factor A domain-containing protein 8), was shared in the list of COVID-19 associated loci

(Table S4) and in our MIS-C associated DNA methylation sites (Table 3). A single nucleotide polymorphism in VWA8 has been linked to hospitalized cases in COVID-19 cases.[26] For MIS-C genetic susceptibility very little is known. Three genes with reported sequence variants for MIS-C (SOCS1, XIAP and CYBB)[27] were not differentially methylated in our cohorts (Table S5), in line with the idea that, for the same candidate target, genetic and epigenetic alteration are usually mutually exclusive.

Testing MIS-C-associated DNA methylation markers in the validation cohort, and development of the EPIMISC signature

The DNA methylation status of single CpG sites linked to the presence of MIS-C in the discovery cohort ($n = 85$) was confirmed in the validation cohort ($n = 42$). Overall, when we individually analyzed the 33 CpGs whose DNA methylation levels differed significantly between the MIS-C and non-MIS-C cases, 20 (60.6%) were also significantly associated with the severe pediatric disorder in the validation cohort (Table S6). Of these 20 CpG sites, seven loci were located in the aforementioned gene coding-containing sequences (35%). Importantly, when we interrogated all the samples as an entire set, comprising the discovery and validation cohorts ($n = 127$), 24 of 33 (72.7%) individual CpG sites remained associated with MIS-C development (Table S6).

The discovery of single DNA methylation sites linked to the presence of MIS-C could be very helpful, but the establishment of an overall epigenomic signature could also be of great value to our understanding of the pathophysiological basis of the diseases and its clinical management. To achieve this, we selected the 33 significantly differential DNA methylation sites that were associated with the occurrence of MIS-C (Table S3) to train our discovery set, using a supervised classification model based on ridge-regularized logistic regression (see Supplementary Methods). By this method, we obtained a DNA methylation signature, hereafter referred to as EPIMISC, that was associated with MIS-C diagnosis (EPIMISC positive). It had a specificity of 98.21% (95% confidence interval CI = 90.45% to 99.95%), a sensitivity of 93.10% (95% CI = 77.23% to 99.15%), and positive and negative predictive values (PPV and NPV) of 96.43% (95% CI = 81.65% to 99.91%) and 96.49% (95% CI = 87.89% to 99.57%), respectively. Its accuracy was 96.47% (95% CI = 90.03% to 99.27%) and the Kappa value was 0.9208 (95% CI = 0.8329 to 1). We also plotted the Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve [AUC = 95.66% (95% CI = 90.65% to 100%)] together with the calibration curve to further assess and visualize the model's performance (Figure S3). Supervised hierarchical clustering using the EPIMISC signature differentiated two branches that were significantly enriched with respect to each condition, MIS-C *vs.* non-MIS-C (Fisher's exact test, $P = 4.3e\text{-}09$) (Figure S4). Most important, we found that the EPIMISC signature kept its value in our validation cohort, being associated with the disease with a specificity of 92.86% (95% CI = 76.50% to 99.12%), a sensitivity of 85.71% (95% CI = 57.19% to 98.22%), and PPV and NPV of 85.71% (95% CI = 57.19% to 98.22%) and 92.86% (95% CI = 76.50% to 99.12%), respectively. The accuracy was 90.48% (95% CI = 77.38% to 97.34%) and the Kappa value was 0.7857 (95% CI = 0.5865 to 0.9849). The ROC curve and AUC [89.29% (95% CI = 78.61% to 99.97%)] alongside the calibration curve were also determined (Figure S5). The EPIMISC signature in the validation cohort also distinguished two branches with respect to MIS-C and non-MIS-C samples (Fisher's exact test, $P = 3.1e\text{-}06$) (Figure S6). Finally, for the entire cohort, EPIMISC was associated with MIS-C diagnosis with a specificity of 96.43% (95% CI = 89.92% to 99.26%), a sensitivity of 90.70% (95% CI = 77.86% to 97.41%), and PPV and NPV of 92.86% (95% CI = 80.52% to 98.50%) and 95.29% (95% CI = 88.39% to 98.70%), respectively. Its accuracy was 94.49% (95% CI = 88.97% to 97.76%) and the Kappa value was 0.8762 (95% CI = 0.7872 to 0.9653). The ROC curve and AUC (93.56% [95% CI = 88.74% to 98.39%)] alongside the calibration curve were also determined (Figure S7). The application of the EPIMISC signature for the entire cohort also classified samples as MIS-C or non-MIS-C (Fisher's exact test, $P = 6.5e\text{-}14$) (Figure 2). The five cases with concomitant viral infections (Table 1) were all of them EPIMISC positive, whereas the epigenomic signature was observed in 2 of 4 (50%) cases classified clinically as MIS-C but without any biological probe of SARS-CoV-2 infection (Table 1).

To further assess the specificity of the EPIMISC signature for the disease, we run our classification model to establish whether it was also overrepresented in available public DNA methylation datasets (GEO data repository) for other distinct pediatric inflammatory disorders. We found that the EPIMISC signature was not present in juvenile localized scleroderma (GEO GSE175379), juvenile systemic sclerosis (GEO GSE175379), or atopic dermatitis (GEO GSE152084). Similarly, the EPIMISC signature was almost non-existent in the general population (0.4%, 1 of 241 donors) (GEO GSE142512; GEO GSE132181). These samples were collected before the emergence of COVID-19, so the donors had never been exposed to the SARS-CoV-2 virus. Our observation that two of the targeted methylated genes within EPIMISC were shared with severe adult COVID-19 cases (AIM2 and PM20D1) prompted us to investigate whether the EPIMISC signature was also present in non-pediatric COVID-19 cases.[12] We found that although EPIMISC was almost completely absent from asymptomatic and

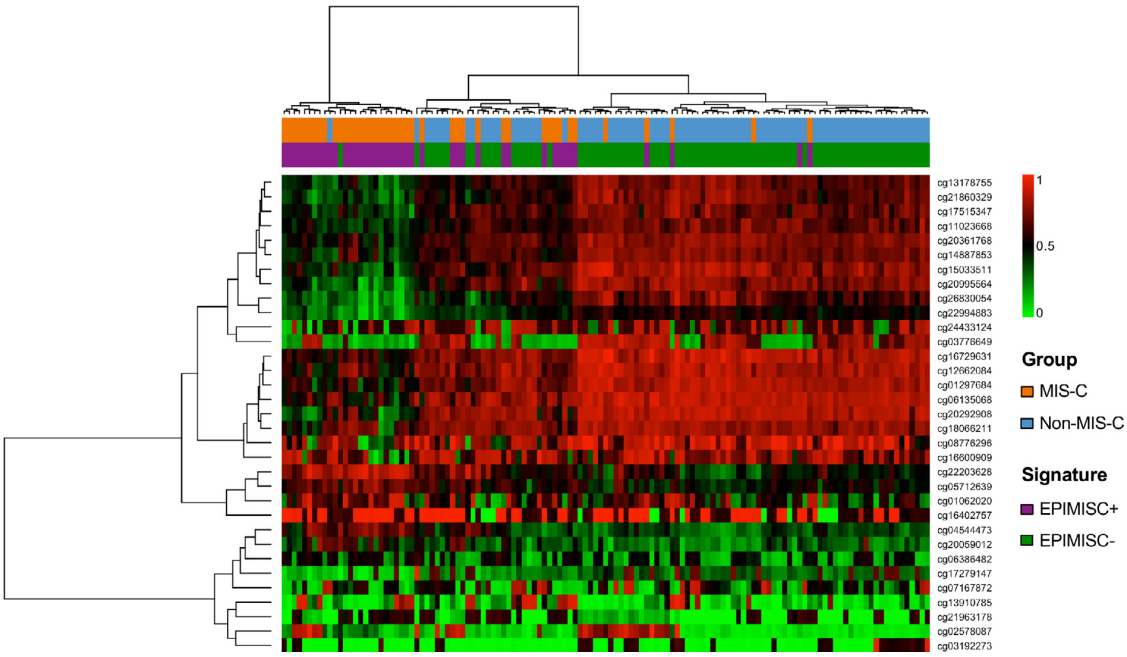


Figure 2. Heatmap representing the entire cohort of MIS-C and non-MIS-C cases clustered by methylation beta values of the 33 CpGs defining the EPIMISC signature. Cluster analysis was performed using the Ward.D clustering method and assuming Manhattan distances.

mild adult COVID-19 patients (1%, 2 of 194), it was present in 24.9% (53 of 213) of adult COVID-19 patients with clinical severity. MIS-C and critically-ill COVID-19 patients show some distinct clinicopathological characteristics, but also some commonalities. This last observation can relate to the targeting of similar cellular networks. For example, the activation of the inflammasome pathobiological pathway represented by the AIM2 gene occurs in COVID-19 adult patients[28] and it was also associated with the severity of the disease in adult cases[12] and, at the same time, the AIM2 gene is a key component of the EPIMISC signature identified herein. But targeting of distinct pathobiological pathways between both disorders also occur. For example the genes in our EPIMISC signature showed a significant enriched Gene Ontology (GO) biological processes (hypergeometric test, adjusted $P < 0.05$) of "regulation of natural killer cell mediated immunity (GO:0002715)" and "peptide antigen assembly with MHC protein complex (GO:0002501)" that were not observed in critically-ill COVID-19 adults.[12] Thus, epigenetic and clinical commonalities and singularities between both disorders occur. Finally, we also wondered if the identified epigenomic profile was overrepresented in diseases involving other viral infections. We observed that the EPIMISC signature was present only in 7.8% (5 out 64) of patients with other viral respiratory infections (GSE167202; Ref.[29]). The interrogated GSE167202 cohort included rhinovirus/enterovirus (33%), influenza A (17%), metapneumovirus (13%), influenza B (11%), other coronavirus (11%), respiratory syncytial virus (9%), parainfluenza (5%), and adenovirus (2%) cases. Most important, the EPIMISC signature was absent in all HIV cases ($n = 70$) analyzed in a recently published cohort (GSE140800; Ref.[30]). Thus, overall, these results further support the specificity of the EPIMISC signature.

MIS-C is considered a new pediatric inflammatory entity associated with SARS-CoV-2 infection, but there is clinical overlap with another disorder, Kawasaki disease,[31–34] a childhood febrile and systemic vasculitis thought to be triggered by exposure to a novel ribonucleic acid, as occurs in viral infections. This is a similar scenario to that presented by SARS-CoV-2 and MIS-C. Remarkably, when we run our classification model to assess the presence of the EPIMISC signature in DNA methylation profiles of Kawasaki disease patients available in the GEO database (GEO GSE84624),[35] we found the EPIMISC signature in 95.8% (23 of 24) of the cases. There are other similarities between the two disorders. For example, beta-catenin contributed to the pathogenesis of Kawasaki disease,[35] and a highly ranked gene of our EPIMISC signature was ZEB2. This gene is involved in the epithelial-mesenchymal transition (EMT), as it also occurs with the Wnt/beta-catenin signaling, but it is also essential for regulating

hematopoiesis.[36] Another example of features common to the two clinical entities was the suggested activation of neutrophils in Kawasaki disease.[37] Using a deconvolution approach to calculate hematological cell populations (Supplementary Methods), we found that our MIS-C cases were also enriched in neutrophils relative to the non-MIS-C cases (Mann–Whitney–Wilcoxon test, $P = 1.7e\text{-}05$). Finally, although the EPIMISC signature was overrepresented in the Kawasaki disease, four CpGs were distinctly methylated between both disorders. One site was not associated with any known gene (cg16729631), and another was located in the GPR111 gene, which was a highly ranked candidate for the MIS-C cases. The other two sites were located in SSUH2 (ssu-2 homolog), a protein chaperone,[38] and RARG (retinoic acid receptor gamma) which is associated with rubella virus-induced cytokine immune responses.[39] These data are germane to similar findings showing that MIS-C and Kawasaki disease share many inflammatory biomarkers, but others are unique such as the high concentration of IFN-gamma-induced CXCL9 in MIS-C cases.[34] The DNA methylation analysis of 33 reported GWAS-derived candidate genes for Kawasaki disease did not show any significant CpG methylation difference in our cohorts (Table S7). Overall, our results highlight the close epigenetic resemblance of MIS-C and Kawasaki disease, further suggesting that a viral infection could unleash the plethora of clinical manifestations that they share.

Discussion

To the best of our knowledge, this is the first study to establish the epigenomic profile of MIS-C patients upon diagnosis. Gene Ontology analyses showed enrichment of the differentially methylated sites in genes associated with an immune response triggered by the SARS-CoV-2 infection. This immune overreaction may well explain the hyperinflammatory phenotype manifested in these children and why multiple body organs and tissues are affected. ZEB2, GPR111, SH2D1B, and HLA-DRB1 are examples of targeted genes, all of which are involved in the generation of immune and inflammatory responses to virus. Interestingly, the EPIMISC signature that we found to be associated with the presence of MIS-C in the discovery and validation cohorts was not linked to other pediatric inflammatory disorders that occur without involvement of a viral agent.

Some of the identified MIS-C epigenetic targets, such as AIM2 and PM20D1, and the EPIMISC signature overall, are also present in some severe adult COVID-19 cases, confirming that both processes (MIS-C in pediatrics and severe acute respiratory distress syndrome in adults) are inflammatory post-infectious complications and probably could be differently treated than the initial phase of the viral infection. Although the gastrointestinal and cardiovascular systems are the most frequently affected in MIS-C, respiratory function is also commonly compromised, with a wide spectrum of consequences, from simple cough and shortness of breath to a requirement for mechanical ventilation.[5,7]

The overlap between the epigenomic landscapes that we found to be associated with MIS-C and Kawasaki disease might have also consequences for understanding the mechanisms involved in the onset of both conditions.[33] Our findings are in line with the reported appearance of Kawasaki's disease-like features in at least 40% of MIS-C patients.[5] MIS-C patients with Kawasaki disease-like features are frequently under 5 years of age,[5] similar to the age of Kawasaki disease patients. In fact, 35% of our patients were younger than 5 years old. The high degree of enrichment of the EPIMISC signature in Kawasaki disease reinforces the invoked role of viral infection in this disorder, as it is also suggested by the peak in cases following the 2009 influenza A H1N1 pandemic.[32] Since MIS-C and Kawasaki disease have similar underlying DNA methylation defects, epigenetic drugs combined with immunomodulatory agents, targeting viral mimicry and inflammation, could be assessed.

Limitations of the study are mainly associated with sample availability, since MIS-C is a rare and novel disease. First, the number of cases is relatively low, although it is in line with previous studies defining molecular profiles in MIS-C.[27,40–47] It should be highlighted, to the best of our knowledge, this is the first epigenomic profiling of MIS-C cases. A second limitation is the lack of ethnic heterogeneity, directly related to the ethnic distribution in the studied population, enriched in West-Eurasia origin. This fact could underestimate key intrinsic features of other populations showing potential enhanced risk of MIS-C in previous studies, such as black children.[7,48] A third limitation to consider is the possible existence of additional unmeasured confounding factor, other than age or those that were not statistically significant in our analysis.

In conclusion, we report that MIS-C patients exhibit a well-defined set of epigenetic loci that are associated with the diagnosis of the disorder and support a direct role of a hyperactivated immune response in the characteristic features of overinflammation and multisystem organ involvement. These DNA methylation sites were used to construct an epigenomic signature, EPIMISC, that is associated with the disease. This profile was absent in non-viral inflammatory processes in children, but present to a certain degree in severe adult COVID-19 cases. The profile overlaps with that of another inflammatory syndrome, Kawasaki disease, where a trigger by viral infection can now also be further strengthened. These findings provide essential clues that will help us to understand the immune mechanisms that go awry in MIS-C cases, to identify patients likely to have worse outcomes, and to suggest actionable

candidate genes for more specific treatments. Together with genetic, serological and clinical parameters,[49] the EPIMISC signature could help in patient stratification and to identify highly susceptible patients who require close attention and early active treatments to prevent the progression of the disease. Most importantly, we have identified new biomarkers for diagnosing MIS-C patients that could be useful as the COVID-19 pandemic progresses and seroconversion increases, reducing the value of knowing the history of exposure and serology for defining the MIS-C clinical entity, a key point once COVID-19 turns into an endemic disease worldwide. Finally, the identified epigenetic sites could be useful for following up these patients, including how we monitor the efficacy of immunomodulation therapies and how we can detect at an early stage the MIS-C cases whose disease will rapidly worsen.

Contributors

VD, CAGP, AP, and ME designed the study, contributed to the analysis, and wrote the first draft of the manuscript. In-depth clinical and pathological characterization and recruitment of patients were carried out by GF, SAA, JVR, ARP, MR, LPS, IJ, IA, PFP, VC, VF, MTV, CR, MMH, ELG, RC, JGR, PSP and AP. All authors helped draft the manuscript or revised it critically for significant intellectual content, and made substantial contributions to the concept and design of the study, and to the acquisition, analysis and interpretation of data.

Data sharing statement

The complete DNA methylation raw data of the all the studied MIS-C and non-MIS-C samples cases are available on the GEO repository under accession number GSE193879.

Declaration of interests

Dr. Esteller declares grants from Ferrer International, personal fees from Quimatryx, outside the submitted work. Dr. Rivière reports personal fees from Grifols, CSL behring and Takeda, outside the submitted work. The other authors declare no conflicts of interest.

Acknowledgements

We thank the Health Department and the Centres de Recerca de Catalunya (CERCA) Programme of the Generalitat de Catalunya, the Josep Carreras Leukaemia Foundation, Fundació La Marató de TV3 and the Cellex Foundation for institutional support. We also wish to thank all the patients, family members and staff from all the units that participated in the study.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2022.101515.

References

1 Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727–733.
2 Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395:1054–1062.
3 Götzinger F, Santiago-García B, Noguera-Julián A, et al. COVID-19 in children and adolescents in Europe: a multinational, multicentre cohort study. *Lancet Child Adolesc Health*. 2020;4:653–661.
4 Whittaker E, Bamford A, Kenny J, et al. Clinical characteristics of 58 children with a pediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2. *JAMA*. 2020;324:259–269.
5 Feldstein LR, Rose EB, Horwitz SM, et al. Multisystem inflammatory syndrome in U.S. children and adolescents. *N Engl J Med*. 2020;383:334–346.
6 Jiang L, Tang K, Levin M, et al. COVID-19 and multisystem inflammatory syndrome in children and adolescents. *Lancet Infect Dis*. 2020;20:e276–e288.
7 Abrams JY, Oster ME, Godfred-Cato SE, et al. Factors linked to severe outcomes in multisystem inflammatory syndrome in children (MIS-C) in the USA: a retrospective surveillance study. *Lancet Child Adolesc Health*. 2021;5:323–331.
8 Penner J, Abdel-Mannan O, Grant K, et al. 6-month multidisciplinary follow-up and outcomes of patients with paediatric inflammatory multisystem syndrome (PIMS-TS) at a UK tertiary paediatric hospital: a retrospective cohort study. *Lancet Child Adolesc Health*. 2021;5:473–482.
9 Harwood R, Allin B, Jones CE, et al. A national consensus management pathway for paediatric inflammatory multisystem syndrome temporally associated with COVID-19 (PIMS-TS): results of a national Delphi process. *Lancet Child Adolesc Health*. 2021;5:133–141.
10 Ahmed M, Advani S, Moreira A, et al. Multisystem inflammatory syndrome in children: a systematic review. *EClinicalMedicine*. 2020;26:100527.
11 Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. *Nat Rev Genet*. 2019;20:109–127.
12 Castro de Moura M, Davalos V, Planas-Serra L, et al. Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine*. 2021;66:103339.
13 Villanueva L, Álvarez-Errico D, Esteller M. The contribution of epigenetics to cancer immunotherapy. *Trends Immunol*. 2020;41:676–691.
14 Scott CL, Omilusik KD. ZEBs: novel players in immune cell development and function. *Trends Immunol*. 2019;40:431–446.
15 Hamann J, Hsiao CC, Lee CS, Ravichandran KS, Lin HH. Adhesion GPCRs as modulators of immune cell function. *Handb Exp Pharmacol*. 2016;234:329–350.
16 O'Connell P, Amalfitano A, Aldhamen YA. SLAM family receptor signaling in viral infections: HIV and beyond. *Vaccines (Basel)*. 2019;7:184.
17 Aldhamen YA, Seregin SS, Schuldt NJ, et al. Vaccines expressing the innate immune modulator EAT-2 elicit potent effector memory T lymphocyte responses despite pre-existing vaccine immunity. *J Immunol*. 2012;189:1349–1359.
18 Curtis VF, Ehrentraut SF, Campbell EL, et al. Stabilization of HIF through inhibition of Cullin-2 neddylation is protective in mucosal inflammatory responses. *FASEB J*. 2015;29:208–215.
19 Kumari P, Russo AJ, Shivcharan S, Rathinam VA. AIM2 in health and disease: Inflammasome and beyond. *Immunol Rev*. 2020;297:83–95.
20 Li X, Zhao X, Xing J, Li J, et al. Different epigenome regulation and transcriptome expression of CD4+ and CD8+ T cells from monozygotic twins discordant for psoriasis. *Australas J Dermatol*. 2020;61:e388–e394.
21 Imran S, Neeland MR, Koplin J, et al. Epigenetic programming underpins B-cell dysfunction in peanut and multi-food allergy. *Clin Transl Immunol*. 2021;10:e1324.

22 Amoroso A, Magistroni P, Vespasiano F, et al. HLA and AB0 polymorphisms may influence SARS-CoV-2 infection and COVID-19 severity. *Transplantation*. 2021;105:193–200.
23 Anzurez A, Naka I, Miki S, et al. Association of HLA-DRB1*09:01 with severe COVID-19. *HLA*. 2021;98:37–42.
24 Castelli EC, de Castro MV, Naslavsky MS, et al. MHC Variants associated with symptomatic versus asymptomatic SARS-CoV-2 infection in highly exposed individuals. *Front Immunol*. 2021;12:742881.
25 Li S, Wu B, Ling Y, et al. Epigenetic landscapes of single-cell chromatin accessibility and transcriptomic immune profiles of T Cells in COVID-19 patients. *Front Immunol*. 2021;12:625881.
26 Mousa M, Vurivi H, Kannout H, et al. Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates. *EBioMedicine*. 2021;74:103695.
27 Chou J, Platt CD, Habiballah S, et al. Mechanisms underlying genetic susceptibility to multisystem inflammatory syndrome in children (MIS-C). *J Allergy Clin Immunol*. 2021;148:732–738. e1.
28 Junqueira C, Crespo Â, Ranjbar S, et al. FcγR-mediated SARS-CoV-2 infection of monocytes activates inflammation. *Nature*. 2022;606(7914):576–584.
29 Konigsberg IR, Barnes B, Campbell M, et al. Host methylation predicts SARS-CoV-2 infection and clinical outcome. *Commun Med (Lond)*. 2021;1(1):42.
30 Oriol-Tordera B, Berdasco M, Llano A, et al. Methylation regulation of antiviral host factors, interferon stimulated genes (ISGs) and T-cell responses associated with natural HIV control. *PLoS Pathog*. 2020;16(8):e1008678.
31 Verdoni L, Mazza A, Gervasoni A, et al. An outbreak of severe Kawasaki-like disease at the Italian epicentre of the SARS-CoV-2 epidemic: an observational cohort study. *Lancet*. 2020;395:1771–1778.
32 Ouldali N, Pouletty M, Mariani P, et al. Emergence of Kawasaki disease related to SARS-CoV-2 infection in an epicentre of the French COVID-19 epidemic: a time-series analysis. *Lancet Child Adolesc Health*. 2020;4:662–668.
33 Sancho-Shimizu V, Brodin P, Cobat A, et al. SARS-CoV-2-related MIS-C: A key to the viral and genetic causes of Kawasaki disease? *J Exp Med*. 2021;218:e20210446.
34 Rodriguez-Smith JJ, Verweyen EL, Clay GM, et al. Inflammatory biomarkers in COVID-19 associated multisystem inflammatory syndrome in children, Kawasaki disease, and macrophage activation syndrome: a cohort study. *Lancet Rheumatol*. 2021;3:e574–e584.
35 Chen KD, Huang YH, Ming-Huey Guo M, et al. The human blood DNA methylome identifies crucial role of β-catenin in the pathogenesis of Kawasaki disease. *Oncotarget*. 2018;9:28337–28350.
36 Wang J, Farkas C, Benyoucef A, et al. Interplay between the EMT transcription factors ZEB1 and ZEB2 regulates hematopoietic stem and progenitor cell differentiation and hematopoietic lineage fidelity. *PLoS Biol*. 2021;19:e3001394.
37 Huang LH, Kuo HC, Pan CT, Lin YS, Huang YH, Li SC. Multiomics analyses identified epigenetic modulation of the S100A gene family in Kawasaki disease and their significant involvement in neutrophil transendothelial migration. *Clin Epigenet*. 2018;10:135.
38 Reinartz A, Ehling J, Franz S, et al. Small intestinal mucosa expression of putative chaperone fls485. *BMC Gastroenterol*. 2010;10:27.
39 Ovsyannikova IG, Dhiman N, Haralambieva IH, et al. Rubella vaccine-induced cellular immunity: evidence of associations with polymorphisms in the Toll-like, vitamin A and D receptors, and innate immune response genes. *Hum Genet*. 2010;127:207–221.
40 Ramaswamy A, Brodsky NN, Sumida TS, et al. Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children. *Immunity*. 2021;54(5):1083–1095.
41 Porritt RA, Binek A, Paschold L, et al. The autoimmune signature of hyperinflammatory multisystem inflammatory syndrome in children. *J Clin Invest*. 2021;131(20):e151520.
42 Beckmann ND, Comella PH, Cheng E, et al. Downregulation of exhausted cytotoxic T cells in gene expression networks of multisystem inflammatory syndrome in children. *Nat Commun*. 2021;12(1):4854.
43 Pfeifer J, Thurner B, Kessel C, et al. Autoantibodies against interleukin-1 receptor antagonist in multisystem inflammatory syndrome in children: a multicentre, retrospective, cohort study. *Lancet Rheumatol*. 2022;4(5):e329–e337.
44 Gruber CN, Patel RS, Trachtman R, et al. Mapping systemic inflammation and antibody responses in multisystem inflammatory syndrome in children (MIS-C). *Cell*. 2020;183(4):982–995.e14.
45 Consiglio CR, Cotugno N, Sardh F, et al. The immunology of multisystem inflammatory syndrome in children with COVID-19. *Cell*. 2020;183(4):968–981.e7.
46 Vella LA, Giles JR, Baxter AE, et al. Deep immune profiling of MIS-C demonstrates marked but transient immune activation compared to adult and pediatric COVID-19. *Sci Immunol*. 2021;6(57):eabf7570.
47 Esteve-Sole A, Anton J, Pino-Ramirez RM, et al. Similarities and differences between the immunopathogenesis of COVID-19-related pediatric multisystem inflammatory syndrome and Kawasaki disease. *J Clin Invest*. 2021;131(6):e144554.
48 Middelburg JG, Crijnen TEM, D'Antiga L, et al. Association of ethnicity with multisystem inflammatory syndrome in children related to SARS-CoV-2 infection: an international case-referent study. *Front Pediatr*. 2021;9:707650.
49 Geva A, Patel MM, Newhams MM, et al. Data-driven clustering identifies features distinguishing multisystem inflammatory syndrome from acute COVID-19 in children and adolescents. *EClinical Medicine*. 2021;40:101112.

Supplementary materials

Supplementary Materials

Supplementary Figure S1. Graphical schema representing the populations of interest and the screening strategy used to identify epigenetic biomarkers of MIS-C.

Supplementary Figure S2. Characteristic laboratory parameters of the MIS-C patients included in the study.

Supplementary Figure S3. EPIMISC performance in the discovery cohort.

Supplementary Figure S4. Heatmap of the discovery cohort samples, clustered by methylation beta values of the 33 CpGs defining the EPIMISC signature.

Supplementary Figure S5. EPIMISC performance in the validation cohort.

Supplementary Figure S6. Heatmap of the validation cohort samples, clustered by methylation beta values of the 33 CpGs defining the EPIMISC signature.

Supplementary Figure S7. EPIMISC performance in the entire cohort.

Supplementary Table S1. Characteristics of the pediatric COVID-19 patients with IgG and/or PCR positive status for SARS-CoV-2, but without MIS-C.

Supplementary Table S2. Characteristics of the children and adolescent control donors collected during the pre-COVID-19 period.

Supplementary Table S3. Description of the 33 CpG sites with a differential DNA methylation status between the MIS-C and non-MIS-C in the discovery cohort, according to the study pipeline described in Supplementary Figure S1.

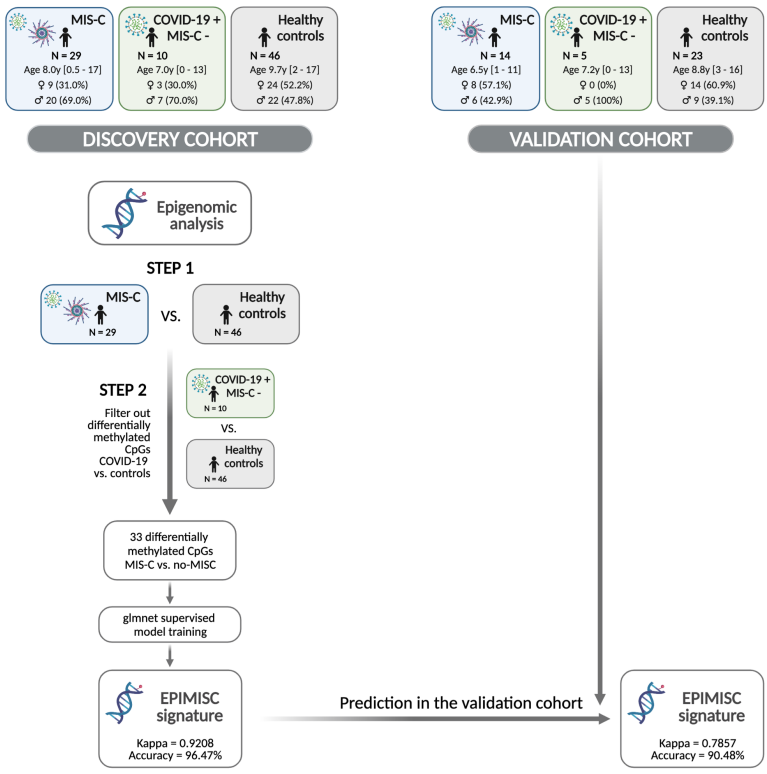
Supplementary Table S4. Description of the 1337 CpG sites in the 47 genes related to COVID-19, based on published studies.

Supplementary Table S5. Description of the 68 CpG sites in the 3 genes related to MIS-C, based on published studies.

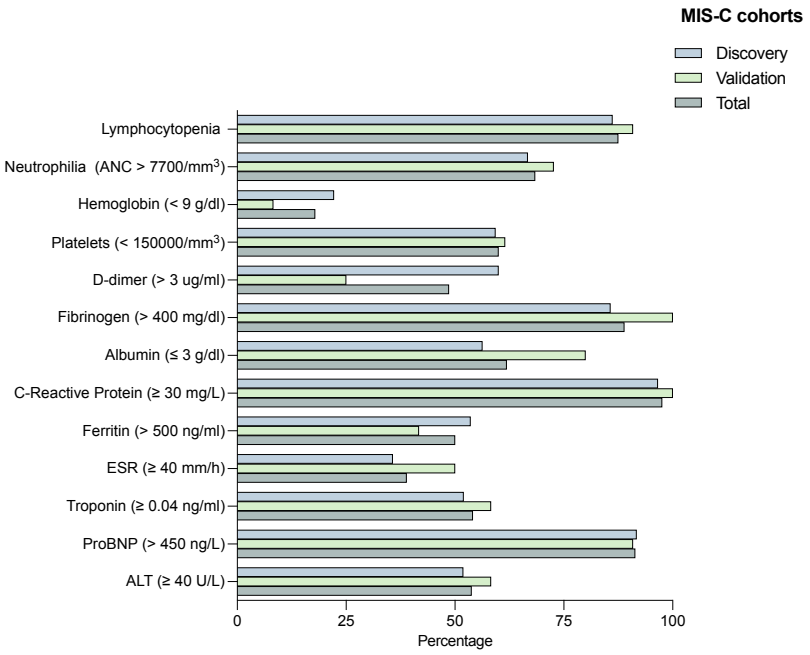
Supplementary Table S6. EPIMISC derived-CpGs in the validation and entire cohorts, according to the study pipeline described in Supplementary Figure S1.

Supplementary Table S7. Description of the 1350 CpG sites in the 33 genes related to Kawasaki disease, based on published studies.

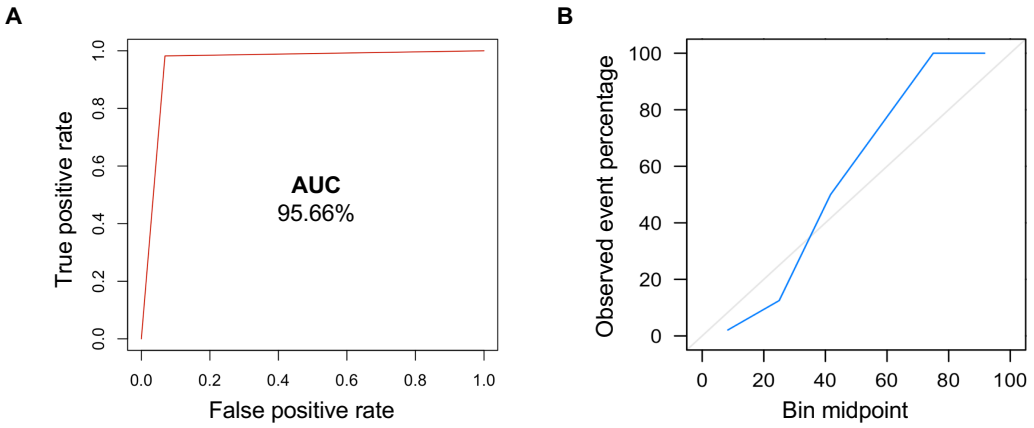
Supplementary Methods. Study protocol and extended methods.



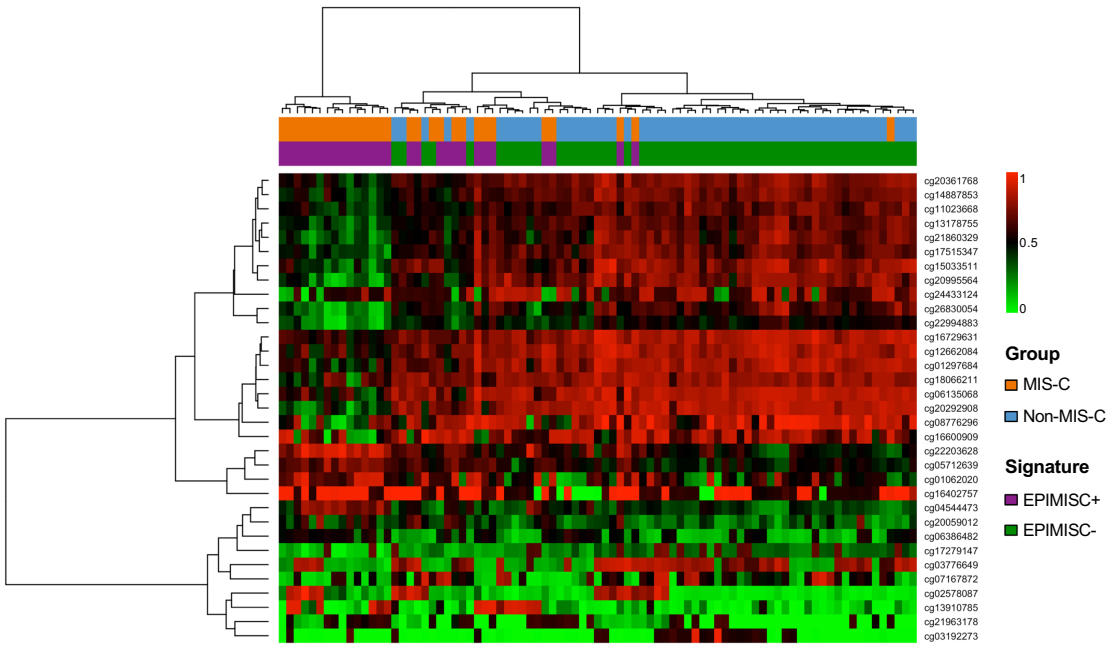
Supplementary Figure S1. Graphical schema representing the populations of interest and the screening strategy used to identify epigenetic biomarkers of MIS-C.



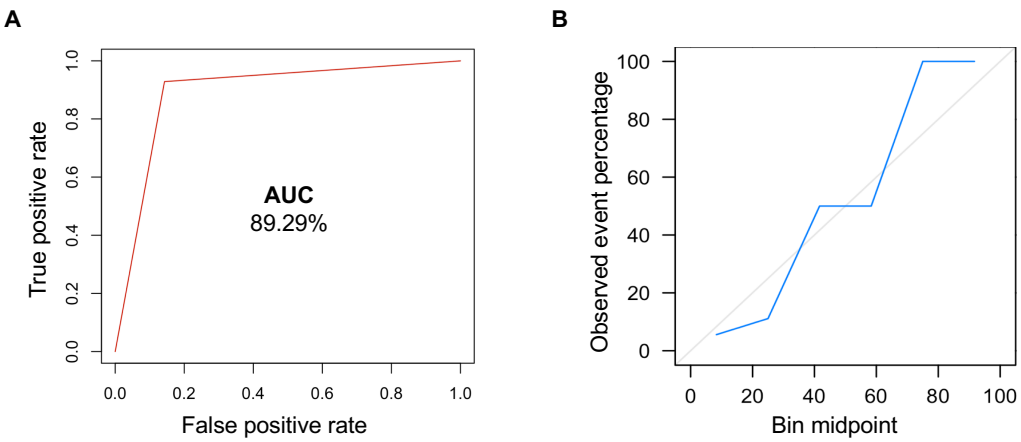
Supplementary Figure S2. Characteristic laboratory parameters of the MIS-C patients included in the study. Following the criteria used for MIS-C patients in Feldstein et al., *N Engl J Med*, 2020. Lymphocytopenia was defined as an absolute lymphocyte count (ALC) of less than 1500 per microliter in patients 8 months of age or older, and less than 4500 per microliter in patients younger than 8 months of age. Abbreviations: ALT, Alanine aminotransferase; ANC, Absolute neutrophil count; ESR, Erythrocyte sedimentation rate; ProBNP, pro-brain natriuretic peptide.



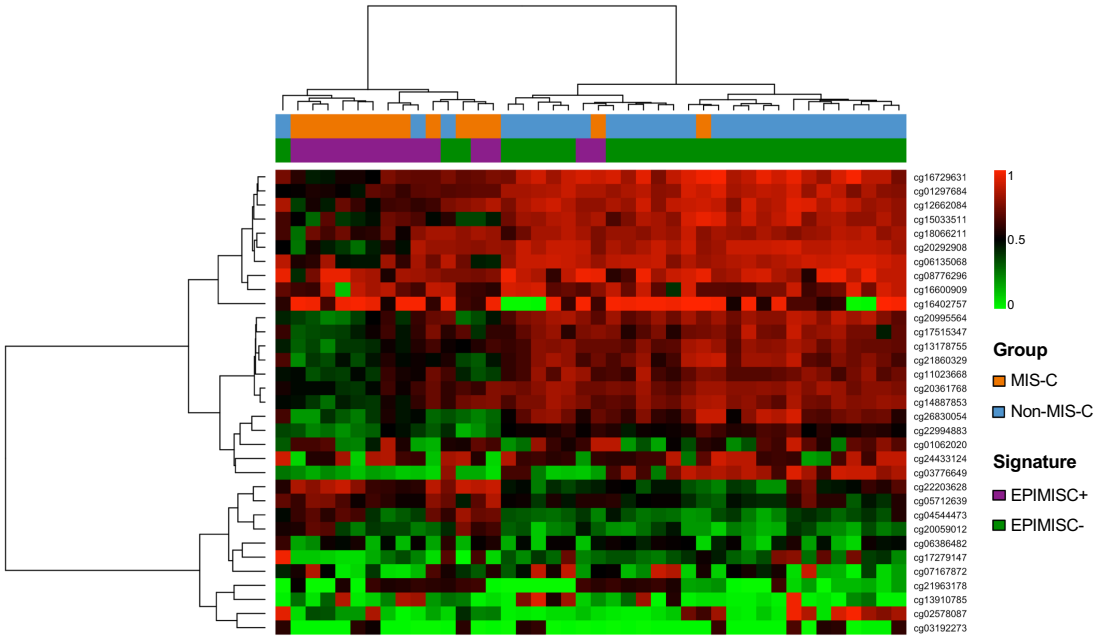
Supplementary Figure S3. EPIMISC performance in the discovery cohort. **A.** Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC = 95.66% (95% confidence interval = 90.65% to 100%)) for the EPIMISC signature. **B**. Calibration curve using 6 bins to characterize the consistency between EPIMISC predicted class probabilities and observed event rates.



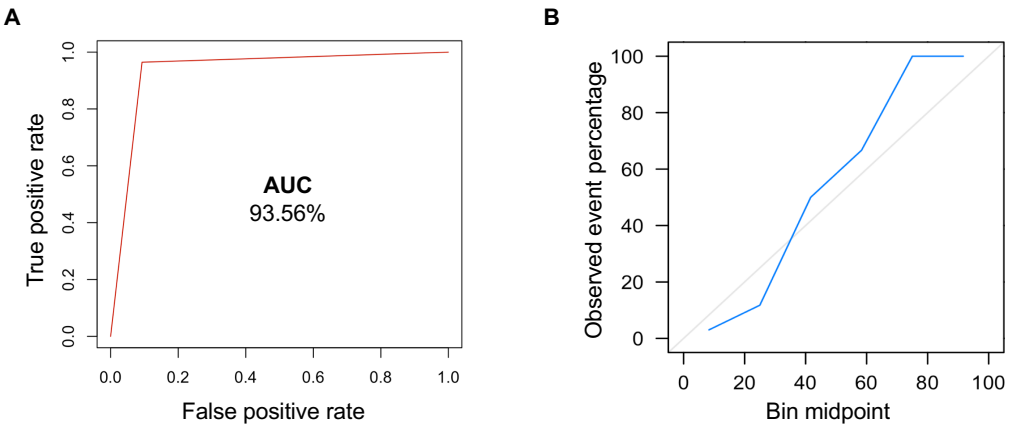
Supplementary Figure S4. Heatmap of the discovery cohort samples, clustered by methylation beta values of the 33 CpGs defining the EPIMISC signature. Cluster analysis was performed using the Ward.D clustering method and assuming Manhattan distances.



Supplementary Figure S5. EPIMISC performance in the validation cohort. **A.** Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC = 89.29% (95% confidence interval = 78.61% to 99.97%)) for the EPIMISC signature. **B**. Calibration curve using 6 bins to characterize the consistency between EPIMISC predicted class probabilities and observed event rates.



Supplementary Figure S6. Heatmap of the validation cohort samples, clustered by methylation beta values of the 33 CpGs defining the EPIMISC signature. Cluster analysis was performed using the Ward.D clustering method and assuming Manhattan distances.



Supplementary Figure S7. EPIMISC performance in the entire cohort. **A.** Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC = 93.56% (95% confidence interval = 88.74% to 98.39%)) for the EPIMISC signature. **B**. Calibration curve using 6 bins to characterize the consistency between EPIMISC predicted class probabilities and observed event rates.

Supplementary Table S1. Characteristics of the pediatric COVID-19 patients with IgG and/or PCR positive status for SARS-CoV-2, but without MIS-C.

	Pediatric COVID-19 cohorts		
Characteristics	**Discovery cohort** (N = 10)	**Validation cohort** (N = 5)	**Entire cohort** (N = 15)
Gender - Frequency (%)			
Female	3 (30.0%)	0 (0%)	3 (20.0%)
Male	7 (70.0%)	5 (100%)	12 (80.0%)
Age (years) - Median [range]	7.0 [0 - 13]	7.2 [0 - 13]	7.1 [0 - 13]
Age group- Frequency (%)			
≤ 2 yr	2 (20.0%)	2 (40.0%)	4 (26.7%)
3 – 5 yr	1 (10.0%)	0 (0%)	1 (6.7%)
6 – 9 yr	5 (50.0%)	1 (20.0%)	6 (40.0%)
10 – 13 yr	2 (20.0%)	2 (40.0%)	4 (26.7%)
Ethnicity - Frequency (%)			
West-Eurasia	10 (100%)	3 (60.0%)	13 (86.7%)
Central-South America	0 (0%)	1 (20.0%)	1 (6.7%)
African	0 (0%)	1 (20.0%)	1 (6.7%)
SARS-CoV-2 status- Frequency (%)			
IgG and/or PCR positive	10 (100%)	5 (100%)	15 (100%)
Highest level of care - Frequency (%)			
Home	2 (20.0%)	1 (20.0%)	3 (20.0%)
Ward	4 (40.0%)	2 (40.0%)	6 (40.0%)
Intensive care unit	4 (40.0%)	2 (40.0%)	6 (40.0%)
Oxygen supplementation - Frequency (%)			
None	3 (30.0%)	3 (60.0%)	6 (40.0%)
Nasal cannula	4 (40.0%)	0 (0%)	4 (26.7%)
Non-Invasive Ventilation or High Flow Oxygen	3 (30.0%)	1 (20.0%)	4 (26.7%)
Mechanical Ventilation	0 (0%)	1 (20.0%)	1 (6.7%)

Supplementary Table S2. Characteristics of the children and adolescent control donors collected during the pre-COVID-19 period.

Characteristics		Pediatric control donor cohorts		
		Discovery cohort (N = 46)	**Validation cohort** (N = 23)	**Entire cohort** (N = 69)
Gender - Frequency (%)				
	Female	24 (52.2%)	14 (60.9%)	38 (55.1%)
	Male	22 (47.8%)	9 (39.1%)	31 (44.9%)
Age (years) - Median [range]		9.7 [2 - 17]	8.8 [3 - 16]	9.4 [2 - 17]
Age group- Frequency (%)				
	≤ 2 yr	4 (8.7%)	0 (0%)	4 (5.8%)
	3 – 5 yr	6 (13.0%)	8 (34.8%)	14 (20.3%)
	6 – 9 yr	10 (21.7%)	6 (26.1%)	16 (23.2%)
	10 – 13 yr	13 (28.3%)	2 (8.7%)	15 (21.7%)
	14 – 18 yr	13 (28.3%)	7 (30.4%)	20 (29.0%)
Ethnicity - Frequency (%)				
	West-Eurasia	31 (67.4%)	11 (47.8%)	42 (60.9%)
	Central-South America	1 (2.2%)	1 (4.3%)	2 (2.9%)
	African	1 (2.2%)	0 (20.0%)	1 (1.4%)
	Unknown	13 (28.2%)	11 (47.8%)	24 (34.7%)
Sampling date - Frequency (%)				
	Pre-COVID-19 pandemic (before December 2019)	46 (100%)	23 (100%)	69 (100%)

Supplementary Table S3. Description of the 33 CpG sites with a differential DNA methylation status between MIS-C and non-MIS-C in the discovery cohort, according to the study pipeline described in Supplementary Figure S1.

CpG ID	Chromosome location	Gene name	P value	Absolute mean methylation difference
cg20292908	chr5:172203421		< 0.001	0.30
cg20995564	chr2:145172035	ZEB2	< 0.001	0.27
cg06135068	chr16:31034016		< 0.001	0.22
cg16402757	chr10:35311004	CUL2	0.010	0.21
cg26830054	chr10:119762394		< 0.001	0.20
cg15033511	chr17:54994023		< 0.001	0.20
cg01062020	chr1:162382848	SH2D1B	0.001	0.19
cg03776649	chr5:125465533		0.023	0.19
cg22203628	chr11:69241075		< 0.001	0.19
cg02578087	chr3:8671361	SSUH2	0.009	0.18
cg01297684	chr12:56069634		< 0.001	0.18
cg20361768	chr3:156819083	LINC00880	< 0.001	0.18
cg06386482	chr6:47624117	GPR111	< 0.001	0.18
cg21963178	chr10:75571738	NDST2	0.001	0.18
cg24433124	chr6:30755968		0.004	0.17
cg13910785	chr6:32549849	HLA-DRB1	0.042	0.17
cg05712639	chr14:52819386		< 0.001	0.17
cg03192273	chr5:150618948		0.028	0.16
cg21860329	chr13:42265546	VWA8	< 0.001	0.16
cg14887853	chr6:139794538	LOC645434	< 0.001	0.16
cg12662084	chr6:17809126	KIF13A	< 0.001	0.16
cg16600909	chr1:173145001		0.012	0.15
cg13178755	chr2:174023580	ZAK	< 0.001	0.15
cg20059012	chr12:53613154	RARG	< 0.001	0.15
cg07167872	chr1:205819463	PM20D1	0.033	0.15
cg16729631	chr8:131000261		< 0.001	0.15
cg17515347	chr1:159047163	AIM2	< 0.001	0.15
cg11023668	chr2:25095040	ADCY3	< 0.001	0.15
cg22994883	chr7:130615334		< 0.001	0.15
cg04544473	chr4:153021978	LOC100996286	< 0.001	0.15
cg18066211	chr4:99369082		< 0.001	0.15
cg17279147	chr6:14739369		0.001	0.15
cg08776296	chr7:134856544	CYREN	0.014	0.15

Notes: CpG ID corresponds to the unique CpG site identifier in the HumanMethylationEPIC array (Illumina). Chromosomal location denoted according human reference assembly GRCh37/hg19. All depicted P values are FDR adjusted P values. CpG-sites with an absolute mean methylation beta value difference > 0.15 and FDR adjusted P value < 0.05 were considered statistically significant.

Supplementary Table S6. EPIMISC derived-CpGs in the validation and entire cohorts, according to the study pipeline described in Supplementary Figure S1.

			Validation cohort		Entire cohort	
CpG ID	**Chromosome location**	**Gene name**	**P value**	**Absolute mean methylation difference**	**P value**	**Absolute mean methylation difference**
cg03776649	chr5:125465533		0.005	0.369	< 0.001	0.248
cg20995564	chr2:145172035	ZEB2	0.001	0.284	< 0.001	0.277
cg22203628	chr11:69241075		0.003	0.265	< 0.001	0.213
cg20292908	chr5:172203421		0.001	0.262	< 0.001	0.285
cg26830054	chr10:119762394		0.003	0.244	< 0.001	0.215
cg15033511	chr17:54994023		0.003	0.237	< 0.001	0.210
cg06135068	chr16:31034016		< 0.001	0.233	< 0.001	0.223
cg21860329	chr13:42265546	VWA8	0.003	0.218	< 0.001	0.182
cg13178755	chr2:174023580	ZAK	0.004	0.192	< 0.001	0.166
cg04544473	chr4:153021978	LOC100996286	0.004	0.189	< 0.001	0.164
cg16729631	chr8:131000261		0.004	0.185	< 0.001	0.163
cg20059012	chr12:53613154	RARG	0.009	0.181	< 0.001	0.162
cg11023668	chr2:25095040	ADCY3	0.004	0.178	< 0.001	0.160
cg17279147	chr6:14739369		0.016	0.176	< 0.001	0.159
cg22994883	chr7:130615334		0.003	0.171	< 0.001	0.157
cg01297684	chr12:56069634		0.003	0.165	< 0.001	0.177
cg14887853	chr6:139794538	LOC645434	0.006	0.164	< 0.001	0.162
cg12662084	chr6:17809126	KIF13A	0.003	0.162	< 0.001	0.159
cg20361768	chr3:156819083	LINC00880	0.003	0.152	< 0.001	0.171
cg17515347	chr1:159047163	AIM2	0.008	0.151	< 0.001	0.151
cg21963178	chr10:75571738	NDST2	0.419	0.114	0.001	0.156
cg18066211	chr4:99369082		0.016	0.114	< 0.001	0.139
cg05712639	chr14:52819386		0.027	0.114	< 0.001	0.150
cg24433124	chr6:30755968		0.393	0.113	0.001	0.152
cg16402757	chr10:35311004	CUL2	0.677	0.087	0.017	0.173
cg16600909	chr1:173145001		0.466	0.074	0.003	0.129
cg01062020	chr1:162382848	SH2D1B	0.295	0.068	< 0.001	0.152
cg08776296	chr7:134856544	CYREN	0.364	0.055	0.003	0.120
cg06386482	chr6:47624117	GPR111	0.587	0.049	0.007	0.104
cg02578087	chr3:8671361	SSUH2	0.373	0.048	0.004	0.139
cg13910785	chr6:32549849	HLA-DRB1	0.987	0.032	0.078	0.104
cg07167872	chr1:205819463	PM20D1	0.930	0.011	0.053	0.105
cg03192273	chr5:150618948		0.995	0.004	0.054	0.111

Notes: CpG ID corresponds to the unique CpG site identifier in the HumanMethylationEPIC array (Illumina). Chromosomal location denoted according human reference assembly GRCh37/hg19. All depicted P values are FDR adjusted P values. CpG-sites with an absolute mean methylation beta value difference > 0.15 and FDR adjusted P value < 0.05 were considered statistically significant.

Supplementary Tables S4, S5 and S7 can be accessed on the publication site at: 10.1016/j.eclinm.2022.101515

Supplementary Methods

Study protocol

Whole blood samples and clinical data from 43 MIS-C patients were retrospectively collected between April 16th, 2020 and August 17th, 2021 in seven Hospitals in Spain organized into collaborative groups. Whole blood samples were also obtained from 15 pediatric COVID-19 cases with no evidence of MIS-C, and from 69 healthy children and adolescents collected during the pre-COVID-19 era (before December 2019). Approvals were obtained by the corresponding Ethical Committees (PR052/21, Hospital Universitari de Bellvitge; PR(AMI)388/2016, Hospital Universitari Vall d'Hebron; PI_2020/35, Complejo Hospitalario de Navarra; CEIm 2314, Hospital Universitario de Burgos; PI-21-160, Hospital Universitari Germans Trias i Pujol). According to the Biomedical Research Law 14/2007, informed consents were signed to donate biological material for research purposes at the reference center. Clinical information has been collected, processed and stored under confidentiality policies, in accordance with the National Organic Law 3/2018, on the protection of personal data and guarantee of digital rights. Clinical data and biological samples arrived at our institution pseudonymized (de-identified) by the clinician or personnel authorized at the healthcare institution. Sensitive patient information showing the identity of the patient was only recorded at the healthcare institution. Biological samples were systematically collected and appropriately preserved for research studies. To this end, peripheral blood samples were drawn in EDTA Vacutainer® blood collection tubes and stored at - 80ºC until DNA extraction.

DNA methylation data

The DNA methylation status of the studied samples was obtained using the Infinium MethylationEPIC Array (~850,000 CpG sites), following the manufacturer's instructions for the automated processing of arrays with a liquid handler (Illumina Infinium HD Methylation Assay Experienced User Card, Automated Protocol 15019521 v01) (Moran et al., 2016). DNA methylation beta values were obtained from the raw IDAT files using the minfi package (v1.36.0) in R. Briefly, the pre-processing of the methylation data performed with the minfi package in R involved removal of erratic probe signals such as failed probes (probes with a detection value of $P > 0.01$), cross-reacting probes and probes that overlapped SNPs within $\pm$1 base pair of CpG sites. Those probes that failed in more than 10% of samples were removed from the analysis, whereas the beta value of the probes that failed in less than 10% of samples was imputed using the median. XY chromosome probes were also removed. Finally, background correction and dye-based normalization were performed using ssNoob algorithm. The genomic analysis presented in the study was performed using the GRCh37 – hg19 human genome reference build, as described in the Illumina manifest file associated with the DNA methylation EPIC microarray.

Computational analyses

The MIS-C epigenetic signature, referred to hereafter as EPIMISC, was obtained by first identifying the probes differentially methylated between MIS-C cases and healthy control donors, filtering out in a second step those probes found to be differentially methylated between pediatric COVID-19 cases and healthy controls (**Figure S1**). This approach enabled us to effectively discover the differentially methylated probes between MIS-C and non-MIS-C cases. This involved deriving a linear model adjusted by the age covariate with the limma R package (v3.46.0) using the methylation values of the discovery dataset. A significance threshold for CpGs with a False Discovery Rate (FDR) adjusted P value < 0.05 and an absolute mean methylation beta difference between groups of > 0.15 was established. The significantly differential DNA methylation sites (**Table S3**) were used to train a supervised classification model based on a ridge-regularized logistic regression to predict MIS-C diagnosis using the glmnet R package (v4.1-1). Thus, the methylation beta values of the differential DNA methylation sites (**Table S3**) are the predictors of the ridge-regularized logistic regression and the outcome is the presence or absence of the EPIMISC signature. The classification model was optimized by tuning parameters (best performance with alpha = 0 from ridge regression, and regularization parameter lambda = 0.1) after resampling with 10-fold cross-validation carried out three times using the caret package in R (v6.0-86). To this end, we used the createFolds caret function to generate balanced cross-validation groupings from the discovery dataset in order to perform the 10-fold cross-validation. Once the model and tuning parameters values have been defined after resampling, our model performance was assessed using the receiver operating characteristic (ROC) and calibration curves. In the next step, the classification model was tested in the validation cohort and a corresponding confusion matrix derived.

Likewise, a supervised classification model was trained and optimized with the same parameters as previously described, but this time using the differentially methylated CpGs present in the Infinium Human Methylation 450K BeadChip to test it in external datasets from the Gene Expression Omnibus (GEO) for which only 450K data were available. The performance of the 450K model was also assessed using the ROC curve of the resamples. Finally, hierarchical clustering analysis was performed using the Ward.D clustering method with Manhattan distances in the gplots (v3.1.1) package in R. All analyses were performed within the R statistical environment (v4.0.3).

Gene Set Enrichment analysis

Gene set enrichment analysis was conducted using the Enrichr tool (Kuleshov et al., 2016) by performing hypergeometrical tests (one-tailed Fisher´s exact test) using Gene Ontology

Biological Process gene set. Enrichments were considered significant at Benjamini & Hochberg False Discovery Rate (FDR) adjusted $P < 0.05$.

Cell type deconvolution analysis

Cell type deconvolution analysis to calculate particular hematological cell populations based on blood-derived DNA methylation signatures (Salas et al., 2018) was performed using estimateCellCounts2 function from FlowSorted.Blood.EPIC (v1.8.0) package in R.

Statistical analyses

The entire cohort (N=127) was divided into a discovery (N=85) and validation cohorts (N=42) using the createDataPartition function from the caret R package (v6.0-86). We used this function to create one hundred (66% discovery and 33% validation) disease status and age-balanced splits of the entire dataset, so that the random sampling occurred within each disease status (MIS-C vs non-MIS-C) and age category (age groups were defined by 5 years intervals), preserving the overall class distribution of the data. We randomly selected one of the 100 balanced splits to divide samples into discovery and validation cohorts. This was done to avoid disparity in the frequencies of the observed classes (disease status and age) that can have a significant negative impact on model fitting.

We estimated the power of our EPIC array DNA methylation study to demonstrate the hypotheses according to an epigenome-wide association study (EWAS) power calculation technique previously described (Mansell et al., 2019). Thus, we determined that 97.4% of sites in the EPIC array (~850,000 sites) have power >90% to detect a 10% mean methylation difference (effect size) with N=85 samples (our discovery cohort sample size) at a significance threshold of $P<0.0001$.

Principal component analysis (PCA) was performed using PCAtools R package (v2.2.0) as an unsupervised method for data exploration in order to detect the greatest sources of variation in our dataset. We explored all the clinical variables available (**Table 1**) and found that disease status and age were the two main sources of variation in our dataset. Thus, we adjusted our analysis by age to correct its confounding effect.

To assess whether the assumptions of linear regression are satisfied, we used the gvlma R package (v1.0.0.3) considering a significance threshold ($P < 9.42 \times 10^{-8}$) previously described for the EPIC array (Mansell et al., 2019). As a result, 99.99% of sites in the EPIC array fulfill the linearity assumption (linear relationship between independent variables (both disease status and age) and dependent variable (CpG site methylation beta value)), 98.9% of sites

show homoscedasticity (constant variance of the residuals), 90.63% of sites present a symmetrical normal distribution of the residuals and 86.21% present a bell-shaped normal distribution of the residuals. Finally, a global test was performed providing an omnibus test of the four individual statistical tests, showing that 83.24% of sites in the EPIC array fulfill all linearity assumptions.

Furthermore, the differential methylation analysis between MIS-C and non-MIS-C cases was applied to beta values employing an empirical Bayesian framework linear model from the limma R package (v3.46.0), a model suitable for DNA methylation data (Mansell et al., 2019). limma operates on a matrix of methylation values, where each row is a probe (EPIC array site) and each column corresponds to a sample, and it fits a linear model to each row of data (Ritchie et al., 2015). For each probe, we have a vector of DNA methylation values and a design matrix that relates these values to some coefficients of interest (disease status and age). Thus, the dependent variable in the linear model is the methylation value and the independent variables are disease status and age.

References

Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 2016; **44** (W1): W90–7.

Mansell G, Gorrie-Stone TJ, Bao Y, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics* 2019; **20**: 366.

Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016 Mar;**8**(3): 389–99.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015 Apr 20; **43**(7): e47.

Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol* 2018; **19**: 64.

References of genetic loci associated to COVID-19 included in the study

1. Bastard P, Rosen LB, Zhang Q, et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* 2020; **370**: eabd4585.
2. Mousa M, Vurivi H, Kannout H, et al. Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates. *EBioMedicine* 2021; **74**: 103695.
3. Pairo-Castineira E, Clohisey S, Klaric L, et al. Genetic mechanisms of critical illness in Covid-19. *Nature* 2020; **591**: 92–98.
4. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, et al. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J* Med 2020; **383**: 1522–34.
5. Singh H, Choudhari R, Nema V, Khan AA. ACE2 and TMPRSS2 polymorphisms in various diseases with special reference to its impact on COVID-19 disease. *Microb Pathog* 2020; **150**: 104621.

6. van der Made CI, Simons A, Schuurs-Hoeijmakers J, et al. Presence of genetic variants among young men with severe COVID-19. *JAMA* 2020; **324**: 1–11.
7. Zhang Q, Bastard P, Liu Z, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* 2020; **370**: eabd4570.

References of genetic loci associated to MIS-C included in the study

1. Chou J, Platt CD, Habiballah S, et al. Mechanisms underlying genetic susceptibility to multisystem inflammatory syndrome in children (MIS-C). *J Allergy Clin Immunol* 2021; **148**(3): 732–738.e1.
2. Lee PY, Platt CD, Weeks S, et al. Immune dysregulation and multisystem inflammatory syndrome in children (MIS-C) in individuals with haploinsufficiency of SOCS1. *J Allergy Clin Immunol* 2020; **146**(5): 1194–1200.e1.

References of genetic loci associated to Kawasaki disease included in the study

1. Buda P, Chyb M, Smorczewska-Kiljan A, et al. Association Between rs12037447, rs146732504, rs151078858, rs55723436, and rs6094136 Polymorphisms and Kawasaki Disease in the Population of Polish Children. *Front Pediatr* 2021; **9**: 624798.
2. Burgner D, Davila S, Breunis WB, et al. A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. *PLoS Genet* 2009; **5**(1): e1000319.
3. Hoggart C, Shimizu C, Galassini R, et al. Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. *Eur J Hum Genet* 2021; **29**(12): 1734-1744.
4. Johnson TA, Mashimo Y, Wu JY, et al. Association of an IGHV3-66 gene variant with Kawasaki disease. *J Hum Genet* 2021; **66**(5): 475-489.
5. Khor CC, Davila S, Breunis WB, et al. Genome-wide association study identifies FCGR2A as a susceptibility locus for Kawasaki disease. *Nat Genet* 2011; **43**(12): 1241-6.
6. Kim JJ, Hong YM, Sohn S, et al. A genome-wide association analysis reveals 1p31 and 2p13.3 as susceptibility loci for Kawasaki disease. *Hum Genet* 2011; **129**(5): 487-95.
7. Kim JJ, Park YM, Yoon D, et al. Identification of KCNN2 as a susceptibility locus for coronary artery aneurysms in Kawasaki disease using genome-wide association analysis. *J Hum Genet*. 2013; **58**(8): 521-5.
8. Kuo HC, Li SC, Guo MM, et al. Genome-Wide Association Study Identifies Novel Susceptibility Genes Associated with Coronary Artery Aneurysm Formation in Kawasaki Disease. *PLoS One* 2016; **11**(5): e0154943.
9. Kwon YC, Kim JJ, Yu JJ, et al. Identification of the TIFAB Gene as a Susceptibility Locus for Coronary Artery Aneurysm in Patients with Kawasaki Disease. *Pediatr Cardiol* 2019; **40**(3): 483-488.
10. Kwon YC, Kim JJ, Yun SW, et al. Male-specific association of the FCGR2A His167Arg polymorphism with Kawasaki disease. *PLoS One* 2017; **12**(9): e0184248.
11. Lee YC, Kuo HC, Chang JS, et al. Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis. *Nat Genet* 2012; **44**(5): 522-5.
12. Onouchi Y, Ozaki K, Burns JC, et al. A genome-wide association study identifies three new risk loci for Kawasaki disease. *Nat Genet* 2012; **44**(5): 517-21.
13. Tsai FJ, Lee YC, Chang JS, et al. Identification of novel susceptibility Loci for kawasaki disease in a Han chinese population by a genome-wide association study. *PLoS One* 2011; **6**(2): e16853.

Chapter IV | Spatial transcriptomics unveils the *in situ* cellular and molecular hallmarks of the lung in fatal COVID-19

Spatial transcriptomics unveils the *in situ* cellular and molecular hallmarks of the lung in fatal COVID-19

Carlos A. Garcia-Prieto,[1,2] Eva Musulen,[1,3] Daniela Grases,[4] Veronica Davalos,[1,5] Gerardo Ferrer,[1,5] Belén Pérez-Miés,[5,6,7] Tamara Caniego-Casas,[5,6] José Palacios,[5,6,7] Xavier Saenz-Sardà,[8] Elisabet Englund,[8] Eduard Porta,[2,4] Manel Esteller[1,5,9,10]

[1]Cancer Epigenetics Group, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Catalonia, Spain
[2]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Catalonia, Spain
[3]Department of Pathology, Hospital Universitari General de Catalunya Grupo-QuirónSalud, Barcelona, Catalonia, Spain
[4]Cancer Immunogenomics Group, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Catalonia, Spain
[5]Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain
[6]Department of Pathology, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain
[7]Department of Medicine and Medical Specialties, University of Alcalá, Spain
[8]Division of Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
[9]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[10]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Catalonia, Spain

Correspondence: Dr Manel Esteller, Josep Carreras Leukaemia Research Institute (IJC), Carretera de Can Ruti, Camí de les Escoles s/n, 08916 Badalona, Barcelona, Catalonia, Spain. mesteller@carrerasresearch.org

ABSTRACT

Severe Coronavirus disease 2019 (COVID-19) can induce progressive diffuse alveolar damage (DAD) leading to fatal outcomes. Here, we leveraged data from Visium spatial transcriptomics for COVID-19 and control lung samples to unravel insights into cellular and molecular events driving lethal COVID-19. We report a progressive loss of endothelial cell types, pneumocytes type I and natural killer cells coupled with a continuous increase of myeloid and stromal cells, mostly peribronchial fibroblasts, over disease progression. Spatial organization analysis also identified specific compartmentalization including immune-specific clusters across DAD spectrum. Importantly, spatially informed ligand-receptor interaction analysis revealed intercellular communication signatures defining COVID-19 induced DAD. Transcription factor activity enrichment analysis identified the TGF-β pathway as DAD driver, highlighting the antagonizing roles of SMAD3 and SMAD7 activity during fibrosis. Integration of these data yielded a signaling kinase pathway in peribronchial fibroblasts with amenable novel targets. Finally, spatio-temporal trajectory analysis characterized an alveolar epithelium regeneration program.

Introduction

Infection by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus caused a worldwide pandemic of the derived coronavirus disease 2019, COVID-19. Beyond 775 million confirmed cases and more than 7 million deaths had been reported (https://data.who.int/dashboards/covid19/). Although most affected individuals exhibit mild clinical manifestations or are asymptomatic, respiratory failure linked to lung damage and acute respiratory distress syndrome could occur, being the most common cause of death[1,2]. Importantly, the maintenance of lung lesions in a subgroup of COVID-19 patients could also be associated to prolonged clinical manifestations[3,4]. Upon SARS-CoV-2 infection in the severe cases, the pathological change is defined by diffuse alveolar damage (DAD)[5] that initiates with an acute stage of early intra-alveolar epithelial lesions, interstitial inflammation and edema, followed by the proliferative stage with a final appearance of pneumocyte hyperplasia and fibroblast proliferation[1,2].

The molecular context of lung damage provoked by SARS-CoV-2 infection is not fully established. Most studies have addressed the bulk transcription landscape[6-8] or a particular lineage such as immune cells[9,10]. A more granulated view of the lung affected by severe COVID-19 has been gained by single-cell gene expression profiles[11-13]. However, all these approaches destroy the anatomical structure of the lung and the spatial cell-cell interactions. Only a limited number of studies have analyzed a spatial component in the lung of COVID-19 cases, using high-parameter imaging mass cytometry for targeted proteins[14] or focusing in regions of interest (ROI)[15-18]. Importantly, larger pathological scrutiny of spatial transcriptomics in COVID-19 patients has only been performed recently[19].

To overcome these issues, we leveraged the recently developed Visium spatial transcriptomics (VisiumST) technology in a cohort of lungs with normal histology and those that underwent DAD upon the course of fatal COVID-19. Using the single-cell RNA expression data of a human cell atlas of the lung to annotate cell types[13], we provide the constellation of shifts in forty-five cell types, the perturbations in cell-to-cell communications and the spatial alterations of molecular pathways that occur upon severe COVID-19.

Results

Identification of lung cell types and cellular compartments in COVID-19 associated DAD progression

To assess how fatal COVID-19 affected cell type composition, cell-cell communication, and global expression patterns across DAD progression, we followed the study design shown in **Fig. 1A**. We first retrieved twenty-three formalin-fixed paraffin-embedded post-mortem lung

tissue samples obtained from nineteen patients with DAD, corresponding to seven cases of acute DAD stage and twelve proliferative DAD stages, and four lung samples from control lungs with normal morphological appearance without SARS-CoV-2 infection (**Table 1**). We then analyzed the spatial transcriptomics patterns using tissue spots on a microarray slide with arrayed oligonucleotides to capture spatial gene expression information following the final generation of next generation sequencing (NGS) libraries[20], adapted by 10x Genomics as '10x Visium' (VisiumST) (**Methods**). In total, 91,068 tissue spots were studied after quality control (QC) and preprocessing (**Methods**). To assess the spatial organization of cell types across tissue slides, we used the integrated Human Lung Cell Atlas (HLCA)[13], as a reference to deconvolute the main cell types present in each spot by applying the validated cell2location pipeline[21]. Uniform Manifold Approximation and Projection (UMAP) using the integrated global spatial transcriptome data showing the spots colored by our three groups of lung samples (control, acute DAD, and proliferative DAD) distinguished these three entities (**Fig. 1B**). Illustrative examples are shown in **Fig. 1B**.

Using cell2location, we annotated forty-five cell types in HLCA defined by derived markers (**Methods**) in our lung samples (**Fig. S1**). A UMAP visualization of these cell populations according to their lineage (epithelial, stroma, immune and endothelial) in the spatial transcriptomics spots for all the integrated samples and illustrative examples are shown in **Fig. 1B**. The mapping of the identified cell types on top of the VisiumST brightfield images stained with Hematoxylin and Eosin (HE) for the described lineages in illustrative control, acute and proliferative DAD cases are shown in **Fig. 1C**. Cell types mapped to their expected locations, matching well-described structures, with epithelial cells lining the airway lumen and stromal cells mapping to blood vessel walls, as validated by endothelial (CD34) and epithelial (CK7) markers immunostaining (**Fig. 1C**).

Overall, we observed that in control lungs the most abundant lineage corresponded to endothelial cell types, as previously described[22], and that acute DAD was characterized by a decrease of the endothelial cell types and an increase in immune infiltrates, whereas the proliferative phase was mostly defined by the large proportion of fibrotic tissue, as previously described[1,2] (**Fig. 1D**). These findings were validated using HE stained sections and specific immunohistochemistry (IHC) markers such as CD34 (endothelial), CD68 (myeloid lineage) and CK7 (epithelial markers) and trichrome staining (fibroblasts) (**Fig. 1E**).

We then moved to characterize a possible uneven distribution of the identified 45 cell types according to their abundance in the control, acute and proliferative DAD lung groups (**Fig. 2A, 2B**). To ease the analyses interpretation, we ordered these cell populations according to their lineage (epithelial, stromal, immune, and endothelial) (**Fig. 2A**). Among the epithelial lineage (10 identified cell types), the most important difference was observed in alveolar type 1 (AT1)

115 cells, comprising between 5-10% of total cells in normal lungs as previously reported[23], that
116 were significantly downregulated in COVID-19 associated proliferative DAD cases in
117 comparison to controls and acute DAD lungs (**Fig. 2A, 2B**). The reverse process was observed
118 in alveolar type 2 (AT2) cells that were upregulated in proliferative cases in comparison to
119 acute DAD and control samples (**Fig. 2A, 2B**). These results fit with the concept that AT1s are
120 the main cell type responsible for provision of the interface for the blood-gas exchange (a
121 function that it is compromised in COVID-19 patients); whereas AT2 cells function as
122 progenitors that repair the injured alveoli epithelium[24]. Regarding the endothelial lineage (7
123 identified cell types), we observed that capillary cell types were the most abundant of lung cells
124 (~30% of total cells), as previously reported[22,23], and that the studied lung samples underwent
125 a progressive loss from normal lung to acute DAD to the final proliferative phase for the
126 abundance of endothelial cells (EC) aerocyte capillary, arterial, general capillary and venous
127 pulmonary (**Fig. 2A, 2B**). These results support the evidence linking SARS-CoV-2 infection to
128 endothelial dysfunction[25]. For the immune lineage (19 identified cell populations), we observed
129 in the myeloid lineage that alveolar macrophages (Mph) CCL3+ and monocyte-derived Mph
130 were significantly overrepresented in COVID-19 associated proliferative DAD cases in
131 comparison to controls and acute DAD lungs (**Fig. 2A, 2B**). Interstitial Mph perivascular
132 showed a progressive increase in the evolution of the disease. In the lymphoid lineage, a
133 decrease in CD4 and CD8 T-cells was observed in proliferative COVID-19 patients in
134 comparison to the control group (**Fig. 2A, 2B**). Interestingly, Natural Killer (NK) cells
135 experienced a significant decrease from control and acute DAD samples to proliferative DAD
136 cases (**Fig. 2A, 2B**). In this regard, these innate effector lymphocytes that respond to acute
137 viral infections have been previously related to COVID-19 severity[26]. Additionally, both alveolar
138 macrophages and plasma cells were overrepresented in proliferative DAD lungs compared to
139 control samples. Finally, regarding the stromal lineage (9 cell types), the most dramatic change
140 was observed for peribronchial fibroblasts (PBFs) that increased in the progression of the
141 disease from control to acute DAD phase~~s~~, skyrocketing in proliferative DAD cases (**Fig. 2A,**
142 **2B**). Not all fibroblast subtypes behaved in a similar manner. Subpleural fibroblasts were
143 overrepresented in COVID-19 associated DAD fatal cases compared to the control group and
144 between acute and proliferative stages, whereas alveolar fibroblasts and pericytes decreased
145 in the proliferative DAD samples compared to control and acute DAD lungs. Lastly, adventitial
146 fibroblasts were overrepresented in proliferative DAD samples compared to controls. UMAPs
147 for the entire set of cases in the cellular populations of AT1, AT2, EC aerocyte capillary, PBF,
148 monocyte-derived Mphs and smooth muscle activated stress response cells are displayed in
149 **Fig. 2C**, showing the unbalanced distribution (**Fig. 1B**). Mapping of the identified cell types on
150 top of the VisiumST brightfield images is shown in **Fig. 2B**. Cell types annotated with the finest
151 granularity level mapped to their expected locations, including subpleural fibroblasts located

next to the pleura, and multiciliated cells lining small airways lumen and smooth muscle cells mapping to blood vessel walls (**Fig. 2B**).

To assess the spatial distribution of the identified forty-five cell types within neighboring compartments, we applied the cell2location algorithm to the VisiumST data. For normal lung, the characterized cell types mapped within physiological cellular microenvironments such as the great compartment defined by immune and endothelial cells, one rich in epithelial cells (excluding AT2 proliferating, suprabasal and deuterosomal cells that shared a common location), another related to smooth muscle related cells and the fibroblast lineage (where PBFs resided in an isolated compartment) (**Fig. 2C**). This compartmentalization underwent an abrupt shift upon DAD progression. The acute DAD stage was characterized by a recruitment of an enriched AT2 proliferating population to the epithelial compartment; the irruption of a spatial cluster of macrophage subtypes and type-2 dendritic cells (DC2s, that promote cytotoxic T-cell responses and helper T-cell differentiation); and the appearance of plasma cells in an isolated population (**Fig. 2C**). Most of these cellular redistributions underwent further compartmentalization in the proliferative DAD stage that additionally exhibited the emergence of a unique compartment for lymphatic mature endothelial cells (**Fig. 2C**).

To further analyze and characterize tissue architecture differences we applied GraphCompass (Graph Comparison Tools for Differential Analyses in Spatial Systems)[27] (**Methods**), a set of designed graph analyses methods for “omics” data to quantitatively determine and compare spatial arrangement of distinct cell types among different biological conditions successfully applied to VisiumST data[27]. The cell-type-specific subgraphs across condition pairs, where the size of the dot indicates the pairwise similarity score variances (**Methods**), is shown in **Fig. 2D**. Among the characterized distinct cells through COVID-19 progression, we further analyzed PBFs, endothelial aerocyte capillary cells and AT2 cells by plotting filtration curves for every sample, as well as the mean curve for each lung stage (**Methods**) (**Fig. 2E**). These analyses reinforced the findings that these cell types underwent antiparallel shifts in their abundance upon DAD progression: PBFs and AT2 cells exhibited an overrepresentation whereas endothelial aerocyte capillary cells were depleted, particularly at the proliferative stage (**Fig. 2E**).

Spatial cell-cell interactions in the spectrum of the disease

One of the most exciting applications of spatial transcriptomics is the potential to analyze cell-cell communication (CCC). CCC is a multicellular and complex process involving multiple mechanisms, including intercellular signaling and intracellular signaling as a downstream response. Related to the intercellular component, cells interact on diverse levels that include direct contact, such as between ligands and surface receptors, and through indirect means, such as the release of soluble factors. For single-cell analyses, the molecular profiles of sender

and receiver cell types allow the inference of underlying cell communication events in a tissue using co-occurrence of ligand and receptor (LR) expression among the candidate communicating cells[28,29] and through gene expression profiles in the receiving cell type related to the extracellular interaction[28,29]. Herein, we used the VisiumST data to identify coordinated cell-cell communication signatures shared across all tissue slides by applying non-negative matrix factorization (NMF) to the estimated local (spot level) ligand-receptor interactions (LRIs), calculated using spatially-weighted Cosine similarity with LIANA+[30,31] (**Methods**). Using the elbow selection procedure, we decomposed the local interactions into three Factors (1, 2 and 3) representing three different intercellular communication signatures. The NMF factor scores indicate the strength of each factor in each spot, representing the degree of influence by the associated signature. The averaged factor scores per tissue slide clustered according to lung status are shown in **Fig. 3A**. Importantly, Factor 3 distinguished the best between control and COVID-19 associated DAD lung tissues, with high mean scores in proliferative DAD (**Fig. 3A**). Factor 1 was most prominent in control samples, whereas Factor 2 was more active in a subset of proliferative DAD.

To provide further biological insight, we performed LRIs pathway enrichment analysis on the distinct interaction loadings contributing to the three factors (**Fig. 3B**) using multivariate linear regression[32] and pathway annotations from PROGENy[33] (**Methods**). We found that DAD associated Factor 3 was significantly enriched in interactions related to the transforming growth factor beta (TGF-β) pathway (**Fig. 3B**), a driver of fibrosis involved in response to tissue injury[34]. Conversely, DAD associated Factor 3 was depleted for the EGFR pathway (**Fig. 3B**). Interestingly, the wingless-related integration site (WNT) pathway was enriched in Factor 1 (characteristic of the control samples) but depleted in Factor 2 (**Fig. 3B**).

Since Factor 3 was the more optimal discriminator between healthy and COVID-19 associated DAD, we mainly focused our analyses in this CCC readout. The top four LR loadings defining DAD associated Factor 3 were the interactions TIMP1^CD63, that promotes lung fibrosis mediated by the TGF-β1/SMAD3 pathway[35]; APP^CD74, highlighting the role of APP (Amyloid-beta precursor protein) as a lung capillary barrier defense during infection[36]; CD99^CD81, both regulators of T-cell and B-cell activity[37]; and LUM^ITGB1, with LUM being linked to extracellular matrix (ECM) remodeling and inflammation-associated fibroblasts[38] (**Table S1**). Noteworthy additional LRIs involved PSAP (PSAP^LRP1) and members of the S100 protein family S100A8 (S100A8^AGER, S100A8^ITGB2) and S100A9 (S100A9^AGER, S100A9^CD68, S100A9^ITGB2), that have been reported to activate macrophages in COVID-19[11,14]; the SPARC protein (SPARC^ENG), a downstream effector of TGF-β upregulated in COVID-19-associated fibrosis[11,39] and idiopathic pulmonary fibrosis (IPF)[40]; and vimentin (VIM^CD44), an attachment factor for SARS-CoV-2 entry into endothelial cells[41]. Beyond the

mentioned ligand proteins, it is also relevant to mention that the three most frequent receptors involved in the LRIs (**Table S1**) were CD44, involved in T-cell abundance and fostering cytokine storm linked to COVID-19 poor prognosis[42]; ITGB1, that associates with the angiotensin-converting enzyme 2 (ACE2) to mediate SARS-CoV-2 entry[43] and regulates ECM remodeling; and LRP1, involved in the overproduction of cytokines and chemokines[42], and enriched in mortal and long COVID-19 patients[44,45]. The NMF factor scores indicating the strength of each factor in each spot are depicted for illustrative examples of VisiumST slide images (**Fig. 3C**).

Regarding intracellular signaling analysis, we studied transcription factor (TF) activities including their downstream transcriptional targets that shift in the progression from normal lung to the COVID-19 stages. We estimated TF activity in each VisiumST spot based on multivariate linear regression using decoupleR[32] and CollecTRI[46] network containing a curated collection of TFs and their targets (**Methods**). **Table S2** shows the TFs activity in each type of lung sample and **Fig. 3D** displays the top 10 scaled TF enrichment scores that discriminate each condition. These data highlight again the critical role of TGF-β pathway in DAD progression, as did also the intercellular LRI analysis (**Fig. 3B**). It is noteworthy to mention the opposite activity landscape of SMAD protein family members. Proliferative DAD stage is characterized by SMAD3 upregulation, a key mediator of TGF-β signaling to promote ECM production, tissue repair, fibrosis and scar formation[47]. Conversely, SMAD7 activity, that exerts antagonizing roles to TGF-β/SMAD3 profibrotic pathway, is downregulated across DAD progression (**Fig. 3E**). The altered TGF-β signaling pathway was further strengthened by two additional components, TGFB1|1 [a marker of contractile smooth muscle cells] and ZEB2 [involved in fibrogenesis[48], that were upregulated in control and proliferative DAD lungs, respectively (**Fig. S2**). Interestingly for the last gene, ZEB2 DNA methylation status has been linked to another severe consequence of SARS-CoV-2 infection, the multisystem inflammatory syndrome in children[49]. Two additional cellular networks were targeted by aberrant TF activity: lung epithelial cell differentiation and NK cells functionality. In the first case, downregulation of the AT2 cell identity regulator ETV5 occurred upon COVID-19 induced DAD progression (**Fig. 3E**), suggesting initiation of epithelial regeneration by AT2 cells[11]; whereas NKX2-1 (regulator of alveolar epithelial progenitors)[50], MYB [involved in airway epithelial cell differentiation[51] and BHLHA15 [linked to acinar cell function] were upregulated in proliferative DADs (**Fig. S2**). Remarkedly, SREBF2, related to surfactant production in AT2 cells, and CIITA, which drives MHCII expression and induces cell resistance to SARS-CoV-2[52], was enriched in acute DAD lungs (**Fig. S2**). For NK cells, where we observed a decrease in COVID-19 progression (**Fig. 2A**), multiple TFs essential for their proper development were also downregulated in the DAD lungs, such as IRF2, IKZF1, NFIL3 and EOMES (**Fig. S2**), supporting that NK cell dysfunction in COVID-19 could be associated to lethality.

Spatial relationships of ligand-receptor interactions and transcription factor activities with cell type abundance

We next leveraged an explainable multi-view modelling approach[53] to decipher the global spatial relationships between LRIs, TFs and the distinct cell type abundances. In this regard, we jointly modelled cell type abundances in each spatial spot using the top 25 local LR loadings from Factor 3 (**Table S1**) and the activity of the top 10 most enriched TFs (**Table S2**) per condition. We observed that across control and DAD lung tissue slides, both LRIs and TFs activity jointly contributed to explain the variance of cell type abundance (mean multi-view R^2 = 0.21) (**Fig. 3F**). The relative contribution to the joint predictive performance was higher for TF activities (median contribution 65%) compared to LRIs (median contribution 35%) across disease progression (**Fig. 3F**). Importantly, the abundance of fibroblasts was best explained by the activity of TFs SMAD3 and SMAD7, including TIMP1^CD63 LRI among top 10 predictors of PBFs (**Fig. 3F**), highlighting their contribution to DAD. Moreover, important differences were also found when explaining cellular composition across control lungs and COVID-19 induced DAD progression, particularly for PBFs, endothelial aerocyte capillary cells, endothelial general capillary cells, alveolar Mphs subtypes, monocyte-derived Mphs, non-classical monocytes and NK cells (**Fig. S3**).

Additionally, we used the predictor importances (coefficients' t-values) from this predictive linear model to infer intracellular signaling networks linking both LRIs and TF activity patterns. In this regard, we applied LIANA+[31] (**Methods**) to infer a putative causal network linking LRIs to TFs focusing on PBFs considering their significant enrichment in proliferative DAD lungs (**Fig. 3G**) and in IPF patients[54]. We found that initial activation of the ITGB1 receptor unleashed a downstream signaling pathway that upregulated the transcription factors SMAD3, MYB and BRCA1 and downregulated SMAD7 and CIITA (**Fig. 3G**). Importantly, the ITGB1 mediated repression CIITA (**Fig. 3G**) could activate collagen expression by lung fibroblasts after injury[55]. Interestingly, we also found that integrin ITGB1 induces an activation of the kinases ILK, AKT1, MAP2K4 and MAPK14 (**Fig. 3G**), previously linked to the disorder[56], becoming amenable candidates for therapies.

To identify local spatial dependencies that might occur only in a sub-region of the studied lung tissues, we leveraged spatially-informed local bivariate similarity metrics, that included spatially-weighted Cosine similarity and global Moran´s R (**Fig. 4A**), to identify pairs of LRIs that are spatially clustered together or apart (**Methods**). The LRI TIMP1^CD63 showed the highest spatial co-clustering pattern by both metrics (**Fig. 4A**), particularly for proliferative DAD lungs within DAD associated Factor 3 boundaries (**Fig. 4B and S4**). Computed permutation-based p-values to assess the significance of the local interactions demonstrated an agreement with the high Cosine similarity regions (**Fig. 4B**). To further categorize TIMP1^CD63 spatial

297 relationship, we identified that for most local category areas, both ligand and receptor were
298 highly expressed (**Fig. 4B**). Interestingly, APP participated in multiple LRIs, including
299 APP^CD74 with the second highest Cosine similarity (**Fig. 4A**). An additional APP LRI,
300 APP^AGER, showed a clear diminishing spatial co-clustering pattern over disease progression
301 (**Fig. S4, S5**), suggesting a biological relationship since AGER is an AT1 marker[54], also
302 observed in our study (**Fig. S1**). Intriguingly, beta-amyloid produced by the infection-mediated
303 lung injury can reach through general circulation other organs originating further defects,
304 including neurocognitive dysfunction[36]. Cognitive impairment in the post-acute phases of
305 COVID-19 is not uncommon[55] and SARS-CoV-2 infection is considered a risk factor for
306 Alzheimer´s disease[56]. Finally, multiple LRIs with proteins involved in the activation of Mphs in
307 COVID-19, such as S100A9, showed increased spatial co-clustering through disease
308 progression (**Fig. S4, S6**). Interestingly, S100A9^CD68, the best LRI predictor of all Mph
309 subtypes, non-classical monocytes and T cells proliferating cell type abundances (**Fig. S6,**
310 **S7**), yielded one of the highest median global Moran´s R scores (**Fig. 4A**), suggesting an
311 important role in aberrant myeloid activation and dysregulated immune response[11,59]. The
312 mapping of the described LRIs on the VisiumST images is shown in **Fig. 4B, S5**, **S6**.
313 We next investigated associations between cell types and TFs activity considering their
314 relevance for tissue function, using also spatially-weighted Cosine similarity and global
315 Moran´s R (**Fig. 4C**) (**Methods**). PBFs, the most abundant cell type in proliferative COVID-19
316 (**Fig. 2A**), were most spatially associated and co-clustered with the pro-fibrotic TF SMAD3
317 activity locations (**Fig. 4C, 4D**), whereas PBFs and the anti-fibrotic TF SMAD7 activity locations
318 were spatially clustered apart (**Fig. 4C, 4D**). These results fit the mutually exclusive location of
319 SMAD3 and SMAD7 activities through DAD progression (**Fig. 3E**); and the top LRI loading,
320 TIMP1^CD63, characterizing DAD associated Factor 3 promoting lung fibrosis through the
321 TGF-β1/SMAD3 pathway (**Fig. 4B**). All these results highlight the central role of TGF-β
322 pathway activation in driving pathological ECM remodeling and repair linked to aberrant
323 activation of PBFs[60] that leads to scar formation and a grossly disrupted lung tissue
324 architecture in the COVID-19 proliferative stages.
325 To provide a second example (beyond PBFs) of local spatial relationships between TFs
326 activity and cell types, it is worth highlighting the myeloid lineage. The activity of the TF MYB
327 was the best predictor of the abundance of myeloid cell types (**Fig. S7**). The spatial co-
328 expression of MYB with the myeloid and T cell proliferating cell fates increases as disease
329 progresses (**Fig. S8**). MYB plays an essential role in many hematopoietic pathways and, most
330 importantly, the E2F/MYB regulatory programs from myeloid cell populations are
331 hyperactivated in severe COVID-19 cases[61]. The interrogation of the immune landscape also
332 unveiled that NK cells, another population critically depleted in our COVID-19 associated DAD
333 cases (**Fig. 2A**), were most influenced by IKZF1 activity showing the highest co-clustering

pattern (**Fig. S8**). IKZF1 is essential for proper NK cell development[62]. Herein we report a loss of the co-clustering pattern of IKFZ1 and NKs across COVID-19 associated DAD progression (**Fig. S8**). Results that strengthen the suggested central role of NK cell dysfunction in fatal COVID-19[63].

Cell-cell communication as a function of niche composition

The potential cell-cell communication events that could occur in lung tissues across the different conditions was assessed not only accounting for LRIs and TFs activity. We also used a graph neural network method (NCEM)[64] that estimates the effect of the inferred spot composition on gene expression variation within cell types across spots to discover intercellular dependencies (**Methods**). To discriminate different intercellular dependencies between control and fatal COVID-19 lung sections, we identified multiple cell types over disease progression (**Methods)** observing a profound reconfiguration of intercellular communication (**Fig. 5A**). We observed a dependency of NK cells in control lung tissues on various cell types, including CD8 T cells, EC aerocyte capillary cells, alveolar Mph CCL3+ and AT1 cells. However, these dependencies were lost in acute and proliferative COVID-19 lung tissues (**Fig. 5A**). We also found that in DAD samples the population of CD4 T cells beared multiple dependencies on various cell types such as non-classical monocytes, SM activated stress response cells, plasma cells and AT1 cells, becoming a prominent receiver node of communication, particularly in acute DAD (**Fig. 5A**). Interestingly, AT1 cells exhibited limited intercellular dependencies in control lung samples, that increased in the acute DAD phase, and were lost in the proliferative DAD stage (**Fig. 5A**). Lastly, CD4 T cells established a dependency on AT2 cells in the proliferative DAD phase (**Fig. 5A**). Additionally, we performed a receiver effect analysis highlighting gene-wise effects of all senders on once receiver cell type to contextualize gene expression differences in some of the couplings (**Fig. 5B**). Furthermore, since CD4 T cells showed important dependencies on multiple cell types in DAD lungs, we performed a sender similarity analysis to characterize the profile of these intercellular dependencies across DAD progression. We observed that in acute DAD lungs, the sender profile mostly conserved lineage cell type identity but was lost in proliferative DAD lungs (**Fig. 5B**).

Spatio-temporal trajectories: AT2-AT1 epithelial regeneration

AT2 cells play an important role as AT1 progenitors during lung injury, proliferating and contributing to alveolar repair and regeneration. By contrast, AT1 cells are fragile, susceptible to damage and unable to proliferate. Therefore, characterizing AT2-AT1 differentiation is important to provide novel insights into cellular processes and tissue repair mechanisms in severe COVID-19. To decipher the dynamic relationships across tissue space and time

between transcriptional states of AT2 and AT1 cells, we leveraged a spatial graph-based method named pseudo-time-space (PSTS) implemented in the stLearn software[65] (**Methods**). By combining spatial and imaging information, representing cell and tissue morphology, with gene expression data, we used the PSTS algorithm to map the spatial changes in AT2 and AT1 cell states, modelling and reconstructing their spatio-temporal trajectories. In this regard, we defined a spatial trajectory for AT2 cells transitioning into AT1 cells across two clusters (clade 6 and clade 9) in a proliferative DAD lung (**Fig. 5C**). The top 10 most upregulated and downregulated genes defining AT2 to AT1 transition in each clade are shown (**Fig. 5C**). Enrichment analysis (**Methods**) revealed that the top 10 upregulated genes in clade 6 were enriched in AT1 cell identity markers including ICAM1, DPYSL2, ANGPTL2 and AGER[22]. Likewise, the top 10 downregulated genes in clade 6 were enriched in AT2 cell identity markers including SFTPB, SCD and CYB5A[22]. Furthermore, the top 10 downregulated genes in clade 9 were also enriched in AT2 cell identity markers including SFTPB, SFTPC, SLC34A2, FASN, CTSH, DBI, MLPH, LPCAT1 and CYB5A[22]. These results further validate the inferred spatio-temporal trajectories in AT2 and AT1 cell states, better characterizing the alveolar epithelial regeneration process after lung injury. Interestingly, when comparing clade 6 and clade 9 (**Fig. 5C**), we found that CAV1, a late AT1 maturation marker[11] was upregulated in clade 9, suggesting a complete transition of AT2 to AT1 cells. This is further supported by the cell type specific gene expression of CAV1 in spatial coordinates among different subtypes of pneumocytes, where CAV1 is expressed only by AT1 cells in distinct locations but not by co-located AT2 and derived AT0 cells during alveolar repair, driving most of CAV1 total expression across all cell types (**Fig. 5C**). Altogether, these results reinforce the notion that, in the lung of patients with severe COVID-19, AT2 cells aim to repopulate AT1 cells upon the activation of alveolar epithelial regeneration programs.

Discussion

The wealth of transcriptomic information in COVID-19 is mostly derived from disaggregated lung tissues where the architecture of the organ is not preserved and, thus, the effect of the surrounding tissue microenvironment on gene expression is lost. This issue has only recently been investigated in COVID-19 affected lung tissues, but mostly for targeted markers or discrete locations[14-19]. In a similar manner that very recently VisiumST has spatially resolved transcriptomes in IPF[66]; our spatial transcriptomic analysis, with the most recently developed bioinformatic tools, delivers a landscape of the lung in fatal COVID-19 defined by profound shifts in specific cellular populations with distorted intercellular communications that finally disrupt important gene networks.

The disbalance between cell types exhibited the most remarkable change for PBFs that experimented an extraordinary increase in the proliferative stage of the disease. Other

subclasses of fibroblasts also underwent upregulation among COVID-19 progression, except for alveolar fibroblasts, that maintain tissue homeostasis with an important role in regulating lung fibrosis[60] and that decreased in a similar fashion of other cell types populating the functional respiratory alveolus such as epithelial AT1 cells and aerocyte capillary endothelial cells. Interestingly, the proliferative DAD phases showed an increase in various subtypes of alveolar Mphs. Furthermore, the myeloid lineage underwent an overall increase, including interstitial and monocyte-derived Mphs and non-classical monocytes, whereas the lymphoid lineage decreased across DAD progression, mostly driven by the loss of NK cells in proliferative DAD lungs. These results match the reported aberrant activation of myeloid cells and impaired T cell and NK cell responses in fatal COVID-19[11,63]. In this regard, we highlight the important role of NK cells in severe COVID-19, characterizing NK cells dysfunction with marked downregulation of essential TFs activity for their functional development and maturation, particularly for IKZF1[62]. Herein, we report a progressive and persistent dysfunction of NK cells throughout DAD spectrum, further implicating their impairment in fatal COVID-19[63]. Besides NK cells dysfunction, a dysregulated immunological repair response to SARS-Cov-2 infection has been proposed as a major contributor to disease progression[59]. Hence, we have described the LRIs and TFs activity related to essential elements of the immune system, identifying S100A9^CD68 LRI and MYB TF activity as major determinants of myeloid cell types abundance, showing increased spatial co-expression patterns over DAD progression, strengthening their role in the activation of Mphs in severe COVID-19[14,61].

Moreover, we have characterized the molecular drivers of pathological responses to lung injury leading to massive fibrosis and grossly disrupted tissue architecture. We report the key role of TGF-β pathway in DAD progression, identifying an upregulation of profibrotic SMAD3 coupled with downregulation of its antagonist SMAD7. Importantly, we have identified TIMP1^CD63 LRI as a major contributor to DAD, emphasizing TIMP1 role as a key regulator of ECM homeostasis and downstream effector of TGF-β pathway activation, becoming a candidate therapy target for pulmonary fibrosis[67]. Furthermore, we inferred an intracellular signaling network in PBFs suggesting that the phenotypic changes and the different targeting of the SMAD TFs involved the activation of integrin ITGB1 receptor and their associated downstream kinases AKT1, MAP2K4 and MAPK14, representing additional potential targets for COVID-19 therapies[56].

Interestingly, when analyzing intercellular dependencies as a function of niche composition, we described a dependency of NK cells on various cell types in control lungs that was lost on DAD lungs, whereas a dependency of CD4 T cells on multiple cell types including CD8 T cells in control lungs shifted towards other lymphoid and myeloid immune cells and stromal cells in DAD, such as non-classical monocytes, plasma cells and smooth muscle activated stress

response cell types, including epithelial AT1 cells in acute DAD and AT2 cells in proliferative DAD, further reflecting the disrupted tissue architecture across the DAD spectrum.

Importantly, our spatio-temporal trajectories analysis helps to characterize the alveolar epithelial regeneration process, highlighting the important role of AT2 cells as AT1 progenitors and identifying markers of AT2 to AT1 differentiation. This process is impaired in severe COVID-19 induced DAD cases, contributing to fatal outcomes in these patients. In this regard, it is tempting to speculate that long COVID-19 could relate to the partial or complete disruption of intercellular communication events and differentiation trajectories that cannot completely restore the functional alveolar gas exchange capacity and/or prevent the persistence of fibrotic scars.

Our findings could also provide and foster the research of small drugs and antibodies targeting some of the cell types, pathways and intercellular communication that characterize the spatial aftermath of severe COVID-19. One example could be related to the combating of the fibrosis associated with DAD progression. Pirfenidone and nintedanib are two antifibrotic drugs approved for the treatment of IPF that could be repurposed to avoid severe COVID-19 associated fibrosis since pirfenidone inhibits TGF-β[68], the main profibrotic pathway underpinned in our study, and nintedanib blocks several tyrosine and serine/threonine kinases[69,70], among them the MAPKs and AKT1 identified in our intracellular pathways analyses of PBFs. In this regard, the herein identified MAPK14 kinase activation within the PBF intracellular signaling pathway leading to SMAD3 activation, suggest the use of MAPK14 inhibitors such as ralimetinib and ARRY-797. both exhibiting SARS-Cov-2 antiviral activity[56]. Interestingly, several inhibitors of the integrin $\alpha v\beta 6$, such as GSK3335103 and BG00011 are also at different levels of preclinical studies or even in clinical trials to treat IPF. Our finding that integrin ITGB1 activation unleashes a profibrotic signaling in PBFs in fatal COVID-19 cases suggests that this receptor can also be another amenable target. A similar case can be drawn for the emerging and hyperactivated population of inflammatory myeloid cells that we have observed. In this regard, Mphs and monocytes engage the NOD-like receptor family pyrin domain containing 3 (NLRP3) inflammasome in COVID-19[71]. Thus, NLRP3 inhibitors, such as NT0793/NT0249 or MCC950[71], can be other useful agents to restrict the activity of these belligerent myeloid cells. Importantly, because our integrative spatial transcriptomics analysis of TFs activity yielded MYB as a major determinant factor for Mphs in severe COVID-19, it could be proposed as a candidate drug target[72]. Similarly, we report S100A9^CD68 LRI as a major determinant of myeloid cell types. Interestingly, S100 protein family members S100A8 and S100A9 have been investigated as potential targets of small molecules such as paquinimod to control aberrant myeloid activation in COVID-19[73].

Related to treatment, we can also briefly mention a provocative thought. Since multiple LRIs involved APP (such as APP^CD74); cognition defects in post-COVID-19[57] could be due to

beta-amyloid liberated to the blood by the lung lesions[36]; and SARS-CoV-2 infection is a risk factor for Alzheimer´s disease[58], we suggest that maybe we can target both processes: the lung injury and the associated cognitive impairment. For example, lysophosphatidic acid receptors (LPAR) from the G protein-coupled receptor family contribute to Alzheimer's disease, but also bind to the viral SPIKE protein; thus, LPAR inhibitors are investigated for this potential double effect[74].

Our study is unique because we provide a spatially informed characterization of the cellular and molecular hallmarks of lung tissue architecture in fatal COVID-19. This detailed spatial transcriptomics study that highlights *in situ* the disease-associated changes in the composition of cellular subsets, their spatial dependencies and disrupted intercellular communication programs also constitutes a proof-of-principle of the potential translational use of the emerging spatial technologies. This transition will require careful benchmark comparison studies among the competing spatial transcriptomic platforms, harmonization of data processing pipelines and design of user-friendly databases where the data can be deposited and interrogated, automatization of sample processing and data analysis workflows leading to a shorter timeframe to deliver the results together with the ongoing reduction of sequencing costs; and scalable computational methods to exploit spatial transcriptomics data. Related to this last point, spatial transcriptomics can constitute one of the entrance points for the application of artificial intelligence in pathology and modern medicine. In this regard, our investigation of the altered cellular and molecular architecture of the lung in fatal COVID-19 could serve as an excellent example of the versatility of spatial transcriptomics to fulfill the promise of how the new genomic technologies could improve our understanding and the personalized management of many human diseases.

Contributors

C. A. G.-P. provided the bioinformatic analyses of the spatial transcriptomics data. E. M., V. D. and G. F. reviewed the clinical data. E. M. performed the immunostainings. D. G. performed the spatial transcriptomic experiments. X. S. S., E. E., B. P. M., T. C. C., J. Pal. and E. M. reviewed the postmortem lung samples. E. P. provided bioinformatic expertise. M. E. designed the study and wrote the manuscript with contributions from all authors. The study was approved by the institutional ethical review boards of Ramón y Cajal University Hospital (Necropsias_Covid19; 355_20) and the Lund Hospital (ref 2020- 02369).

Declaration of interests

Dr. Esteller declares past grants from Ferrer International and Incyte and personal fees from Quimatryx, outside the submitted work.

Data sharing

The data discussed in this publication, including all Visium datasets, sample metadata, raw and processed data (Space Ranger output), have been deposited in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE271370 (https://www.ncbi.nlm.nih.gov./geo/query/acc.cgi?acc=GSE271370) (reviewer access with token oraxwgqmllmxzeb).

Code availability

The code used for preprocessing and downstream data analyses performed in Python and used to produce all analyses and figures presented in this publication is available at the GitHub repository via https://github.com/carlosgarciaprieto/SpatialCOVID19.

Acknowledgements

We thank CERCA Programme / Generalitat de Catalunya for institutional support. The Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia has provided funding to ME (2021 SGR01494). ME has also received funding from the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033/ERDF 'A way to make Europe' (PID2021-125282OB-I00), Cellex Foundation (CEL007) and "la Caixa" Foundation (LCF/PR/HR22/00732). M.E. is an ICREA Research Professor. GF received support by Fundacio La Marato de TV3 (ref 202131-32). BP-M and JP is supported by Instituto de Salud Carlos III (ISCIII) (PI22/01892, PMP22/00054, PMP21/00107). EE is supported by Region Skane funds.

References

1. Milross, L. et al. Post-mortem lung tissue: the fossil record of the pathophysiology and immunopathology of severe COVID-19. *Lancet Respir. Med.* **10**, 95-106 (2022).
2. Bridges, J.P. et al. Respiratory epithelial cell responses to SARS-CoV-2 in COVID-19. *Thorax* **77**, 203-209 (2022).
3. Adeloye, D. et al. The long-term sequelae of COVID-19: an international consensus on research priorities for patients with pre-existing and new-onset airways disease. *Lancet Respir. Med.* **9**, 1467-1478 (2021).
4. Davis, H.E. et al. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133-146 (2023).
5. Erjefält, J.S. et al. Diffuse alveolar damage patterns reflect the immunological and molecular heterogeneity in fatal COVID-19. *EBioMedicine*. **83**, 104229 (2022).
6. Blanco-Melo, D. et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036-1045.e9 (2020).
7. Pinto, B.G.G. et al. ACE2 Expression Is Increased in the Lungs of Patients With Comorbidities Associated With Severe COVID-19. *J. Infect. Dis.* **222**, 556-563 (2020).
8. D'Agnillo, F. et al. Lung epithelial and endothelial damage, loss of tissue repair, inhibition of fibrinolysis, and cellular senescence in fatal COVID-19. *Sci. Transl. Med.* **13**, eabj7790 (2021).
9. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842-844 (2020).
10. Wilk, A.J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070-1076 (2020).
11. Melms, J.C. et al. A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114-119 (2021).
12. Delorey, T.M. et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107-113 (2021).
13. Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563-1577 (2023).
14. Rendeiro, A.F. et al. The spatial landscape of lung pathology during COVID-19 progression. *Nature*. **593**, 564-569 (2021).
15. Desai, N. et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat. Commun* **11**, 6319 (2020).
16. Margaroli, C. et al. Spatial mapping of SARS-CoV-2 and H1N1 lung injury identifies differential transcriptional signatures. *Cell Rep. Med.* **2**, 100242 (2021).
17. Park, J. et al. System-wide transcriptome damage and tissue identity loss in COVID-19 patients. *Cell Rep. Med*. **3**, 100522 (2022).
18. Milross, L. et al. Distinct lung cell signatures define the temporal evolution of diffuse alveolar damage in fatal COVID-19. *EbioMedicine* **99**, 104945 (2024).
19. Mothes, R. et al. Distinct tissue niches direct lung immunopathology via CCL18 and CCL21 in severe COVID-19. *Nat. Commun*. **14**, 791 (2023).
20. Rao, A. et al. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211-220 (2021).
21. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661-671 (2022).
22. Travaglini, K.J., et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619-625 (2020).
23. Crapo, J.D. et al. Cell number and cell characteristics of the normal human lung. *Am. Rev. Respir. Dis.* **126**, 332-337 (1982).
24. Chan, M. & Liu, Y. Function of epithelial stem cell in the repair of alveolar injury. *Stem Cell Res. Ther.* **13**, 170 (2022).
25. Xu, S.W., Ilyas, I. & Weng, J.P. Endothelial dysfunction in COVID-19: an overview of evidence, biomarkers, mechanisms and potential therapies. *Acta. Pharmacol. Sin.* **44**, 695-709 (2023).
26. Maucourant, C. et al. Natural killer cell immunotypes related to COVID-19 disease severity. *Sc.i Immunol.* **5**, eabd6832 (2020).
27. Ali, M., et al. GraphCompass: spatial metrics for differential analyses of cell organization across conditions. *Bioinformatics* **40**, i548-i557 (2024).
28. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
29. Efremova, M. et al. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

30. Dimitrov, D. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).
31. Dimitrov, D. et al. LIANA+: an all-in-one cell-cell communication framework. *bioRxiv* 2023.08.19.553863; doi: https://doi.org/10.1101/2023.08.19.553863.
32. Badia-I-Mompel, P. et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform. Adv.* **2**, vbac016 (2022).
33. Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).
34. Chanda, D. et al. Developmental pathways in the pathogenesis of lung fibrosis. *Mol. Aspects Med.* **65**, 56-69 (2019).
35. Duch, P. et al. Aberrant TIMP-1 overexpression in tumor-associated fibroblasts drives tumor progression through CD63 in lung adenocarcinoma. *Matrix Biol.* **111**, 207-225 (2022).
36. Balczon, R. et al. Lung endothelium, tau, and amyloids in health and disease. *Physiol. Rev*. **104**, 533-587 (2024).
37. Pata, S. et al. Association of CD99 short and long forms with MHC class I, MHC class II and tetraspanin CD81 and recruitment into immunological synapses. *BMC Res. Notes*. **4**, 293 (2011).
38. Tao, Z., Huang, J. & Li. J. Comprehensive intratumoral heterogeneity landscaping of liver hepatocellular carcinoma and discerning of APLP2 in cancer progression. *Environ. Toxicol.* **39**, 612-625 (2024).
39. Pérez-Mies, B. et al. Progression to lung fibrosis in severe COVID-19 patients: A morphological and transcriptomic study in postmortem samples. Front Med (Lausanne). **9**, 976759 (2022).
40. Conforti, F. et al. Paracrine SPARC signaling dysregulates alveolar epithelial barrier integrity and function in lung fibrosis. *Cell Death Discov.* **6**, 54 (2020).
41. Amraei, R. et al. Extracellular vimentin is an attachment factor that facilitates SARS-CoV-2 entry into human endothelial cells. *Proc. Natl. Acad. Sci USA*. **119**, e2113874119 (2022).
42. Zick, Y. Galectin-8, cytokines, and the storm. *Biochem Soc. Trans.* **50**, 135-149 (2022).
43. Zhang, Y. et al. An antibody-based proximity labeling map reveals mechanisms of SARS-CoV-2 inhibition of antiviral immunity. *Cell Chem. Biol.* **29**, 5-18.e6 (2022).
44. Razaghi, A. et al. Proteomic Analysis of Pleural Effusions from COVID-19 Deceased Patients: Enhanced Inflammatory Markers. *Diagnostics* (Basel). **12**, 2789 (2022).
45. Gu, X. et al. Probing long COVID through a proteomic lens: a comprehensive two-year longitudinal cohort study of hospitalised survivors. *EBioMedicine* **98**,104851 (2023).
46. 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).
47. Finnson, K.W. et al. Endoglin differentially regulates TGF-β-induced Smad2/3 and Smad1/5 signalling and its expression correlates with extracellular matrix production and cellular differentiation state in human chondrocytes. *Osteoarthritis Cartilage* **18**, 1518-1527 (2010).
48. Teraishi, M. et al. Critical involvement of ZEB2 in collagen fibrillogenesis: the molecular similarity between Mowat-Wilson syndrome and Ehlers-Danlos syndrome. *Sci. Rep.* **7**, 46565 (2017).
49. Davalos, V. et al. Epigenetic profiling linked to multisystem inflammatory syndrome in children (MIS-C): A multicenter, retrospective study. *EClinicalMedicine* **50**, 101515 (2022).
50. Toth, A. et al. Alveolar epithelial progenitor cells require Nkx2-1 to maintain progenitor-specific epigenomic state during lung homeostasis and regeneration. *Nat. Commun.* **14**, 8452 (2023).
51. Pan, J.H. et al. Myb permits multilineage airway epithelial cell differentiation. *Stem Cells* **32**, 3245-3256 (2014).
52. Bruchez, A. et al. MHC class II transactivator CIITA induces cell resistance to Ebola virus and SARS-like coronaviruses. *Science* **370**, 241-247 (2020).
53. Tanevski, J. et al. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol.* **23**, 97 (2022).
54. Madissoon, E. et al. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat. Genet.* **55**, 66-77 (2023).
55. Xu, Y. et al. The effect of class II transactivator mutations on bleomycin-induced lung inflammation and fibrosis. *Am. J. Respir. Cell. Mol. Biol.* **44**, 898-905 (2011).
56. Bouhaddou, M. et al. The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **182**, 685-712.e19 (2020).
57. Wang, W. et al. Cognitive Impairment in the Post-Acute Phases of COVID-19 and Mechanisms: An Introduction and Narrative Review. *J. Alzheimers. Dis. Rep.* **8**, 647-658 (2024).

58. Bonhenry, D. et al. SARS-CoV-2 infection as a cause of neurodegeneration. *Lancet Neurol.* **23**, 562-563 (2024).
59. Merad, M. et al. The immunology and immunopathology of COVID-19. *Science* **375**, 1122-1127 (2022).
60. Tsukui, T. et al. Wolters, P.J. & Sheppard, D. Alveolar fibroblast lineage orchestrates lung inflammation and fibrosis. *Nature* **631**, 627–634 (2024).
61. Lam, M.T.Y. et al. Dynamic activity in cis-regulatory elements of leukocytes identifies transcription factor activation and stratifies COVID-19 severity in ICU patients. *Cell Rep. Med.* **4**, 100935 (2023).
62. Goh, W. et al. IKAROS and AIOLOS directly regulate AP-1 transcriptional complexes and are essential for NK cell development. *Nat. Immunol.* 25, 240-255 (2024).
63. Krämer, B. et al. Early IFN-α signatures and persistent dysfunction are distinguishing features of NK cells in severe COVID-19. *Immunity* **54**, 2650-2669.e14 (2021).
64. Fischer, D.S., Schaar, A.C. Theis, F.J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol.* **41**, 332-336 (2023).
65. Pham D, et al. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat Commun.* **14**, 7739 (2023).
66. Franzén, L. et al. Mapping spatially resolved transcriptomes in human and mouse pulmonary fibrosis. *Nat. Genet.* https://doi.org/10.1038/s41588-024-01819-2 (2024).
67. Almuntashiri, S. et al. TIMP-1 and its potential diagnostic and prognostic value in pulmonary diseases. *Chin. Med. J. Pulm. Crit. Care Med.* **1**, 67-76 (2023).
68. Sansores, R.H. et al. Prolonged-release pirfenidone in patients with pulmonary fibrosis as a phenotype of post-acute sequelae of COVID-19 pneumonia. Safety and efficacy. *Respir. Med.* **217**, 107362 (2023).
69. Landi, C. et al. Idiopathic Pulmonary Fibrosis Serum proteomic analysis before and after nintedanib therapy. *Sci. Rep.* **10**, 9378 (2020).
70. Umemura, Y. et al. Efficacy and safety of nintedanib for pulmonary fibrosis in severe pneumonia induced by COVID-19: An interventional study. *Int. J. Infect. Dis.* **108**, 454-460 (2021).
71. Diarimalala R.O. et al. Inflammasomes during SARS-CoV-2 infection and development of their corresponding inhibitors. *Front. Cell Infect. Microbiol.* **13**, 1218039 (2023).
72. Uttarkar, S. Frampton, J. & Klempnauer, K.H. Targeting the transcription factor Myb by small-molecule inhibitors. *Exp. Hematol.* **47**, 31-35 (2017).
73. Mellett, L. & Khader, S.A. S100A8/A9 in COVID-19 pathogenesis: Impact on clinical outcomes. *Cytokine Growth Factor Rev.* **63**, 90-97 (2022).
74. Malar, D.S. et al. Network analysis-guided drug repurposing strategies targeting LPAR receptor in the interplay of COVID, Alzheimer's, and diabetes. *Sci. Rep.* **14**, 4328 (2024).

Figure Legends

Fig.1: **Study design and spatial transcriptomics profiling of fatal COVID-19**. **A,** Overview of the study design, including sample processing workflow and spatial transcriptomics data analysis pipeline. **B,** UMAP representing sample integration and spot-wise most abundant cell type assignment. **C,** Mapping cell type deconvolution results on top of VisiumST brightfield images stained with H&E across disease progression. Illustrative examples of lung structures are shown matching cell types to expected structures. **D,** Bar plot showing cell type lineage abundance per condition. **E,** Immunohistochemistry staining with H&E and specific markers of the four main cell type lineages: CK7 (alveolar epithelial cells), CD34 (endothelial cells), CD68 (myeloid lineage) and trichrome (fibrosis). H&E: hematoxylin and eosin. DAD: diffuse alveolar damage.

Fig. 2: Cell type assignment and cellular compartments. **A,** Bar plot and stacked bar plot showing the proportion of the 45 cell types identified across disease progression. Significant values indicate credible differences between pairs of conditions (FDR < 0.05) denoted with a different colored star for each pairwise comparison. **B,** UMAP representation of representative cell types for the different conditions and mapping of cell type deconvolution results on VisiumST images across disease progression. An illustrative example showing cell type density matching expected lung structures is shown. **C,** Cellular compartments identified across disease progression are represented as factors on the x-axis. NMF factor loadings for cell types are represented as dot plot normalized per cell type, indicating the proportion of cells of each type present in each compartment. **D,** Cell type specific subgraph comparison using the portrait method across condition pairs. Dot size is indicative of the similarity score variance over samples. **E,** Filtration curves of three illustrative cell types. A filtration curve is plotted for every sample as well as the mean curve for every condition identified by the thicker and darker line. Large vertical steps towards the left of the plot indicate low density, whereas large vertical steps towards the right of the plot indicate high density. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered outliers. AT: alveolar type. Mph: macrophage. MT: metallothionein. DC: dendritic cell. EC: endothelial cell. NK: natural killer. SM: smooth muscle. TB: terminal bronchiole. DAD: diffuse alveolar damage.

Fig. 3: Intercellular and intracellular communication programs in fatal COVID-19. **A,** Heatmap representing average factor scores per lung tissue slide according to ligand-receptor interaction scores. Ward clustering method and Euclidean distance were used to perform hierarchical clustering. **B,** Pathway enrichment analysis of ligand-receptor loadings. Statistically significant enrichment scores (p-value < 0.05) are denoted with a star (*). **C,** Factor 1, Factor 2 and Factor 3 scores mapped in illustrative VisiumST samples across disease progression. **D,** Heatmap representing transcription factor activity enrichment score. Top 10 transcription factors are shown per condition. Enrichment scores were scaled. **E,** SMAD3, SMAD7 and ETV5 transcription factor enrichment scores in illustrative VisiumST samples across disease progression. **F,** Cell type abundance variance (R^2) explained by the joint ligand-receptor interactions and transcription factor activity predictive model. Additionally, the contribution of ligand-receptor interactions and transcription factor activity to the predictive performance is shown. Lastly, the top 10 predictors of peribronchial fibroblasts abundance with their corresponding importances as defined by the coefficients' t-values calculated by ordinary

least squares t-values in the predictive model are shown. **G**, Causal intracellular signaling network in peribronchial fibroblasts connecting deregulated intercellular ligand-receptor communication events with downstream transcription factors. Activatory events with upregulated proteins and transcription factors are colored in red while inhibitory events and downregulated proteins and transcription factors are colored in blue. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered outliers. AT: alveolar type. Mph: macrophage. MT: metallothionein. DC: dendritic cell. EC: endothelial cell. NK: natural killer. SM: smooth muscle. TB: terminal bronchiole. DAD: diffuse alveolar damage.

Fig. 4: Spatial local dependencies between ligand-receptor interactions and between cell types and transcription factors activity. A, Local bivariate similarity metrics score, including spatially-weighted global mean cosine similarity and bivariate global Moran´s R for the top 25 ligand-receptor loadings defining DAD associated Factor 3 in all studied samples. **B**, Mapping of Factor 3 scores, cosine similarity, permutation based p-values and local categories for TIMP1^CD63 interaction in illustrative VisiumST samples across disease progression. **C**, Local bivariate similarity metrics score, including spatially-weighted global mean cosine similarity and bivariate global Moran´s R for the spatial co-occurrence of the top 10 enriched transcription factors activity per condition and peribronchial fibroblasts abundance in all studied samples. **D**, Mapping of SMAD3 and SMAD7 activities local interactions with peribronchial fibroblasts abundance, including cosine similarity and local categories in selected illustrative VisiumST samples across disease progression. High-high interactions (red) and high-low or low-high interactions (blue) are depicted in local categories plots. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered outliers. PBFs: peribronchial fibroblasts. DAD: diffuse alveolar damage.

Fig. 5: Intercellular dependencies as a function of niche composition and spatio-temporal trajectories. A, Type coupling analysis with edge proportional to strength of directional dependencies by means of fold changes of differentially expressed genes for each pair of sender and receiver cell types across disease progression. Only those cell types with credible differential abundances and the most abundant cell types were considered. Only edges with at least 500 genes are shown. Results for intercellular

796 dependencies across disease progression are shown. **B,** Sender effect analysis of the
797 CD8 T cells – NK cells axis in control samples, AT1-CD4 T cells axis in acute DAD lungs,
798 and non-classical monocytes and CD4 T cell axis in proliferative DAD lungs. Shown is
799 the estimated fold change that the sender cell type on the y-axis induces in the gene on
800 the x-axis in receiving cells. Additionally, a sender similarity analysis based on a
801 correlation of the coefficient vectors of each sender type with respect to CD4 T cell
802 receivers across disease progression is shown. **C,** Spatio-temporal trajectory of AT2 to
803 AT1 cell type differentiation in a proliferative DAD lung tissue identified two clades of
804 transdifferentiating cells. Transition genes positively (blue) or negatively (red) correlated
805 with the predicted trajectory and extracted by Spearman correlation test with adjusted p-
806 value <0.05 and correlation coefficient > 0.4 or < -0.4 are shown for each transitioning
807 clade. A comparison between clade markers is additionally shown. Lastly, total and cell
808 type specific gene expression of CAV1 is shown for AT1, AT2 and AT0 cells. AT: alveolar
809 type. Mph: macrophage. MT: metallothionein. EC: endothelial cell. NK: natural killer. SM:
810 smooth muscle. DAD: diffuse alveolar damage.
811

812 **Methods**

813 **Patient inclusion criteria**

814 The inclusion criteria for the COVID-19 associated DAD cohort were: patients > 18
815 years old, with polymerase chain reaction (PCR) positive for SARS-CoV-2 and complete
816 clinical information of disease history, comorbidities and follow-up, showing clinical
817 pulmonary involvement and COVID-19-related death. The inclusion criteria for the
818 control cohort were: individuals > 18 years old with complete clinical information about
819 comorbidities, without clinical evidence of SARS-CoV-2 infection, and sudden death
820 due to cardiovascular disease, except one case died due to cancer dissemination, and
821 one sample from a normal lung biopsy was included.
822

823 **Generation of Visium spatial transcriptomics data from formalin fixed paraffin**
824 **embedded COVID-19 lung samples**

825 First, the RNA integrity of the FFPE samples was assessed by extracting RNA from
826 freshly collected tissue sections and evaluating the percentage of RNA fragments above
827 200 base pairs (DV200). Briefly, Tissue blocks were placed in the microtome
828 (ThermoScientific HM340E) and trimmed to expose the tissue. 4 sections 10 µm thick
829 were placed in a chilled Eppendorf tube and the RNA was extracted using a protocol

from Qiagen (Rneasy FFPE Kit 73504), following extraction, the product was analyzed by TapeStation. Samples with DV200 ≥ 22% were selected for experiments.

Selected samples were placed in the microtome and sectioned 7 µm thick, each section was then placed in a water bath floating at 42 °C, sections were collected and mounted onto a 6.5 × 6.5 mm capture area of the Visium Spatial Gene Expression slide (2000233, 10X Genomics). Capture areas contain approximately 5000 barcoded spots, providing an average resolution of 1–10 cells. After sectioning, the slides were dried at 42°C for 3 hours. The slides were then placed inside a slide mailer, sealed with parafilm, and left overnight at Room temperature.

At the next day, the slides were deparaffinized by successive immersions in xylene and ethanol followed by H&E staining according to Demonstrated Protocol (CG000409, 10X Genomics). Brightfield images were taken using a 10X objective (Plan APO) on a Nikon Eclipse Ti2, images were stitched together using NIS-Elements software (Nikon) and exported as tiff files. After imaging, the glycerol and cover glass were carefully removed from the Visium slides by holding the slides in an 800 ml water beaker and letting the glycerol diffuse until the cover glass detached and density changes were no longer visible in the water. The slides were then dried at 37°C and incubated for decrosslinking for 60 min.

Following decrosslinking step, over-night probe hybridization was performed, and libraries were prepared according to the Visium Spatial Gene Expression for FFPE User Guide (CG000407, 10X Genomics). Libraries were sent for sequencing in Macrogen Korea using 1 lane of HiSeq X 150PE (2x 150bp) per sample, applying 1% Phix. Sequencing was performed using the specific for FFPE following read protocol: read 1: 28 cycles; i7 index read: 10 cycles; i5 index read: 10 cycles; read 2: 50 cycles.

Immunohistochemistry analysis

FFPE tissue sections were analyzed using standard IHC techniques. The primary antibodies used were anti-CD34 (clone QBEnd 10, Agilent Technologies, Santa Clara, CA, USA), anti-CD68 (clone KP1, Agilent Technologies, Santa Clara, CA, USA) and anti-CK7 (clone OV-TL 12/30, Agilent Technologies, Santa Clara, CA, USA). Immunostaining was performed automatically using a DAKO Autostainer Link 48 machine (Agilent Technologies, Santa Clara, CA, USA). Anti-CD34 was positive in endothelial cells, anti-CD68 was expressed in the cytoplasm of intraalveolar macrophages and CK7 was used as a marker for pneumocytes. A trichrome stain was used to evaluate the amount of fibrosis.

Computational analysis

Visium spatial gene expression libraries mapping

Visium Spatial Gene Expression libraries for Formalin Fixed Paraffin Embedded (FFPE) tissue samples were analyzed with spaceranger count pipeline using Space Ranger version 2.0.0 from 10x Genomics. First, a manual fiducial alignment and tissue boundary identification, including manual selection of spots covering tissue regions, were performed for each single library FFPE sample using Loupe Browser version 6.0 on the brightfield image. A probe set reference file compatible with FFPE workflow and human reference genome GRCh38 were downloaded from 10x Genomics and used to map Visium gene expression libraries.

Visium ST data preprocessing

We performed quality control (QC) steps including filtering of low-quality spots defined by a low number of detected genes with positive counts, low number of counts (library size) and high proportion of mitochondrial counts. These metrics were computed using scanpy[75]. As QC automatic filtering threshold, we utilized median absolute deviations (MAD) to identify outliers, as defined by differences in 5 MADs for number of detected genes and library size and 3 MADs for mitochondrial counts (including mitochondrial counts exceeding 8%) per tissue slide[76].

We next applied normalization to the raw counts by scaling the counts followed by the shifted logarithm transformation to stabilize variance in gene expression between cells. To filter out uninformative genes with mostly zero counts, we performed feature selection using deviance to select informative genes[76] using scry R package and selecting the top 6,000 highly deviant genes, as inspired by the preprocessing workflow utilized by the Human Lung Cell Atlas[13].

Finally, we performed dimensionality reduction using principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) with scanpy default parameters to reduce data complexity and for visualization purposes. To identify cellular structure, we cluster cells applying the Leiden algorithm to the previously computed neighborhood graph with scanpy. Lastly, individual sample objects were joined into a single object using the anndata concat() function. After the concatenation, we re-normalized raw counts on the joined object using global scaling by the total counts per barcode and applying the shifted logarithm transformation, followed by feature selection using deviance to select the top 6,000 highly deviant genes, dimensionality reduction and clustering as previously

898 described. A total of 91,068 spots, including 77,580 spots from fatal COVID-19 samples
899 and 13,488 spots from control samples were profiled after QC.

900 **Human Lung Cell Atlas (HLCA) reference processing**

901 We leveraged the HLCA[13] single-cell RNA sequencing (scRNA-seq) reference dataset
902 and consensus cell type annotations for spatial mapping and annotation. To this end, we
903 downloaded and processed the HLCA core and data object, selected lung parenchyma
904 tissue cell types with at least 150 total cells at the finest level of annotation for a more
905 robust and reliable reference model training, including a total of 333,011 cells in the
906 filtered dataset. The mitochondrial genes were removed for spatial mapping. Marker
907 gene selection was performed for every cell type by ranking genes using scanpy tool
908 rank_genes_groups() function with default parameters using the t-test and computing a
909 hierarchical clustering based on gene expression values for visualization with scaled
910 expression for easier identification of differences.

911 **Spatial mapping of cell types with cell2location**

912 Both, our joined Visium ST and the filtered HLCA reference datasets, were subset to the
913 same gene set as baseline for the mapping between single cell and spatial data, using
914 default parameters to select a total number of 5,850 Ensembl gene identifiers. First, a
915 reference model was fitted to estimate the reference cell type signature derived from the
916 HLCA scRNA-seq data with cell2location[21] and using the finest level of cell type
917 annotation reported. Cell2location uses a Negative Binomial regression model to
918 estimate signatures, while accounting for batch effect and covariates. Hence, we
919 included the following variables from the HLCA core object as covariates in our model:
920 “assay”, “donor_id”, “tissue_sampling_method” and “tissue_dissociation_protocol”. To
921 train the regression model we used default parameters to perform training on all cells in
922 the dataset. A maximum number of 250 epochs were sufficient to achieve convergence.

923 For the subsequent spatial mapping, cell2location requires two user-provided
924 hyperparameters based on the tissue and experiment QC, including expected number of
925 cells per spot, that we set to 20, and regularization parameter of within slide or batch
926 variation in RNA detection sensitivity, set to 20 (default), as previously described for the
927 profiling of human lung tissue with Visium ST[54]. The model was trained using full data
928 until convergence with 40,000 iterations and loss function (ELBO) was used.
929 Reconstruction accuracy plots were inspected to assess model quality. Cell abundance
930 mapped to spatial coordinates was derived using the 5% quantile of the posterior
931 distribution. To ease visualization of cell type abundances, the most abundant cell type

per spot, including aggregation by cell type lineage, were represented. We constructed a K-nearest neighbor (KNN) graph (size of local neighborhood set to 8) using estimated cell abundances and applied Leiden clustering (resolution parameter set to 0.5) to jointly cluster all Visium spots, making cluster identities directly comparable. We used the KNN graph representing location composition similarity to build a joint integrated UMAP representation (minimum distance between embedded points set to 0.3 and spread value set to 1) of all VisiumST samples. Additionally, cell-type specific expression of every gene at every spatial location was computed and used as input for cell-cell communication analysis with NCEM and for inferring intracellular signaling networks in peribronchial fibroblasts.

Lastly, we used non-negative matrix factorization (NMF) on cell2location mapping results to identify spatial co-occurrence of cell types. NMF was trained for a range of factors, selecting 8 factors for cellular compartments identification and visualization per condition with NMF factor loadings being represented.

Differential analysis of cell populations

To evaluate how cell populations changer across the studied biological conditions, we used scCODA model[77] that employs a Bayesian model to perform compositional data analysis on the estimated cell-type abundances. The scCODA model determine statistically credible effects. We set the cutoff between credible and non-credible effects on a false discovery rate level (FDR) < 0.05. Estimated cell type abundances were used, including abundance aggregation by cell type lineage, and compositional data visualization was performed using stacked barplots and boxplots. To find a reference cell type that preserves changes in relative abundance across samples we used automatic reference cell type estimation, and Migratory dendritic cells (DCs) were deemed as reference category. Differences between conditions were computed using control samples as control group for control vs acute DAD and control vs proliferative DAD comparisons, whereas acute DAD samples were used as control group for acute DAD vs proliferative DAD comparisons.

Analysis of the spatial arrangement of cell types

A comparison of tissue architecture across conditions was performed leveraging novel statistical and computational approaches to compare cell spatial organization at the level of cell types and samples using GraphCompass[27]. Samples are modelled as graphs of cells and cell-type specific graphs between conditions were compared with distances being computed using the portrait method and cell-type specific similarity scores were

jointly visualized. Furthermore, comparisons between entire sample graphs for selected cell types and condition were performed using filtration curves as previously described[27]. For computing spatial graphs, we used default parameters for Visium ST samples and defined the cluster key as the spot-wise most abundant cell-type.

Cell-cell communication (CCC) analysis

To analyze intercellular communication events, we looked for potential ligand-receptor interactions (LRI) on our Visium ST slides using LIANA+[30,31]. To assess the spatial co-occurrence of LRIs we used spatially-informed local (individual spot-level) bivariate similarity metrics, including spatially-weighted Cosine similarity and local Moran´s R[31]. These metrics use weights based on the spatial connectivity between spots, defined as radial kernels using the inverse Euclidean distance. Since we are assessing LRIs, we only considered interactions in which ligands, receptors and their subunits were expressed in at least 5% of the spots.

Furthermore, local interactions were categorized according to their magnitude and sign, allowing identification of local categories, where interactions are further classified into high-high (both variables are highly expressed) or high-low (one variable is highly and the other lowly expressed) interactions. Additionally, we used spot label permutations (N=100) to generate a Null distribution and to compute empirical local p-values to assess statistical significance of local metrics.

In addition to the local bivariate scores and to obtain “global” summaries of the local interaction results, we obtained global scores for each pair of variables, including Global mean (average) Cosine similarity and Global Moran´s R[31], to identify pairs of variables that cluster together or apart and to select the best candidates for visualization and downstream analysis.

Beyond LRIs, we applied the same approach to assess spatial relationships between transcription factors (TFs) activity (see below) and cell type abundances. To make the distributions comparable, we z-scaled TF activities and cell type abundances.

To identify coordinated cell-cell communication signatures, we used NMF on the local LRI scores. The heuristic elbow procedure selected three factors as the optimal component number. Average NMF factor scores per tissue slide clustered samples according to disease status and were visualized using a heatmap representation and hierarchical clustering using Euclidean distances and Ward´s method.

Pathway enrichment analysis of LRI loadings was performed to biologically characterize the three identified factors. To this end, pathway annotations from the PROGENy resource[33] were converted into ligand-receptor sets as previously described[31], assigning LRIs to specific pathways. Next, we performed enrichment using multivariate linear regression with decoupleR[32]. For LRI analysis we used LIANA+ consensus resource.

Transcription factor activity analysis

To infer TF activity based on prior knowledge, we used CollecTRI resource[46] containing a curated collection of TFs and their transcriptional targets with interactions weighted by their regulation mode (activation or inhibition). To estimate TF enrichment scores, we run a multivariate regression model using decoupleR for each spot and each TF, and a linear model was fitted to predict gene expression based on the interaction weights. The obtained t-value of the slope is the score, indicating activation or inactivation of the TF if positive or negative, respectively.

TFs enriched in each condition were identified using decoupleR rank_sources_groups() function with a t-test that overestimates variance of each group. The top 10 TFs per condition were extracted and represented in a heatmap using scaled TF enrichment scores between 0 and 1, meaning for each TF, subtract the minimum and divide each by its maximum. Moreover, TF enrichment scores without scaling were represented on top of illustrative Visium ST slides.

Learning spatial relationships between LRIs and TFs activity with cell type abundance

To learn spatial dependencies between local LRIs, TFs activity and cell types abundance, we used MISTy[53], an explainable multi-modelling approach, as previously described[31]. We selected the top 25 local ligand-receptor loadings from Factor 3 and the enrichment scores of the top 10 TFs per condition, to jointly modelled in a spatially informed manner the estimated cell type abundances in each spot. We specifically utilize a linear model, since the coefficients' t-values (predictor importances), as calculated by ordinary least squares (OLS), are signed and comparable. Importantly, we bypassed predicting the intraview (intra spot), and for each target (cell type) we assessed not only the predictor importance, but how well LRIs and TFs explained cell type abundance using the joined multi-view R^2 (goodness of fit) and evaluated their relative contribution to the joint predictive performance.

Intracellular signaling network in peribronchial fibroblasts

We inferred intracellular signaling networks from prior knowledge by linking identified LRIs and TF activity scores. LIANA+ approach[31] considers the direction of deregulation, including activation and inhibition of receptors and TFs, and the sign and direction of edges (activating or inhibiting) from prior knowledge of protein-protein interactions and of TFs and their targets obtained from Omnipath[78]. Using CORNETO, a putative causal network connecting LRIs to TFs was inferred for peribronchial fibroblasts. Hence, we used the coefficients´ t-values of the predictive linear model for peribronchial fibroblasts abundance as input (LRIs) and output (TFs) nodes of the intracellular signaling network. To this end, we computed the median of the t-values across samples and ranked them by absolute median value. For LRIs, the receptors were selected and when the same receptor was involved in multiple LRIs, the largest absolute median t-value was kept for that receptor and used as input node. Additionally, we obtained a prior knowledge network (PKN) with signed protein-protein interactions from Omnipath and used peribronchial fibroblasts specific gene expression computed with cell2location to calculate gene expression proportions within our target cell type, using those to generate weights for the nodes in the PKN.

Analysis of intercellular dependencies as a function of spot composition

Since cell-cell communications events are not limited to LRIs, we further assessed spot/niche composition effects on all genes using NCEM[64]. To this end, we utilized cell-type specific gene expression computed with cell2location. Hence, NCEM models expression variation within cell types across spots as function of the inferred spot composition. To focus the analysis on biologically relevant genes, we selected gene sets described in the WikiPathways database from the Molecular Signature Database (MSigDB) using decoupleR. We filtered out cell type specific marker genes computed using rank_genes_groups_df() function in scanpy using adjusted p-value < 0.05 and minimum log fold change > 2. Finally, 1614 genes were shared within our dataset. Additionally, 22 cell types were considered for NCEM analysis, including cell types with credible differential abundances computed by scCODA and the most abundant cell types.

Type coupling analysis was performed on the filtered dataset to compute sender and receiver effects based on a Wald test on the parameters learned by the linear NCEM model, using the full dataset and optimized with OLS. We further dissected these couplings based on gene-wise effects of particular interactions, including effects of all senders on one receiver (receiver effect analysis), reporting dependencies with at least 500 differentially expressed with q-value < 0.05 and absolute log fold change > 0.8.

Finally, a sender similarity analysis was performed to characterize sender profiles on CD4 T cells across conditions.

Spatio-temporal trajectories analysis

To characterize AT2-AT1 differentiation process, we leveraged stLearn tool[65] that employs a spatial graph-based method named pseudo-time-space (PSTS) that combines spatial and imaging information with gene expression to map spatial changes cell states, modelling and reconstructing their spatio-temporal trajectories. To accurately identify transition genes positively or negatively correlated with the predicted trajectory for AT2-AT1 differentiation, we reported genes with a Spearman correlation <-0.4 or > 0.4. Furthermore, we only selected spots enriched in AT1 and AT2 cells, where these cell types represented at least 70% of the maximum abundance of the inferred spot composition. Additionally, gene set enrichment analysis on the identified transition genes using cell type signature gene sets in the MSigDB was performed and the false discovery rate (FDR) q-values derived from the hypergeometric test were used.

References

75. Wolf, F.A., Angerer, P., Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

76. Heumos L, et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet.* **24**, 550-572 (2023).

77. Büttner, M. et al. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun.* **12**, 6876 (2021).

78. Türei, D. et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Sys.t Biol.* **17**, e9923 (2021).

Figure 1

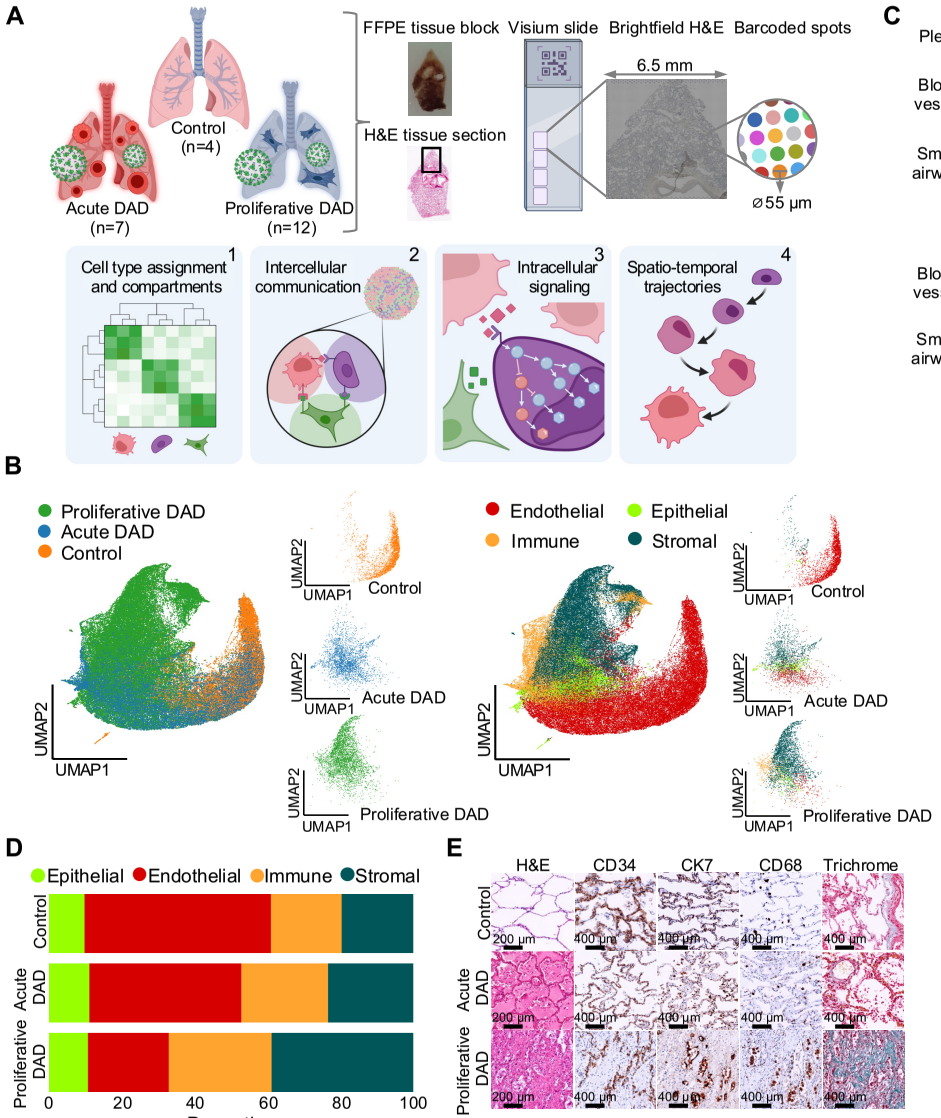

A
Control (n=4)
Acute DAD (n=7)
Proliferative DAD (n=12)
FFPE tissue block
H&E tissue section
Visium slide
Brightfield H&E
Barcoded spots
6.5 mm
⌀55 µm
1 Cell type assignment and compartments
2 Intercellular communication
3 Intracellular signaling
4 Spatio-temporal trajectories
B
Proliferative DAD
Acute DAD
Control
Endothelial
Epithelial
Immune
Stromal
UMAP1
UMAP2
D
Epithelial
Endothelial
Immune
Stromal
Control
Acute DAD
Proliferative DAD
0
20
40
60
80
100
Proportion
E
H&E
CD34
CK7
CD68
Trichrome
C
Control
Pleura
Blood vessels
Small airways
Acute DAD
Bronchial cells
Bronchus
Proliferative DAD
Blood vessel
Pneumocytes hyperplasia

Figure 2

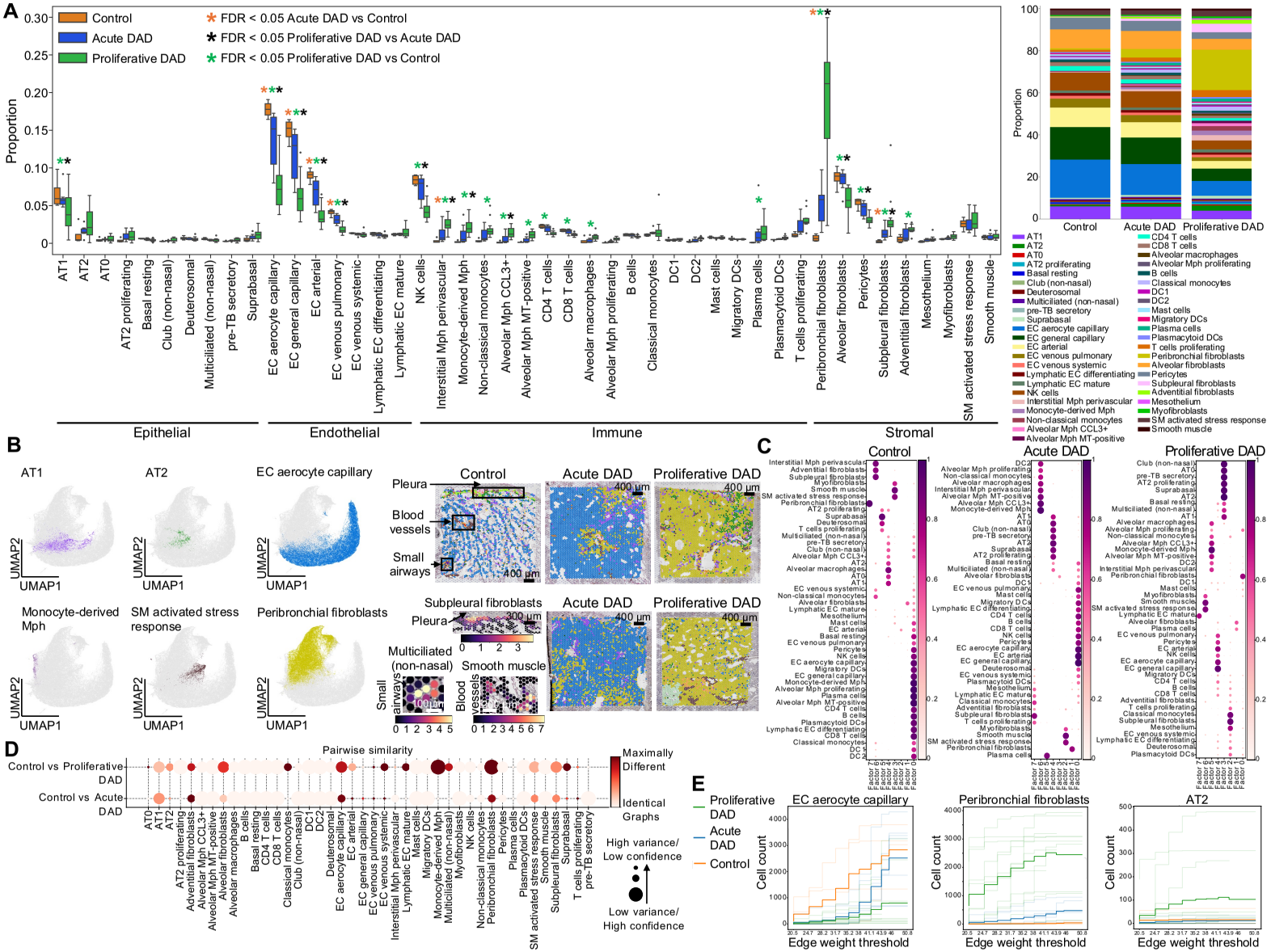

A
Control
Acute DAD
Proliferative DAD
* FDR < 0.05 Acute DAD vs Control
* FDR < 0.05 Proliferative DAD vs Acute DAD
* FDR < 0.05 Proliferative DAD vs Control
Proportion
AT1
AT2
AT0
AT2 proliferating
Basal resting
Club (non-nasal)
Deuterosomal
Multiciliated (non-nasal)
pre-TB secretory
Suprabasal
EC aerocyte capillary
EC general capillary
EC arterial
EC venous pulmonary
EC venous systemic
Lymphatic EC differentiating
Lymphatic EC mature
NK cells
Interstitial Mph perivascular
Monocyte-derived Mph
Non-classical monocytes
Alveolar Mph CCL3+
Alveolar Mph MT-positive
CD4 T cells
CD8 T cells
Alveolar macrophages
Alveolar Mph proliferating
B cells
Classical monocytes
DC1
DC2
Mast cells
Migratory DCs
Plasma cells
Plasmacytoid DCs
T cells proliferating
Peribronchial fibroblasts
Alveolar fibroblasts
Pericytes
Subpleural fibroblasts
Adventitial fibroblasts
Mesothelium
Myofibroblasts
SM activated stress response
Smooth muscle
Epithelial
Endothelial
Immune
Stromal
B
AT1
AT2
EC aerocyte capillary
Monocyte-derived Mph
SM activated stress response
Peribronchial fibroblasts
UMAP1
UMAP2
Control
Acute DAD
Proliferative DAD
Pleura
Blood vessels
Small airways
400 μm
Subpleural fibroblasts
Multiciliated (non-nasal)
Smooth muscle
100 μm
C
Control
Acute DAD
Proliferative DAD
D
Pairwise similarity
Control vs Proliferative DAD
Control vs Acute DAD
Maximally Different
Identical Graphs
High variance/ Low confidence
Low variance/ High confidence
E
Proliferative DAD
Acute DAD
Control
EC aerocyte capillary
Peribronchial fibroblasts
AT2
Cell count
Edge weight threshold

Figure 3

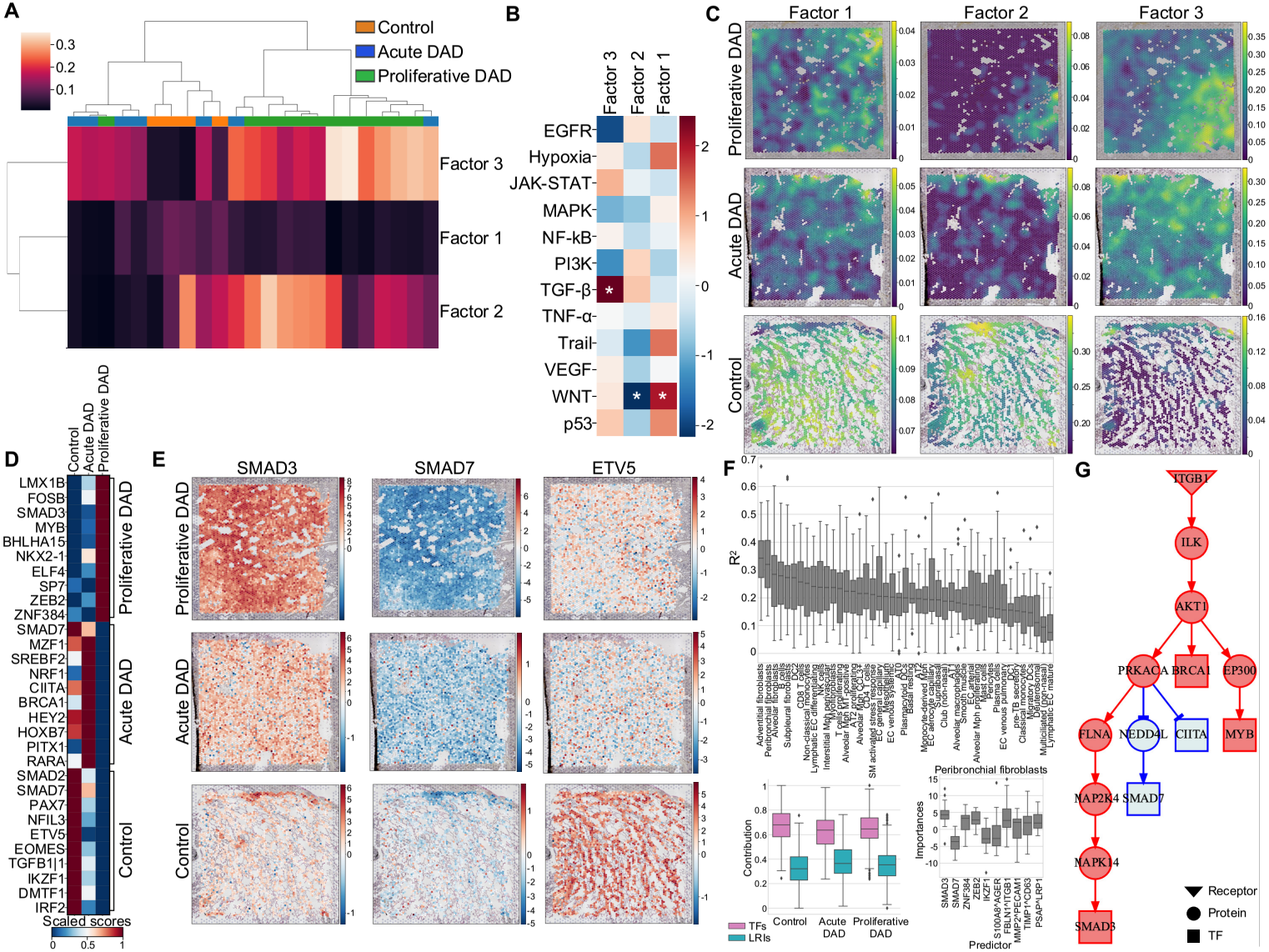

A
Control
Acute DAD
Proliferative DAD
Factor 3
Factor 1
Factor 2
B
Factor 3
Factor 2
Factor 1
EGFR
Hypoxia
JAK-STAT
MAPK
NF-kB
PI3K
TGF-β
TNF-α
Trail
VEGF
WNT
p53
C
Factor 1
Factor 2
Factor 3
Proliferative DAD
Acute DAD
Control
D
Control
Acute DAD
Proliferative DAD
LMX1B
FOSB
SMAD3
MYB
BHLHA15
NKX2-1
ELF4
SP7
ZEB2
ZNF384
SMAD7
MZF1
SREBF2
NRF1
CIITA
BRCA1
HEY2
HOXB7
PITX1
RARA
SMAD2
SMAD7
PAX7
NFIL3
ETV5
EOMES
TGFB1|1
IKZF1
DMTF1
IRF2
Scaled scores
E
SMAD3
SMAD7
ETV5
Proliferative DAD
Acute DAD
Control
F
R²
Peribronchial fibroblasts
Contribution
Importances
TFs
LRIs
Control
Acute DAD
Proliferative DAD
Predictor
G
ITGB1
ILK
AKT1
PRKACA
BRCA1
EP300
FLNA
NEDD4L
CIITA
MYB
MAP2K4
SMAD7
MAPK14
SMAD3
Receptor
Protein
TF

Figure 4

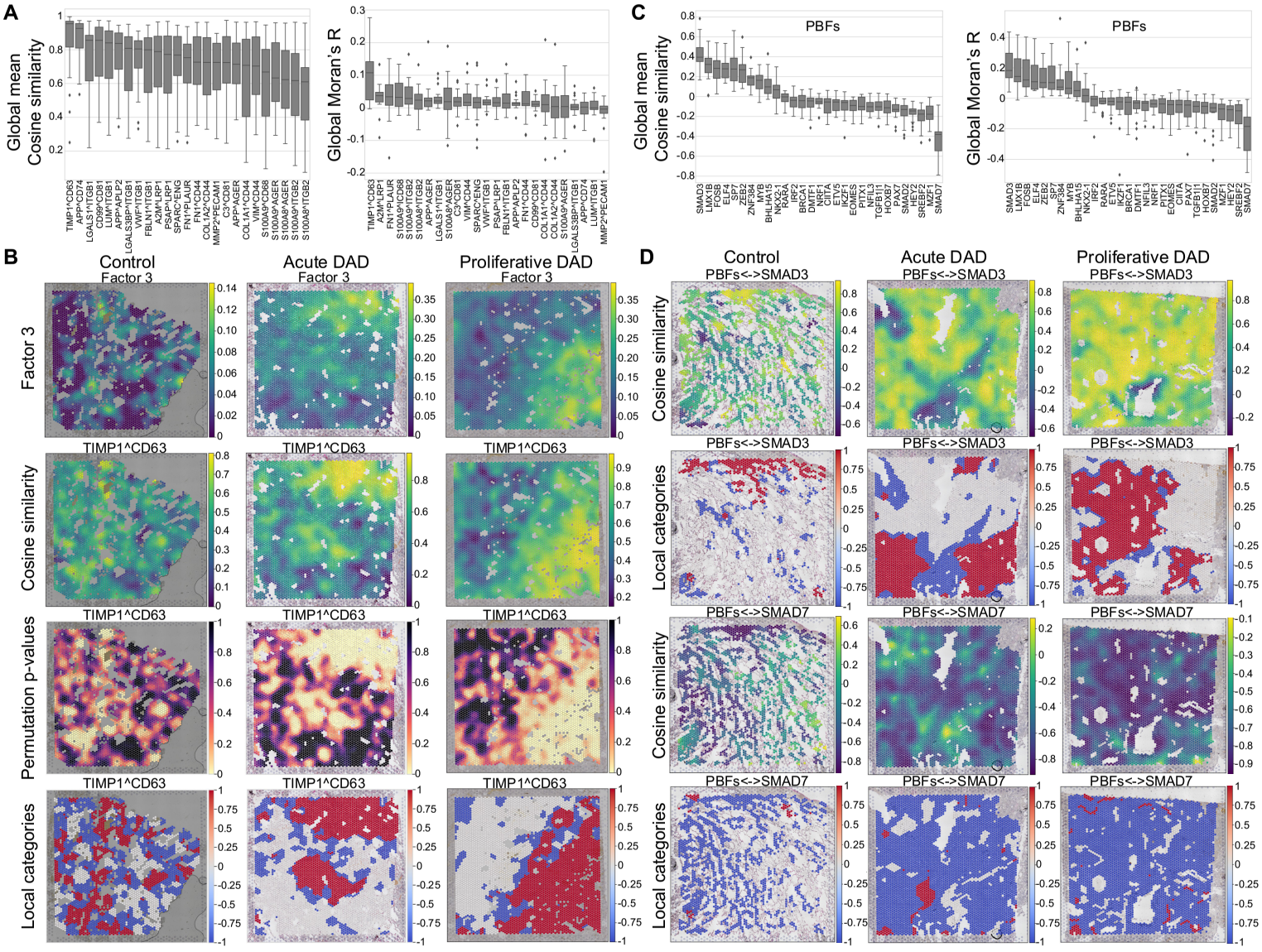

A
Global mean Cosine similarity
Global Moran's R
C
PBFs
Global mean Cosine similarity
PBFs
Global Moran's R
B
Control
Acute DAD
Proliferative DAD
Factor 3
Cosine similarity
Permutation p-values
Local categories
TIMP1^CD63
D
Control
Acute DAD
Proliferative DAD
Cosine similarity
Local categories
PBFs<->SMAD3
PBFs<->SMAD7

Figure 5

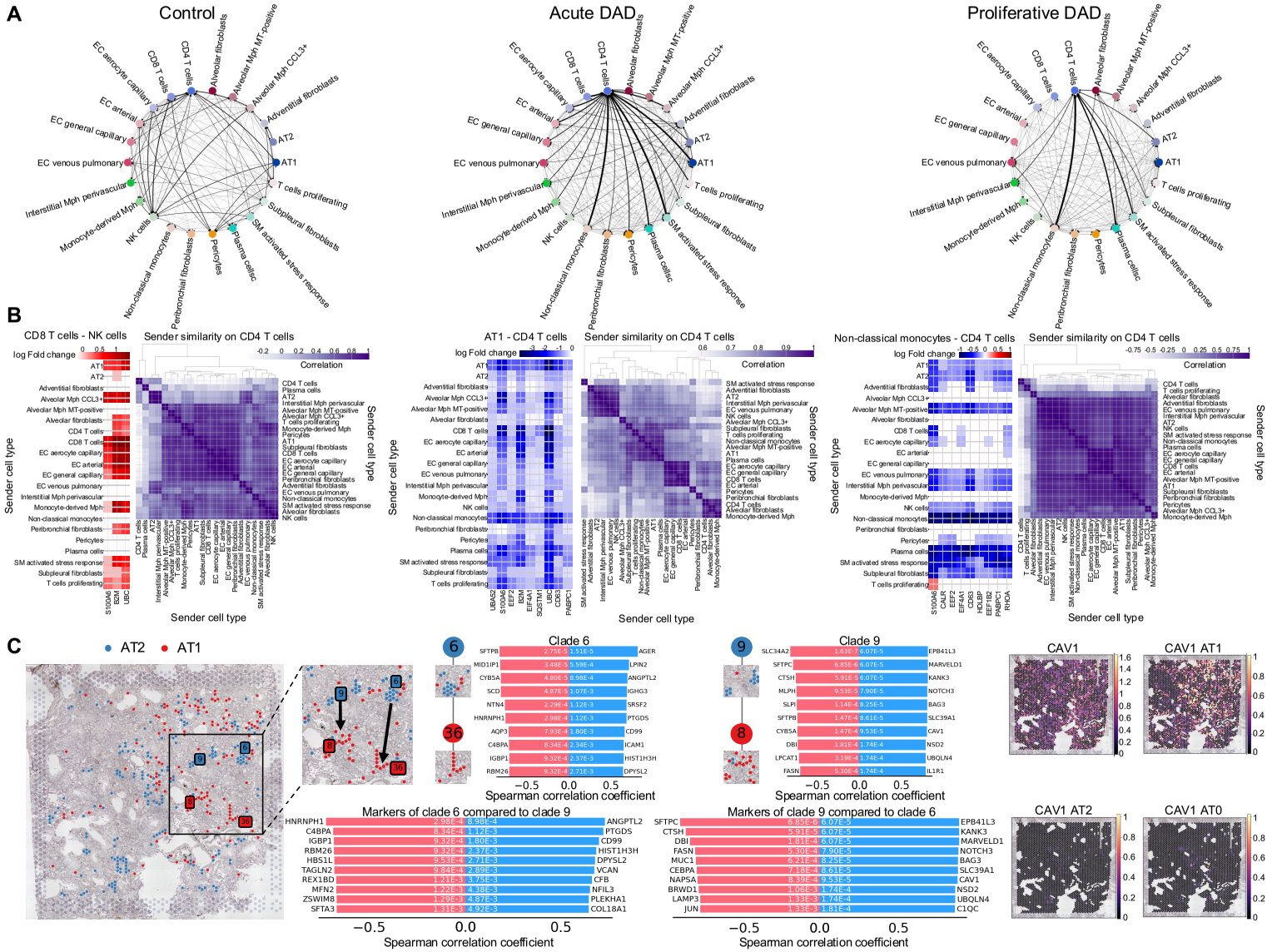

A
Control
Acute DAD
Proliferative DAD
B
CD8 T cells - NK cells
Sender similarity on CD4 T cells
AT1 - CD4 T cells
Non-classical monocytes - CD4 T cells
log Fold change
Correlation
Sender cell type
C
AT2
AT1
Clade 6
Clade 9
Spearman correlation coefficient
Markers of clade 6 compared to clade 9
Markers of clade 9 compared to clade 6
CAV1
CAV1 AT1
CAV1 AT2
CAV1 AT0

Table 1. Clinicopathological characteristics of the studied COVID-19 associated DAD patients and control group.

Characteristics	COVID-19 DAD cohort (N = 19)[a]	Control cohort (N = 4)[b]	p-value*
Gender - Frequency (%)			
Female	5 (26.3)	2 (50.0)	p=0.557
Male	14 (73.7)	2 (50.0)	
Age (years) - Median [range]	68.0 [52 - 91]	65.5 [39 - 72]	p=0.409
Underlying conditions - Frequency (%)			
Smoking	5 (38.5)	1 (25.0)	p=1.000
Hypertension	12 (63.2)	2 (50.0)	p=1.000
Diabetes mellitus	3 (15.8)	1 (25.0)	p=1.000
Obesity	9 (52.9)	1 (25.0)	p=0.586
Respiratory disease[c]	5 (26.3)	0 (0.00)	p=0.539
Cardiac disease[d]	13 (68.4)	0 (0.00)	p=0.024
Chronic kidney disease	2 (10.5)	0 (0.00)	p=1.000
Chronic neurological or neuromuscular disease	5 (26.3)	1 (25.0)	p=1.000
Cancer	7 (36.8)	2 (50.0)	p=1.000
Immunocompromised state[e]	3 (15.8)	2 (50.0)	p=0.194
Number of comorbidities - Frequency (%)			
0-2	6 (31.6)	2 (50.0)	p=0.589
3-6	13 (68.4)	2 (50.0)	
Pulmonary disease - Frequency (%)			
Acute DAD	7 (36.8)	NA	
Proliferative DAD	12 (63.2)	NA	
Cause of death - Frequency (%)			
Multiorgan failure/Respiratory distress	14 (73.7)	0 (0)	p=0.014
Pancreatitis	2 (10.5)	0 (0)	p=1.000
Intestinal necrosis	1 (5.3)	0 (0)	p=1.000
Cardiopathy[f]	1 (5.3)	1 (25.0)	p=0.324
Pyelonephritis	1 (5.3)	0 (0)	p=1.000
Cancer dissemination	0 (0)	1 (25.0)	p=0.174
Acute pulmonary embolism	0 (0)	1 (25.0)	p=0.174
Alive[g]	0 (0)	1 (25.0)	NA

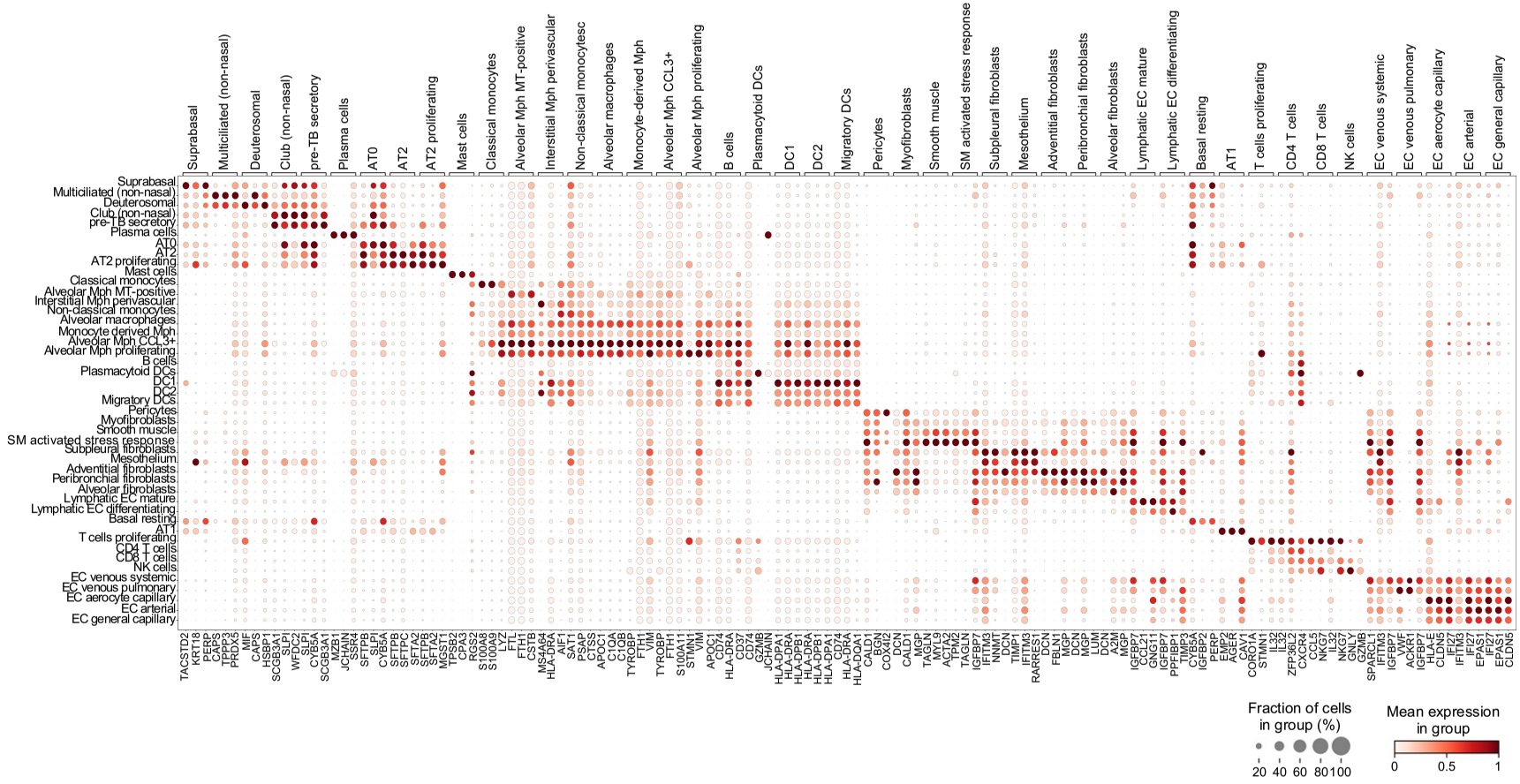
Abbreviations: DAD, diffuse alveolar damage; NA, Not applicable. [a]Inclusion criteria for the COVID-19 cohort: Patients >18 years old, PCR positive for SARS-CoV-2 with complete clinical information of disease history, comorbidities and follow-up, showing clinical pulmonary involvement and COVID-19-related death; [b]Inclusion criteria for the control cohort: >18 years old individuals with complete clinical information about comorbidities, without clinical evidence of SARS-CoV-2 infection, and sudden death due to cardiovascular disease, except one case died due to cancer dissemination and one post-surgery; [c]Chronic obstructive pulmonary disease or asthma; [d]Coronary artery disease, heart failure or atrial fibrillation; [e]Immunocompromised state due to autoimmune disease or other cause; [f]Acute myocardial infarction, cardiac fibrosis, aortic dissection, hemopericardium or myocarditis; [g]A normal lung biopsy was included in the Control group.*p-values were calculated using Fisher's exact test or Mann-Whitney test for dichotomous or continuous variables, respectively. p-values under 0.05 represent statistical significant association between co-variables.

Supplementary materials

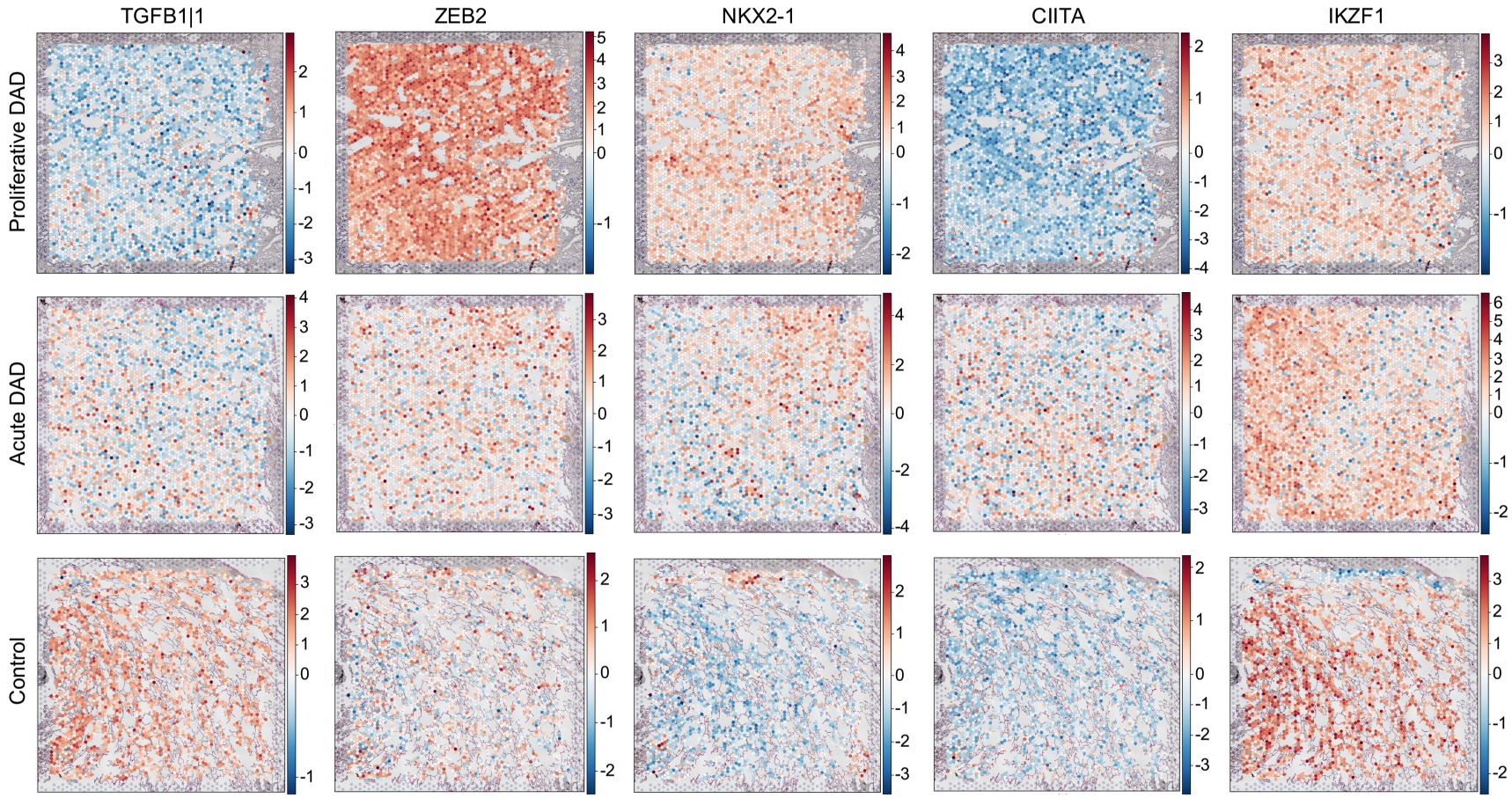
Table S2: Top 10 transcription factor enrichment scores per condition.

group	reference	names	statistic	mean change	pvals	pvals_adj
Control	rest	SMAD2	149.287	2.111	<0.001	<0.001
Control	rest	SMAD7	129.858	2.454	<0.001	<0.001
Control	rest	PAX7	122.506	1.602	<0.001	<0.001
Control	rest	NFIL3	115.582	1.321	<0.001	<0.001
Control	rest	ETV5	102.882	1.549	<0.001	<0.001
Control	rest	EOMES	99.829	1.109	<0.001	<0.001
Control	rest	TGFB1I1	95.669	1.139	<0.001	<0.001
Control	rest	IKZF1	95.512	1.192	<0.001	<0.001
Control	rest	DMTF1	92.678	1.098	<0.001	<0.001
Control	rest	IRF2	88.707	1.028	<0.001	<0.001
Acute	rest	SMAD7	94.634	1.424	<0.001	<0.001
Acute	rest	MZF1	92.890	0.727	<0.001	<0.001
Acute	rest	SREBF2	87.107	0.759	<0.001	<0.001
Acute	rest	NRF1	83.497	0.664	<0.001	<0.001
Acute	rest	CIITA	82.463	0.765	<0.001	<0.001
Acute	rest	BRCA1	76.606	0.682	<0.001	<0.001
Acute	rest	HEY2	75.841	0.636	<0.001	<0.001
Acute	rest	HOXB7	75.729	0.591	<0.001	<0.001
Acute	rest	PITX1	74.006	0.564	<0.001	<0.001
Acute	rest	RARA	72.190	0.600	<0.001	<0.001
Proliferative	rest	LMX1B	246.008	2.462	<0.001	<0.001
Proliferative	rest	FOSB	232.635	2.155	<0.001	<0.001
Proliferative	rest	SMAD3	193.746	1.922	<0.001	<0.001
Proliferative	rest	MYB	188.634	1.423	<0.001	<0.001
Proliferative	rest	BHLHA15	186.785	1.007	<0.001	<0.001
Proliferative	rest	NKX2-1	179.464	1.408	<0.001	<0.001
Proliferative	rest	ELF4	177.739	1.375	<0.001	<0.001
Proliferative	rest	SP7	171.968	1.175	<0.001	<0.001
Proliferative	rest	ZEB2	161.070	1.062	<0.001	<0.001
Proliferative	rest	ZNF384	160.747	0.985	<0.001	<0.001

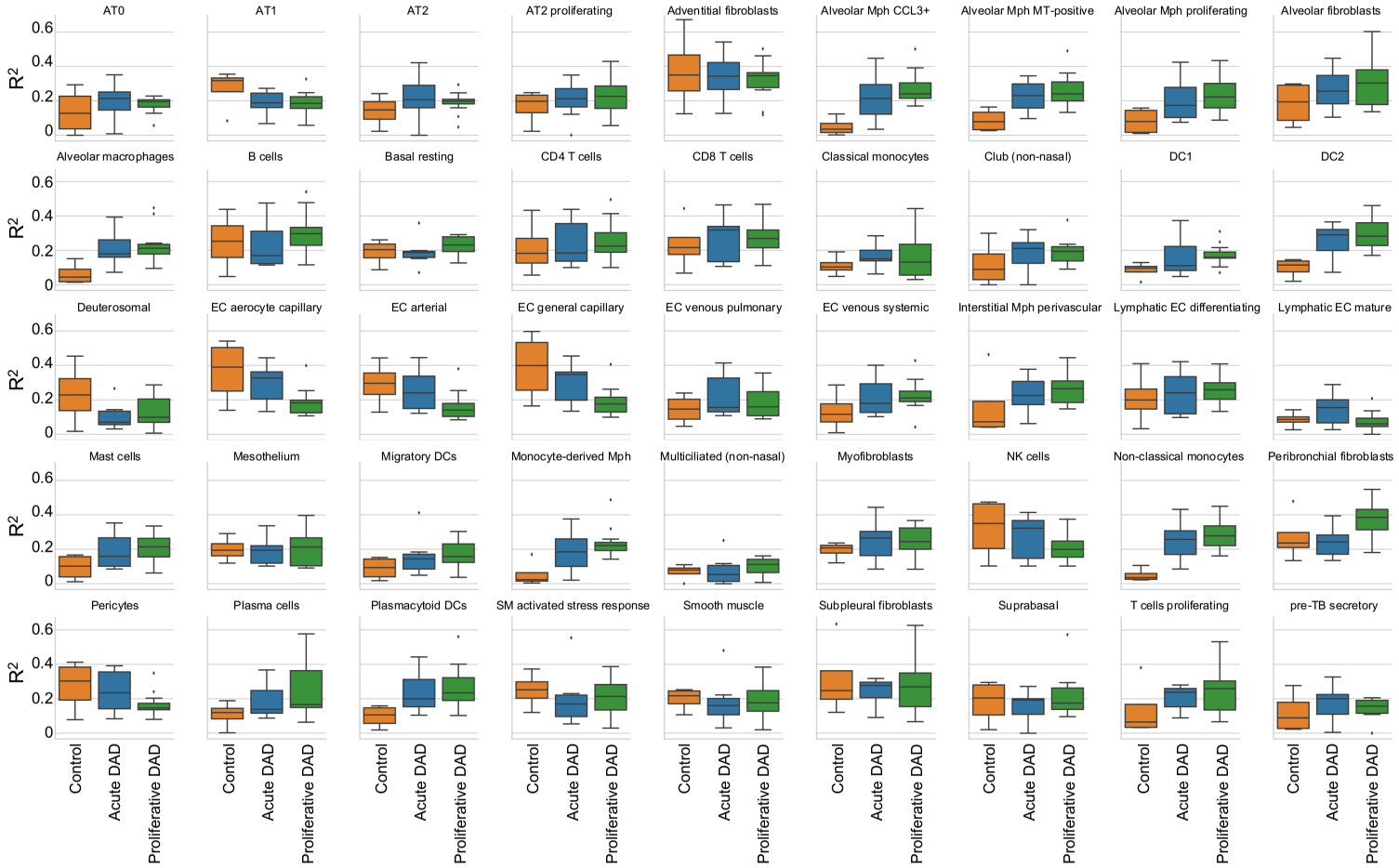
Supplementary Table S1 can be accessed on *bioRxiv* at:
https://doi.org/10.1101/2024.07.03.601404



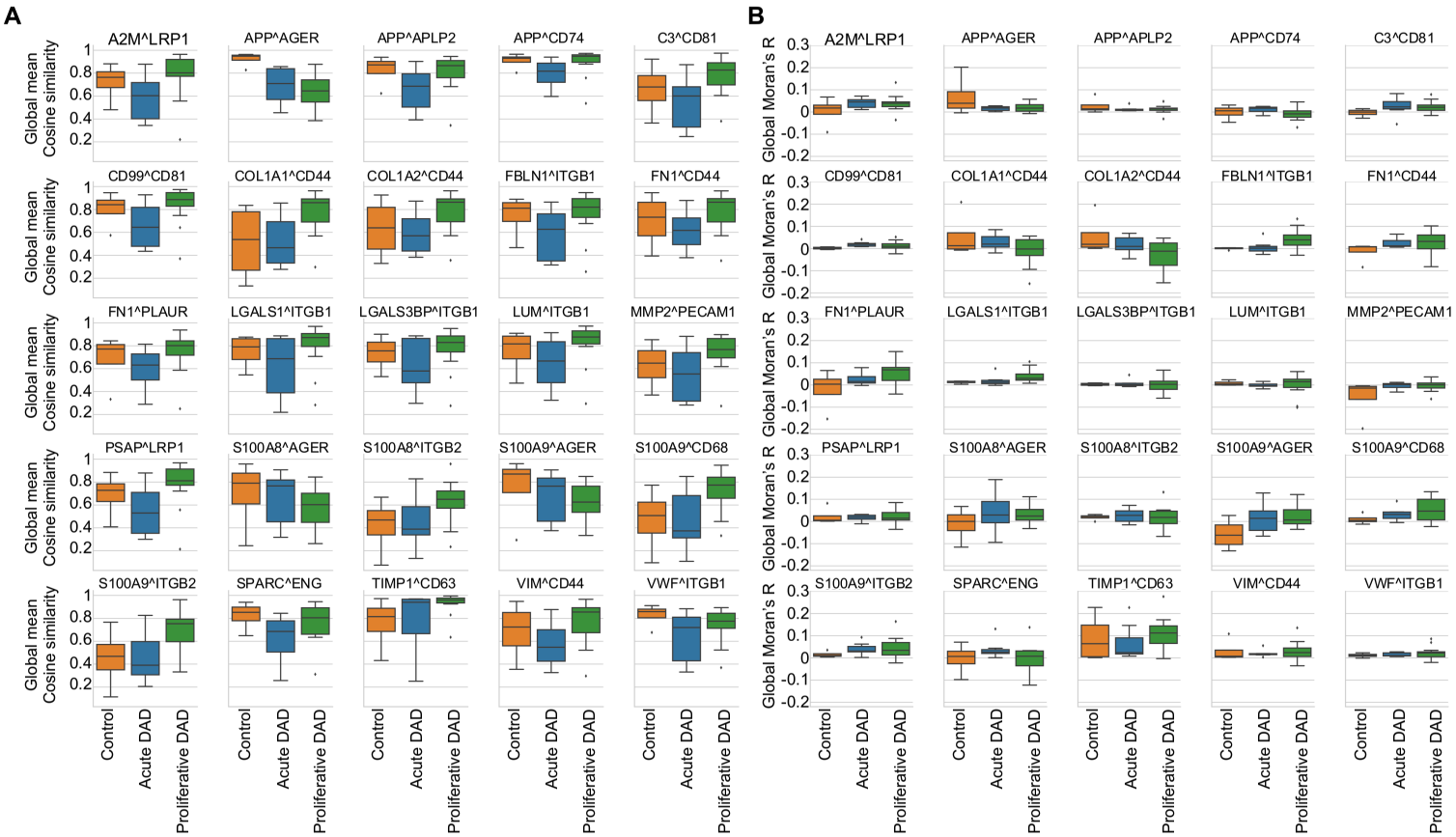
Supplementary Figure 1: Marker gene expression of reference signatures derived from the HLCA core for all 45 cell types identified. Gene expression was normalized such that the maximum group expression of cells for each marker was set to 1. Dot size indicates the fraction of cells in group. AT: alveolar type. Mph: macrophage. MT: metallothionein. DC: dendritic cell. EC: endothelial cell. NK: natural killer. SM: smooth muscle. TB: terminal bronchiole.



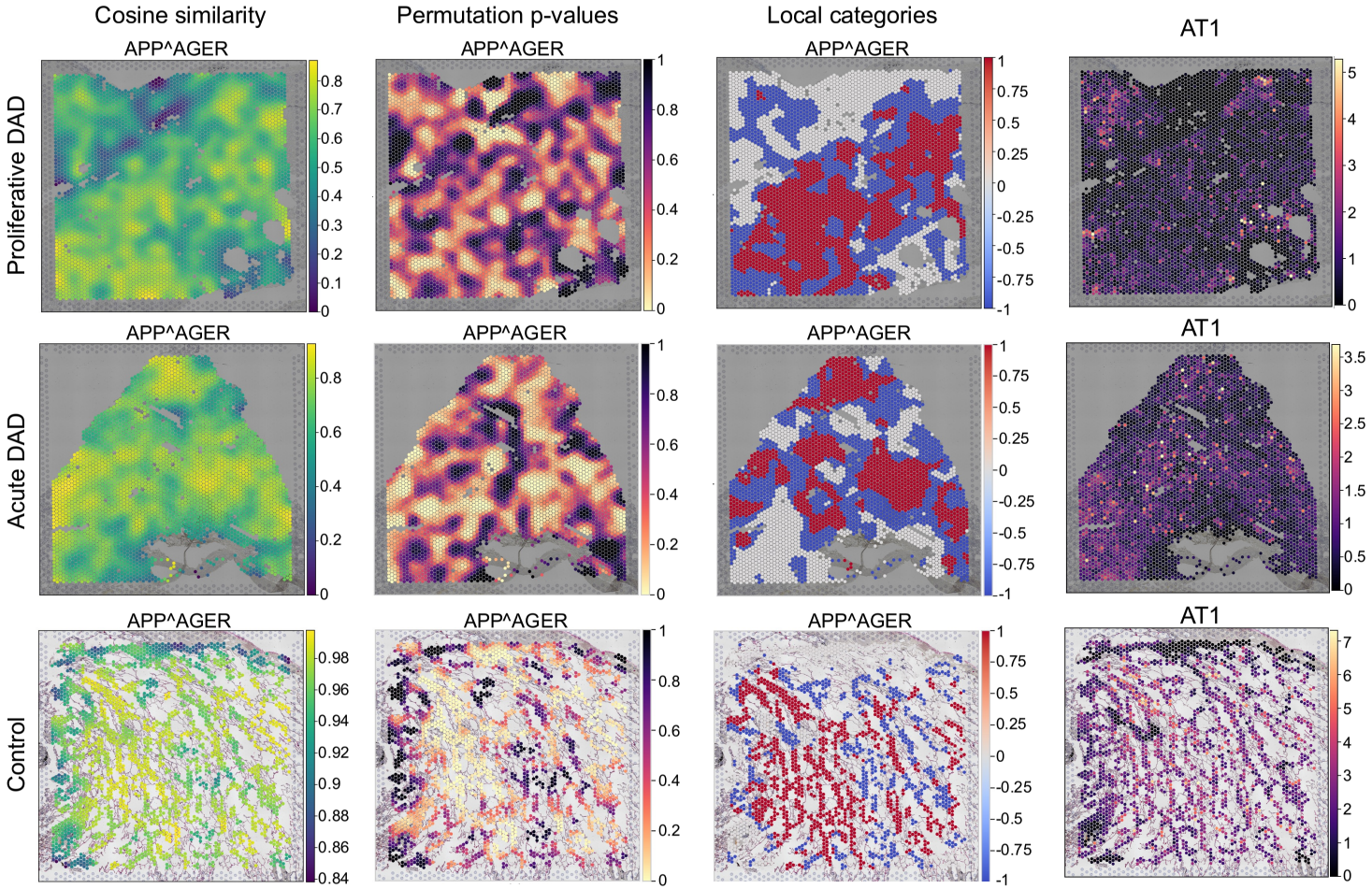
Supplementary Figure 2: Transcription factors activity across disease progression. Enrichment scores of selected transcription factors across disease progression in illustrative Visium ST lung samples are shown. DAD: diffuse alveolar damage.



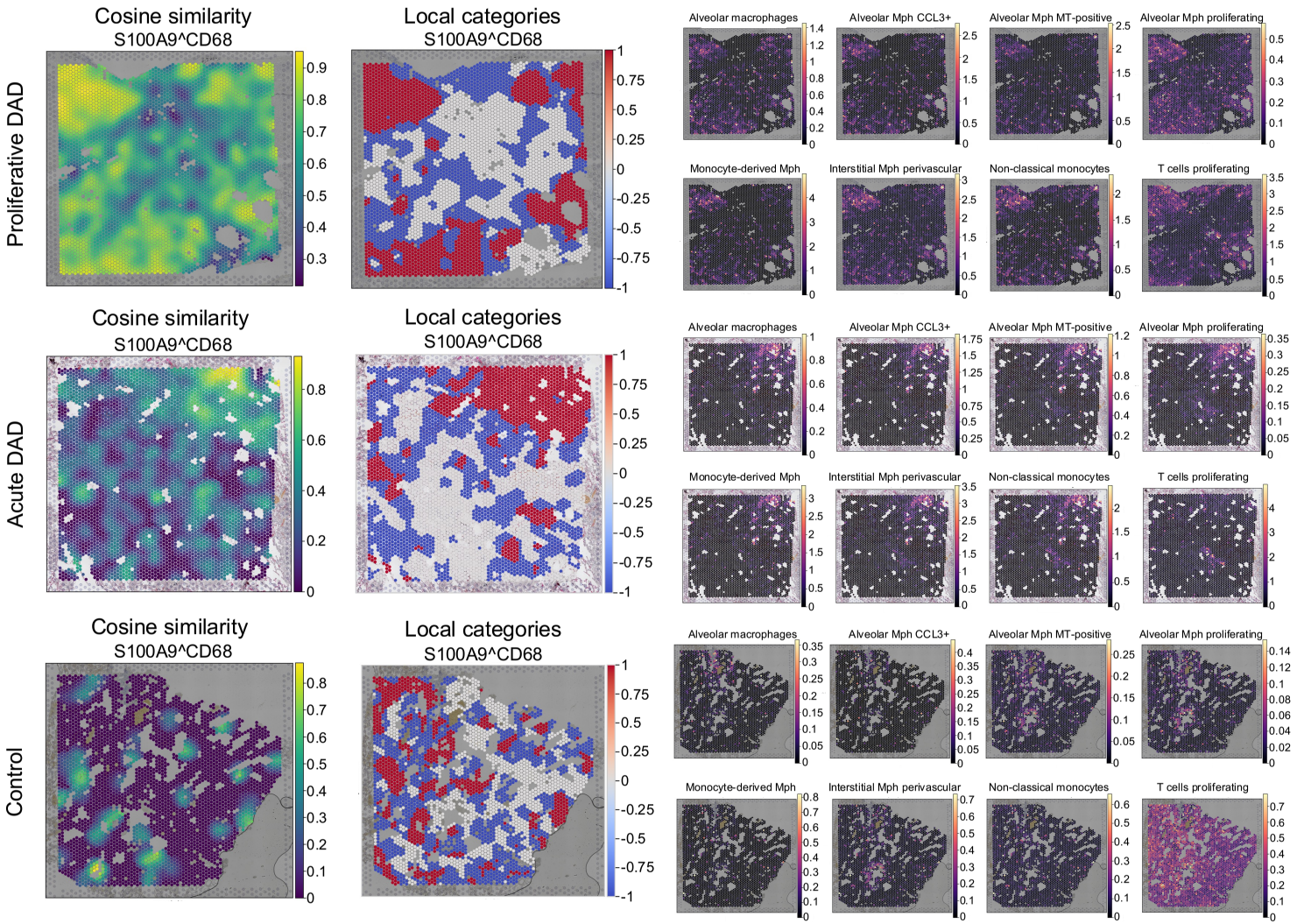
Supplementary Figure 3: Variance of cell type abundance explained by the predictive model. The joint ligand-receptor interactions and transcription factors activity multi-view R^2, contributing to explain cell type abundance variance across disease porgression is shown. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered to be outliers. AT: alveolar type. Mph: macrophage. MT: metallothionein. DC: dendritic cell. EC: endothelial cell. NK: natural killer. SM: smooth muscle. TB: terminal bronchiole. DAD: diffuse alveolar damage.



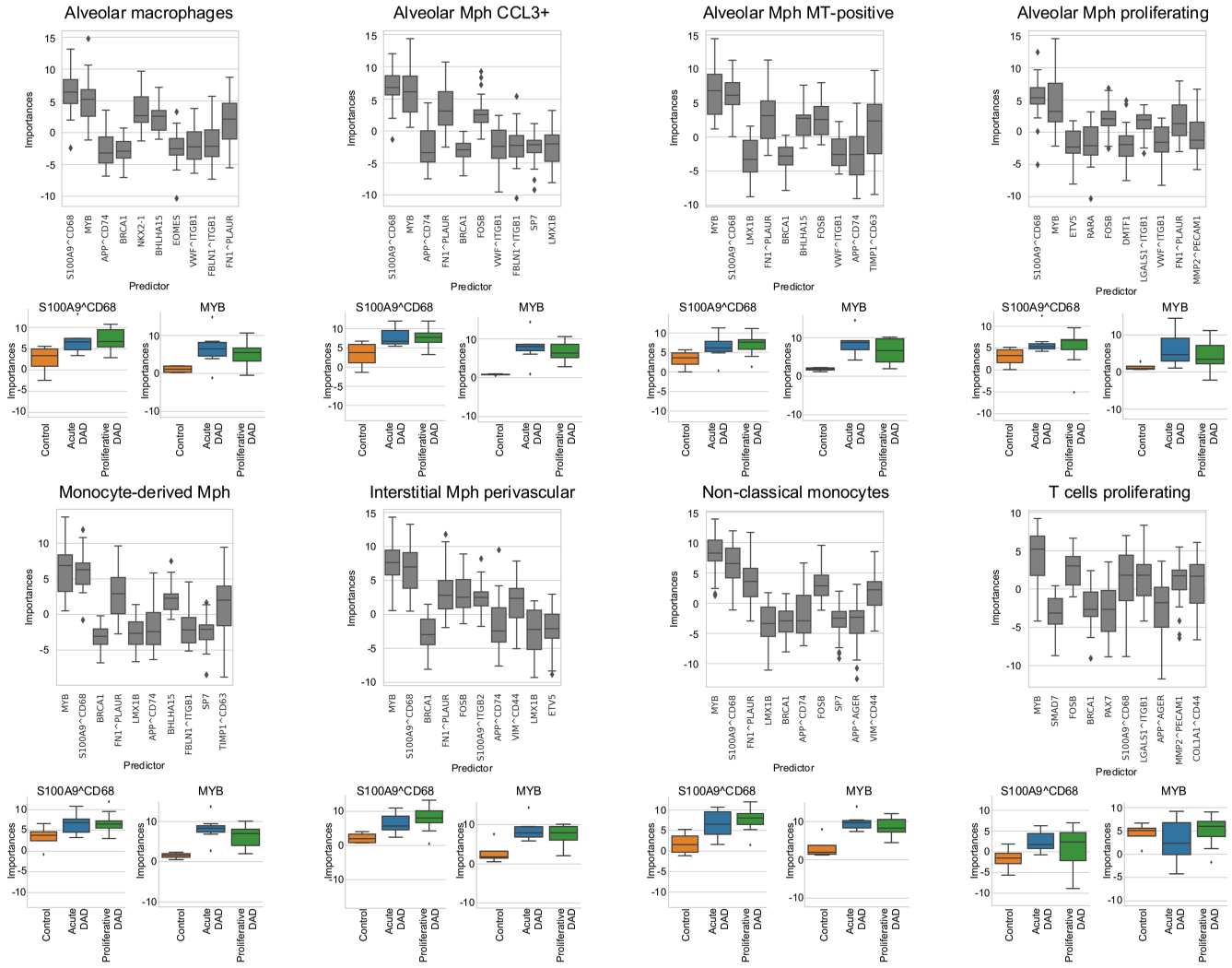
Supplementary Figure 4: Local spatial relationships of ligand-receptor interactions across disease progression. A) Global mean Cosine similarity and **B)** Global Moran's R for the top 25 ligand-receptor interactions defining DAD associated Factor 3 across disease progression are shown. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered to be outliers. DAD: diffuse alveolar damage.



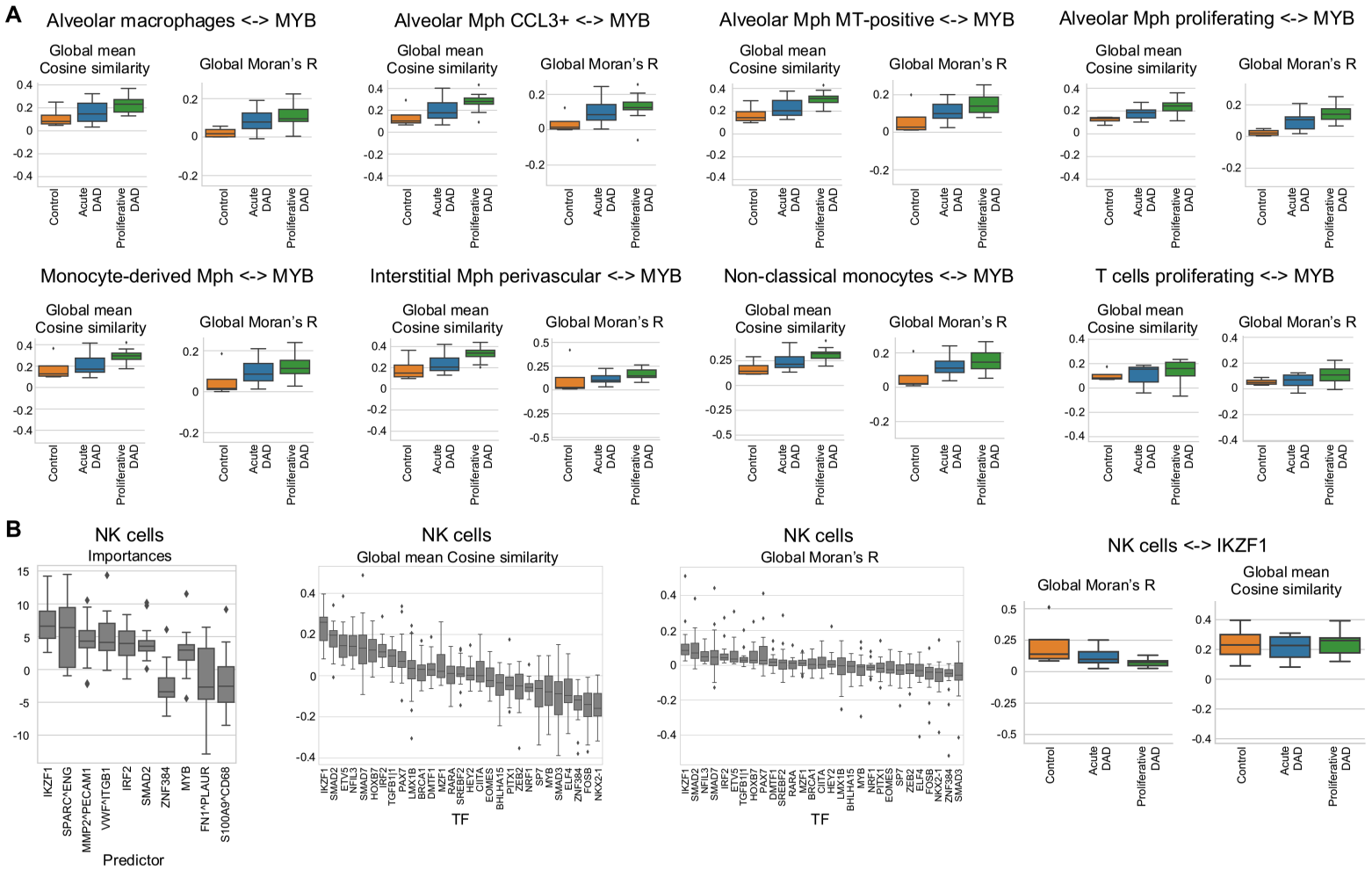
Supplementary Figure 5: Local spatial relationship of APP^AGER ligand-receptor interaction. Cosine similarity, permutation p-values and local categories of APP^AGER interaction, and AT1 cell type abundance on illustrative Visium ST lung samples across disease progression are shown. High-high interactions (red) and high-low or low-high interactions (blue) are depicted in local categories plots. DAD: diffuse alveolar damage.



Supplementary Figure 6: Local spatial relationship of S100A9^CD68 ligand-receptor interaction. Cosine similarity and local categories of S100A9^CD68 interaction, and myeloid and T cells proliferating cell type abundances on illustrative Visium ST lung samples across disease progression are shown. High-high interactions (red) and high-low or low-high interactions (blue) are depicted in local categories plots. Mph: macrophage. MT: metallothionein. DAD: diffuse alveolar damage.



Supplementary Figure 7: S100A9^CD68 and MYB predictive performance for myeloid and T cells proliferating cell types abundances. The top 10 predictors of myeloid cell types and T cells proliferating abundances are shown. Additionally, S100A9^CD68 interaction and MYB activtiy predictor importances (coefficients' t-values as calculated by ordinary least squares) for the same cell types across disease progression are shown. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered to be outliers. Mph: macrophage. MT: metallothionein. DAD: diffuse alveolar damage.



Supplementary Figure 8: Local spatial relationships of MYB and NK cells. A) Global mean Cosine similarity and Global Moran's R for myeloid cell types and T cells proliferating spatial co-clustering with MYB activity are shown across disease progression. **B)** Top 10 predictors of NK cells abundance including their importances (coefficients' t-values as calculated by ordinary least squares), Global mean Cosine similarity and Global Moran's R for NK cells spatial co-clustering with the top 10 transcription factors enriched per condition, and NK cells spatial co-clustering with IKZF1 activity across disease progression are shown. For all box plots, the boxes show the median and interquartile range while the whiskers extend to show the rest of the distribution, except for data points more than 1.5 times the interquartile range outside the low and high quartile, that are considered to be outliers. Mph: macrophage. MT: metallothionein. DAD: diffuse alveolar damage. TF: transcription factor.

Discussion

Discussion

1. The role of variant calling in cancer genomics

The advent of NGS technologies has revolutionized cancer genomics, allowing for the identification of a wide array of somatic mutations that drive cancer development. Large-scale efforts such as TCGA[47] and ICGC[48] have mapped the mutational landscapes of numerous cancers, revealing critical oncogenic mutations and cancer driver genes that have laid the foundation for precision oncology. However, the accuracy of these discoveries depends heavily on the reliability of the variant calling process, which remains one of the most crucial yet variable steps in cancer genome analysis.

Several variant calling tools, such as MuTect2[56], MuSE[57], SomaticSniper[58], and VarScan2[59], have been developed, each with strengths and weaknesses based on the cancer type and research goals. As shown in Chapter I, variant calling decisions significantly impact downstream analyses, including the identification of cancer driver genes and CAVs.

This variability underscores the need for harmonizing variant calling strategies across studies to ensure consistent detection of cancer drivers and actionable mutations. It also highlights the importance of tailoring variant calling strategies to the specific cancer type and research objectives, as different strategies may be better suited for different contexts.

1.1 Impact of variant calling on cancer driver gene detection

The accurate detection of cancer driver genes is a cornerstone of cancer genomics, as these genes often serve as key targets for therapeutic interventions. Our analysis revealed significant variability in the performance of different variant calling tools, which poses challenges in reliably identifying cancer drivers. Notably, only about half of the 3.5 million somatic mutations reported by TCGA were consistently identified across all variant callers, underscoring the need to employ multiple variant callers or combine their outputs to ensure comprehensive mutation detection.

Our findings suggest that using the union of all somatic mutations detected by different callers yielded the best results for cancer driver gene detection using IntOGen[14]. This

result contrasts with the more conservative 'consensus' approaches, often employed in large consortia such as MC3[64] and PCAWG[65], which prioritize specificity over sensitivity. While reducing false positives is essential, the union approach showcases IntOGen's robustness in effectively managing a broader mutation spectrum. This highlights that a balance between sensitivity and specificity is crucial for comprehensive cancer driver gene detection.

The strength of IntOGen lies in its multi-step approach[14], which filters out samples with abnormalities (e.g., hypermutator phenotypes), integrates multiple state-of-the-art driver discovery methods, and combines results using a weighted vote system based on method credibility. This process culminates in a post-processing step that filters out spurious candidates, maintaining high sensitivity without compromising specificity.

The 'wisdom of crowds' approach used by projects like MC3, which aggregates only mutations detected by multiple callers[64], while effective in generating high-confidence variant sets, tends to sacrifice sensitivity. Our findings indicate that requiring mutations to be called by at least three variant callers leads to the loss of significant mutations, resulting in the suboptimal detection of cancer driver genes. A more effective strategy involves less stringent thresholds, such as the union of all somatic mutations or requiring calls from at least two callers.

Using multiple variant calling tools allows for a more nuanced capture of the complexity inherent in cancer genomes. Our analysis provided insights into how the optimal variant calling strategy may vary by cancer type. For instance, in prostate adenocarcinoma, nearly twice as many cancer driver genes were detected depending on the variant call set, highlighting the influence of the underlying mutational landscape on cancer driver gene detection. Moreover, we provide a comprehensive guide for researchers to select the most appropriate variant calling strategy for cancer driver gene detection, tailored to specific cancer types and goals. This guidance is crucial, as it offers a systematic approach to navigating the complexity of variant calling tools based on the specific cancer under investigation. From a translational perspective, pre-selecting the optimal variant calling strategy for each cancer type could significantly enhance clinical decision-making, potentially improving treatment outcomes by ensuring that key drivers are not missed. This approach emphasizes the need for customizing variant calling strategies to the cancer

type, as differences in the mutational landscape can profoundly impact the detection of clinically relevant genes.

Our results further showed that the number of cancer driver genes detected correlated with both mutation burden per megabase and cohort sample size. Larger cohorts inherently provide greater statistical power, enabling the detection of less common driver genes. One of the ultimate goals of cancer genomics is to elucidate the full compendium of cancer driver genes[14]. While over 700 driver genes have been identified[83], significant gaps remain, particularly regarding underrepresented populations, different tumor stages (especially metastatic disease), and cancer types that are less studied. Expanding the diversity of datasets is essential for uncovering rare cancer driver genes that might play important roles in tumor biology.

A notable limitation of our study, similar to others like MC3[81], is its focus on cancer driver genes detected primarily through point mutations, specifically SNVs and short indels. While these alterations provide critical insights, they exclude other key types of genomic alterations, such as copy-number variations, genomic rearrangements, and epigenetic modifications, including DNA methylation silencing. Moreover, non-coding mutations, which have remained largely unexplored due to challenges in modeling their background mutation rates[14], also contribute to tumorigenesis. A more comprehensive understanding of cancer requires a catalogue that includes all types of driver alterations, both coding and non-coding mutations, structural variants, and epigenetic events.

Finally, an open question remains: How many more cancer driver genes are yet to be discovered? While it is likely that most frequently mutated driver genes have already been identified, uncovering drivers with lower mutation frequencies will require larger, more diverse datasets. Greater representation of different ethnic backgrounds in cancer genomic studies is crucial to achieve a complete understanding of cancer drivers. Addressing these gaps will not only advance cancer research but also open new avenues for therapeutic interventions, contributing to a more personalized approach to cancer treatment.

1.2 Impact of variant calling on mutational signatures quantification

Mutational signatures represent a pivotal advancement in cancer genomics, offering deep insights into the biological processes driving tumor development. These signatures reflect mutational mechanisms such as defective DNA repair or mutagenic exposures, providing valuable information on cancer development. Their use as biomarkers has the potential to transform our understanding of tumor etiology and guide therapeutic decision-making.

Our study demonstrated that mutational signatures are highly robust to variant calling variability. Despite significant discrepancies in somatic mutation detection across different callers, the identification of mutational signatures remained remarkably consistent. This robustness emphasizes the value of mutational signatures as reliable biomarkers, even when variant calling strategies differ.

For example, we consistently detected ubiquitous flat signatures such as SBS5 and SBS40, which are present across multiple cancer types, as well as signatures like SBS1, associated with the spontaneous deamination of methylated cytosines[84,85,87]. Although SBS1 is a source of false-positive calls due to germline mutations at CpG sites[227], it was detected uniformly across all cohorts, reinforcing the stability of mutational signatures as biomarkers that are largely unaffected by the choice of variant calling strategy.

One limitation of our study was its focus on five TCGA cancer types, selected to represent diverse mutational processes, purity levels, mutation rates, and cohort sizes. While computational constraints limited the scope of this analysis, future research should expand to include additional cancer types and mutational signatures, particularly doublet-base substitutions and short indels[85], to provide a more comprehensive understanding of their biological and clinical relevance.

Looking ahead, the potential of mutational signatures goes beyond their current applications. While signatures such as MSI-High/dMMR and TMB-High (≥10 mutations per megabase) have already demonstrated their clinical utility in predicting immunotherapy responses and received FDA approval as tissue-agnostic biomarkers[93], future research should prioritize the discovery of novel signatures that are directly linked to treatment response, resistance, and prognosis. These discoveries could significantly

enhance the predictive accuracy of biomarker-based tests, offering more personalized therapeutic strategies for patients.

The robustness of mutational signature analysis across variant calling strategies supports their expanded use in clinical settings. A key challenge for future studies will be to identify new mutational signatures that predict treatment responses, particularly in the context of immunotherapy or targeted therapies. These signatures could inform tumor-agnostic therapeutic approaches, enhancing precision medicine by guiding treatment decisions based on mutational patterns rather than histological classifications.

Incorporating mutational signatures into routine genomic profiling will require refining the tools used for analysis, particularly for WES and WGS[93]. While WES provides important insights, WGS has the advantage of capturing a wider range of mutational signatures, including single- and doublet-base substitutions as well as short indels patterns, which may be more predictive of treatment responses than individual gene mutations alone. Expanding the use of mutational signatures, particularly through WGS, holds the potential to deepen our understanding of drug responses and advance personalized cancer care.

1.3 Impact of variant calling on clinically actionable variant identification

Identifying CAVs is essential in precision oncology, as these variants guide therapeutic decisions by serving as biomarkers for drug sensitivity, resistance, and prognosis. Our study revealed significant variability in the detection of somatic mutations across different variant calling strategies, including those that are clinically relevant. This variability was particularly evident in the detection of missense and nonsense mutations within cancer driver genes, with only around 60% of these mutations being consistently detected across all variant callers. For example, in the TCGA cohorts, up to 22% of uterine corpus endometrial carcinoma patients (196 patients) and 27% of pancreatic adenocarcinoma patients (49 patients) displayed discrepancies in their *PTEN* and *KRAS* mutational statuses, respectively, depending on the variant caller used.

Importantly, CAVs represent only a small fraction of all somatic mutations. In our study, out of over 3.5 million somatic variants analyzed, only 1% were identified as clinically

actionable or biologically relevant, and even fewer were directly associated with therapy response or disease prognosis. This highlights the difficulty in interpreting the vast majority of variants, which are of unknown biological and clinical significance[93]. Furthermore, just over half of all CAVs were detected consistently across all variant calling strategies. Such variability may lead to missed treatment opportunities, as accurate detection of CAVs is essential for guiding targeted therapy.

Our analysis revealed that 10% of all CAVs, including 6% of FDA-approved variants, were exclusively detected by MuTect2, many of which had low VAFs. After adjusting VAFs for cancer DNA fraction and ploidy[55], we interpreted these low VAF mutations as subclonal, providing a clearer view of variant calling performance across different mutational landscapes. Detecting subclonal mutations is crucial, as they can significantly influence patient-specific treatment strategies. The high read depth at these loci suggests that these calls are unlikely to be false positives, further indicating MuTect2's superior sensitivity in identifying subclonal variants compared to other tools[56]. This was especially evident in prostate adenocarcinoma, a cancer characterized by high intra-tumor heterogeneity, where greater variability in cancer driver gene detection likely reflected the presence of diverse subclonal populations. Importantly, many of these variants were absent from the MC3 call set[64], illustrating the limitations of the 'wisdom of crowds' approach.

Moreover, we observed significant discrepancies in the detection of MSI-associated CAVs, an FDA-approved predictive biomarker for response to anti-programmed cell death protein 1 (PD-1) therapy[93]. Accurate detection of MSI-associated variants is crucial for identifying patients eligible for immunotherapy. However, only 20% of these variants were consistently detected across all variant callers. This variability likely stems from the fact that only two of the four callers, MuTect2 and VarScan2, can detect short indels, which are important for determining MSI status. These two callers exclusively identified 70% of samples with MSI-associated CAVs classified as MSI-High, underscoring the important role of short indel detection in guiding treatment decisions. By comparing the performance of different variant callers, we were able to quantify the degree to which their capabilities influence the detection of MSI-associated CAVs. This underscores the importance of selecting the right tools to capture clinically relevant MSI variants for precision oncology.

Despite these insights, one limitation of our study is the absence of a thorough assessment of how sequencing coverage affects variant detection. While we provided coverage information in our plots to aid in assessing the reliability of mutation calls, particularly in low VAF ranges, we did not explore how varying sequencing depths might affect variant detection in clinical practice. In clinical settings, targeted sequencing panels are commonly used, offering deeper coverage on specific genes and mutations[93]. As a result, the discrepancies observed between variant callers in our study may be less pronounced in clinical scenarios where high coverage enables more reliable detection of specific mutations. Furthermore, we did not include copy number or structural variants in our analysis, limiting the scope of our analysis. Future studies should account for these factors to better understand their impact on variant calling performance, particularly in clinical contexts.

Lastly, it is important to note that variant calling is part of a broader, multi-step computational process that includes important preprocessing tasks, such as mapping sequencing reads to a reference genome and performing quality assessments[55]. Although our study focused on variant caller selection, differences in these earlier steps can also influence outcomes. We relied on the harmonized pipelines provided by the Genomic Data Commons[67], which standardizes many of these processes. However, variations in preprocessing workflows across different studies can lead to discrepancies in results.

2. Epigenetic determinants of CAR T-cell therapy response

CAR T-cell therapies have been highly successful in treating R/R B-ALL and B-NHL malignancies. However, challenges remain in optimizing therapeutic outcomes, specifically improving response rates, extending remission durability, and minimizing toxicities[169]. Epigenetic profiling, particularly DNA methylation, has emerged as a powerful tool for identifying molecular and cellular factors that influence CAR T-cell therapy efficacy. DNA methylation profiling stands out for its ability to reveal markers linked to T-cell functionality, persistence, and therapeutic outcomes[184]. Investigating these epigenetic modifications provides valuable insights into the mechanisms underlying CAR T-cell fitness and antitumor activity. Understanding these mechanisms presents opportunities to predict treatment responses more accurately, refine manufacturing processes, and ultimately improve CAR T-cell efficacy.

In Chapter II, the EPICART signature is introduced as a key epigenetic marker associated with CAR T-cell therapy outcomes. Derived from DNA methylation patterns, EPICART indicates critical T-cell functional states that directly affect CAR T-cell efficacy. CAR T-cell products with an EPICART-positive signature are associated with CR and improved clinical outcomes, driven primarily by a higher proportion of naïve and early memory T-cell phenotypes, which correlate with favorable therapeutic responses[169]. Given the heterogeneity in CAR T-cell infusion products, these findings underscore the importance of T-cell fitness in determining CAR T-cell persistence and antitumor effectiveness. Unlike transcriptional or metabolic changes, epigenetic modifications provide a stable and long-lasting reflection of cell fate[185], making DNA methylation a valuable tool for understanding CAR T-cell functionality and exhaustion.

2.1 Potential use of EPICART for CAR T-cell manufacturing optimization

Our findings highlight the potential for using epigenetic programs, such as those identified by EPICART, to optimize CAR T-cell manufacturing (**Figure 4**). Early detection of epigenetic biomarkers in the infusion product offers a unique opportunity to adjust the manufacturing process, improving CAR T-cell functionality and persistence. These interventions could enhance response rates and result in more durable remissions by ensuring higher quality in the final CAR T-cell product. Given that less-differentiated naïve and early memory T cells are linked to better clinical outcomes, targeting these phenotypes through specific cytokine environments or pathway inhibition during cell culture could improve efficacy[169].

Recent advances in DNA-targeting technologies, such as clustered regularly interspaced short palindromic repeats (CRISPR)-based epigenome editing, provide a promising approach for fine-tuning T-cells at the epigenetic level without introducing DNA breaks[228]. This precise gene modulation technique presents a safe and stable way to engineer CAR T-cells that are more potent, durable, and resistant to exhaustion, paving the way for next-generation CAR T-cell therapies (**Figure 4**).

Moreover, prolonged *ex vivo* culture times are associated with increased T-cell exhaustion and the development of less favorable phenotypes, underscoring the need to

minimize culture duration[229]. The changes in DNA methylation patterns observed under different culture conditions emphasize the importance of exploring various CAR constructs, cytokine combinations, and viral integration vectors. In our EPICART study, we profiled three academic CAR T-cell products, each with distinct CAR designs and manufacturing processes, differing in culturing times (7 to 14 days), media (IL-2, IL-7 and IL-15), and viral integration vectors (lentivirus and gammaretrovirus). Continuous monitoring of methylation patterns during the manufacturing process allows for real-time adjustments, optimizing CAR T-cell functionality and persistence, aiming to improve therapeutic outcomes.

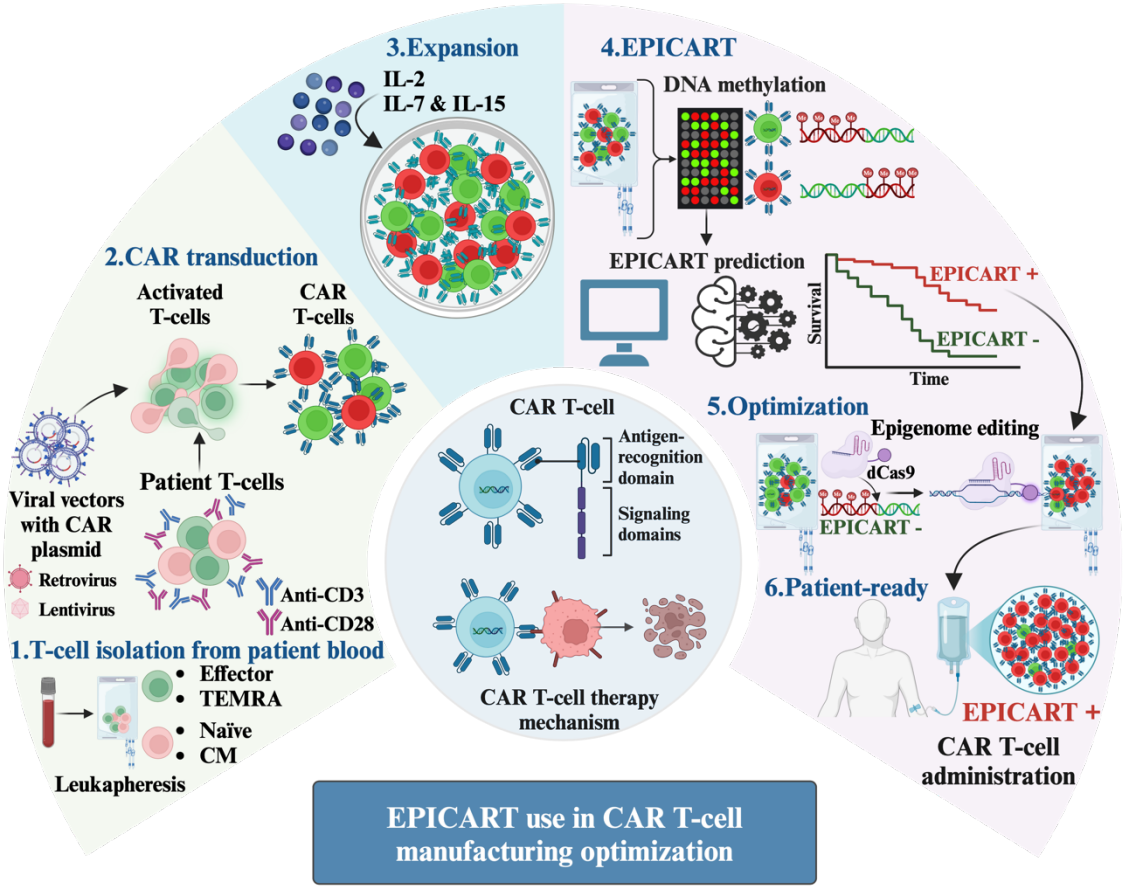


Figure 4. Role of EPICART in optimizing the CAR T-cell manufacturing process. The integration of EPICART into the CAR T-cell manufacturing process aims to enhance quality control and optimize the infusion product for therapeutic efficacy. The steps involved include: (**1**) T-cell isolation from patient blood, (**2**) CAR transduction using viral vectors, (**3**) expansion of transduced cells, (**4**) EPICART-based prediction of infusion product quality, (**5**) epigenome editing to convert EPICART-negative products to EPICART-positive, and (**6**) final preparation of EPICART-positive infusion product for patient administration. EPICART serves as an essential quality control checkpoint, ensuring the product is ready for effective therapeutic use. CAR: chimeric antigen receptor; TEMRA: terminally differentiated effector T-cells; CM: central memory T-cells; IL: interleukin; dCas9: dead CRISPR-associated protein 9. Created in BioRender.com

2.2 Clinical relevance of EPICART

Our analysis included patients diagnosed with both B-ALL and B-NHL, representing a diverse group of pediatric and adult populations. Overall, 65% of patients achieved CR, which aligns with reported CAR T-cell therapy outcomes[169]. EPICART demonstrated strong classification performance, with an area under the receiver operating characteristic (ROC) curve of 0.80, indicating its effectiveness in classifying clinical responses. Furthermore, the obtained Cohen's kappa score of 0.6, indicating moderate agreement, reflects the alignment between EPICART's predicted clinical responses and the actual observed outcomes, taking into account the possibility of agreement occurring by chance and going beyond simply predicting the dominant class[230]. This provides a more balanced and reliable measure of EPICART's performance. While this result strengthens EPICART's potential, further validation in larger, independent cohorts is needed to confirm its effectiveness, especially across B-NHL subtypes and different age groups.

Younger patients, particularly pediatric populations, tend to experience better outcomes following CAR T-cell therapy compared to adults. This is largely due to the higher proportion of naïve and early memory T-cell phenotypes in younger patients, which contribute to immune cell plasticity and enhanced persistence, critical factors for the long-term efficacy of CAR T-cell therapies[169]. In contrast, older patients, affected by thymic involution, exhibit a decline in naïve T-cells and an accumulation of exhausted memory T-cells, limiting CAR T-cell effectiveness[231]. To provide deeper insights into these dynamics, EPICART captures the broader epigenetic landscape, revealing a continuum of T-cell states[232] that may not be fully apparent through surface markers detected by flow cytometry, which traditionally define T-cell phenotypes. This more comprehensive understanding of T-cell heterogeneity offers a nuanced view of functional T-cell states, ultimately leading to better-informed therapeutic decisions and more personalized treatment strategies for CAR T-cell therapies.

While EPICART is predictive of CR to CD19 CAR T-cell therapy, it is also associated with better long-term clinical outcomes, particularly overall survival. Importantly, disease relapse can still occur even after MRD-negative remissions, underscoring that a deep initial response may not always guarantee long-term remission. This highlights the importance of assessing long-term survival to fully validate EPICART's clinical utility.

Furthermore, it is important to consider the differing disease courses of B-NHL and B-ALL patients. B-NHL patients, while less likely to achieve CR, tend to maintain sustained remission once CR is achieved. In contrast, B-ALL patients, although more likely to achieve CR, are at higher risk of relapse[169]. While our study did not specifically stratify these subgroups due to sample size limitations, EPICART was associated with enhanced overall survival across the cohort, reinforcing its broader clinical relevance. This highlights EPICART's potential as a valuable tool for predicting CAR T-cell therapy outcomes across various B-cell malignancies.

2.3 Validation of EPICART for clinical implementation

Our study demonstrated that DNA methylation arrays are a robust tool for profiling differentially methylated loci associated with CAR T-cell functionality and therapeutic outcomes. EPICART's findings were supported by multiple methods, including flow cytometry, which revealed the higher proportion of naïve and central memory T-cell phenotypes in EPICART-positive products, and pyrosequencing and bisulfite genomic sequencing, which verified the methylation status of specific CpG sites identified by EPICART.

The functional relevance of EPICART was further validated by correlating DNA methylation changes with gene expression alterations. This was confirmed in a subset of CpG sites using quantitative reverse transcription–polymerase chain reaction (qRT-PCR) and Western blot analyses, demonstrating the impact of DNA methylation changes on transcriptional and protein-level alterations. These validations not only strengthen EPICART's biological significance but also underscore its potential as a clinically relevant biomarker for assessing CAR T-cell functionality and predicting therapeutic outcomes.

Despite these validations, phenotype heterogeneity remains a significant challenge in CAR T-cell therapy, contributing to variability in treatment outcomes. A limitation of our study was the use of bulk tissue analysis, which may have introduced noise by mixing different cell populations, potentially masking important signals and leading to an averaging effect. To overcome this, future studies employing single-cell technologies will

allow for high-resolution exploration of individual cellular phenotypes and states, deepening our understanding of CAR T-cell behavior at the single-cell level[184].

For broader clinical implementation, it is crucial to validate EPICART using PCR-based assays, which are widely available in hospital settings[233]. PCR offers a faster and more accessible method, potentially streamlining EPICART's adoption in routine clinical workflows by reducing turnaround times. However, developing a PCR-based version of EPICART will require further technical and biological validation, including training a new signature version based on PCR readouts to ensure its accuracy and reliability on this new platform.

The consistent correlation of EPICART with clinical outcomes across diverse academic CAR T-cell products underscores its robustness and utility. However, for EPICART to serve a truly universal predictive signature, it must be validated across independent cohorts and FDA-approved commercial CAR T-cell products. Comparing academic and commercial CAR T-cell designs will be essential to identify potential differences and ensure EPICART's broad applicability. Additionally, profiling DNA methylation changes at various stages of CAR T-cell manufacturing, particularly during activation and expansion, may provide deeper insights into optimizing CAR T-cell products for enhanced therapeutic efficacy.

2.4 Future challenges of EPICART and CAR T-cell therapy

Minimizing toxicities and predicting adverse events remain critical challenges in CAR T-cell therapy. Although EPICART successfully identified DNA methylation patterns associated with therapeutic outcomes, our efforts to detect reliable biomarkers for key toxicities, such as CRS and ICANS, were inconclusive. While some trends emerged, they did not reach statistical significance, underscoring the complexity of predicting these side effects.

Toxicities like CRS and ICANS limit the broader clinical application of CAR T-cell therapies. Future developments of EPICART and similar models will benefit from integrating multi-omics data[184]. By refining EPICART with multidimensional approaches, we may better anticipate adverse events, enabling earlier intervention

strategies and ultimately improving the safety and scalability of CAR T-cell therapy in clinical practice.

EPICART's relevance may extend beyond autologous CAR T-cell therapies, offering potential as a biomarker for 'off-the-shelf' CAR T-cell products derived from healthy donors[184]. Using T-cells from healthy individuals can reduce manufacturing costs, shorten production times, and mitigate issues related to the impaired functional fitness of patient-derived T-cells. This approach, combined with EPICART, could improve the effectiveness and scalability of CAR T-cell therapies.

Additionally, longitudinal profiling post-infusion is essential to understand the determinants of CAR T-cell expansion and persistence *in vivo*. There is evidence that viral integration vectors used in adoptive cell transfer therapies may be susceptible to epigenetic silencing via vector hypermethylation, potentially influencing clinical outcomes[197]. Our pyrosequencing analysis of the retroviral vector in transduced T-cells showed no methylation before infusion, but post-infusion methylation changes warrant further investigation.

Furthermore, integrating T-cell receptor (TCR) sequencing with transcriptomic profiling can link specific TCR clonotypes to different T-cell phenotypes, providing valuable insights into T-cell clonality and dynamics over time[184]. This combined analysis could refine therapeutic strategies and improve long-term outcomes. Emerging evidence also suggests that administering CAR T-cell therapies earlier in treatment lines, before multiple rounds of chemotherapy, may result in better outcomes due to more robust immune systems[169]. Ongoing trials are exploring the efficacy of CAR T-cell therapies in less refractory patients, potentially capitalizing on healthier immune systems for improved responses[169].

Beyond scientific and technical challenges, addressing the high cost of CAR T-cell therapies is critical[234]. Most commercial CAR T-cell products are developed by pharmaceutical companies, leading to prohibitively high treatment costs. Academic institutions, in collaboration with hospitals, are pioneering CAR T-cell development at significantly lower costs[235]. This collaborative approach not only fosters innovation, as demonstrated in our work on EPICART, but also aims to make life-saving therapies more

accessible, ensuring that financial barriers do not prevent patients from receiving essential care.

However, translating academic innovations like EPICART into routine clinical practice often requires partnerships with pharmaceutical companies. While academic research drives innovation and reduces costs, commercial collaborations are vital for scaling and validating these technologies across diverse clinical settings. To this end, we have filed a patent to protect EPICART and have successfully licensed it to a pharmaceutical company. This partnership will allow the validation of EPICART across various CAR constructs and patient cohorts, assessing its robustness and adaptability. By leveraging this collaboration, we aim to explore EPICART's clinical utility in real-world settings, expanding its application and potential to improve patient outcomes.

Ultimately, making a lasting impact requires collaboration across a broad range of stakeholders, including academic researchers, pharmaceutical companies, and clinical practitioners. This multisectoral approach enhances the likelihood that technologies like EPICART are validated, optimized, and integrated into clinical practice, transforming academic innovations into practical tools that benefit patients across diverse healthcare settings.

3. Epigenetic insights into MIS-C pathophysiology

In Chapter III, the EPIMISC signature is introduced to characterize the distinct epigenetic alterations driving the hyperinflammatory response and immune dysregulation in MIS-C. Using DNA methylation analysis, similar to our prior work on CAR T-cell therapies, we sought to dissect the molecular mechanisms underlying MIS-C's hyperinflammatory response. Given the severity of this condition, which is marked by multi-organ inflammation, understanding these mechanisms holds potential for earlier diagnosis and more targeted therapeutic interventions.

Our study reveals a distinct epigenetic landscape associated with MIS-C, differentiating it from the generally mild pediatric response to COVID-19. EPIMISC consists of methylation sites linked to immune and inflammatory pathways, playing a pivotal role in driving the severe multisystem inflammation characteristic of MIS-C. Importantly, EPIMISC was largely absent in healthy children and pediatric COVID-19 cases without

MIS-C. However, it was detected in a subset of severe adult COVID-19 patients, further reinforcing the reported overlap between immune cell signatures and inflammatory parameters in MIS-C and severe adult COVID-19[210]. This shared epigenetic profile points to common mechanisms driving the hyperinflammatory responses in both conditions.

A potential shared mechanism involves *ZEB2*, identified as a key gene in EPIMISC, which plays an important role in driving terminal differentiation of effector T-cells and plasmacytoid dendritic cells (pDCs)[236,237]. In MIS-C, the immune profile shows elevated neutrophils, reduced pDCs, fewer naïve T-cells, and a higher proportion of activated memory T-cells[210], reflecting a more cytotoxic and hyperinflammatory state. Interestingly, *ZEB2*, which is known to be involved in EMT[236], a key process in fibrogenesis, was also identified in our study of fatal COVID-19 as playing a significant role in proliferative DAD lungs. This suggests that *ZEB2* may serve as a common link in the hyperinflammatory responses seen in both MIS-C and severe adult COVID-19 cases, contributing to immune activation and tissue damage in both conditions.

3.1 Clinical relevance and diagnostic potential of EPIMISC

The identification of EPIMISC as a novel biomarker signature linked to the onset of MIS-C provides valuable insights into the pathophysiological mechanisms underlying the syndrome. EPIMISC demonstrated strong diagnostic potential, with an area under the ROC curve of 89.3% and a Cohen's kappa score of 0.79, underscoring its reliability in distinguishing MIS-C from pediatric COVID-19 cases without MIS-C and healthy children. Notably, EPIMISC is largely absent in other pediatric inflammatory disorders and viral infections, including multiple respiratory viruses, further highlighting its specificity and potential as a diagnostic tool.

EPIMISC revealed a significant overlap with KD, a childhood inflammatory vasculitis, with nearly all analyzed KD cases exhibiting the EPIMISC signature. While the onset of MIS-C is linked to prior exposure to SARS-CoV-2, the exact triggers of KD remain elusive, though viral or bacterial infections have been suggested[210]. The identification of this overlap is important as it supports the hypothesis that viral triggers, such as SARS-CoV-2 in MIS-C or other agents in KD, may be key drivers of the pathogenesis in both hyperinflammatory syndromes[210].

Despite prior reports suggesting a potential genetic susceptibility to MIS-C and KD[210,212], our study did not identify significant methylation differences in known genetic loci. This implies that genetic predisposition alone may not fully explain disease pathogenesis. Our findings instead emphasize the role of epigenetic factors, specifically DNA methylation of inflammatory and immunoregulatory genes, in driving disease onset. These results reinforce the importance of epigenetic mechanisms, which have already been recognized as pivotal in KD[238], as key contributors to MIS-C development. Additionally, host factors such as immune system immaturity and epigenetic background may explain why some children develop MIS-C following SARS-CoV-2 exposure, with DNA methylation playing a significant role in modulating immune and inflammatory responses, contributing to both the development and severity of the syndrome.

However, these findings may be influenced by certain limitations. The rarity of MIS-C constrained our sample size, although it is in line with other molecular studies of the disease[212,214]. Additionally, the ethnic homogeneity of our cohort, predominantly of West-Eurasian origin, may limit the generalizability of these findings. Given the higher incidence of MIS-C in children of African and Hispanic heritage[210], future studies must focus on more ethnically diverse cohorts to better understand the full spectrum of risk factors and epigenetic contributors.

Despite these constraints, our study offers important insights into the immune mechanisms underlying MIS-C. EPIMISC, when integrated with clinical, genetic, and serological markers, holds promise for enhancing patient stratification and early detection. Additionally, these biomarkers could prove useful in monitoring therapy effectiveness and identifying early signs of disease progression.

4. Machine learning approaches in epigenomic biomarker discovery

The integration of ML into biomedical research has the potential to transform the discovery of biomarkers, driving significant advancements in precision medicine[233]. In our studies involving EPICART and EPIMISC, several ML algorithms were tested, ranging from random forests and k-nearest neighbors to support vector machines and Extreme Gradient Boosting (XGBoost), to identify DNA methylation signatures

predictive of CAR T-cell therapy outcomes and diagnostic of MIS-C. Despite the capability of these models to uncover complex patterns in high-dimensional data, ridge-regularized logistic regression emerged as the most effective approach in our analyses.

The better performance of logistic regression[239] in our studies supports its use in biomedical research due to several key factors such as interpretability, simplicity, and computational efficiency. The inclusion of ridge regression with L2 regularization helps mitigate overfitting by penalizing large coefficients and shrinking less informative ones. This penalization proves beneficial in managing small sample sizes[240], which is critical in rare conditions like MIS-C, ensuring both stability and generalizability. Ridge regularization also addresses multicollinearity[241], a common issue in DNA methylation data where CpG sites are often highly correlated. By preventing overfitting, ridge regularization improves the model's ability to generalize[242] to new, unseen data, which is essential for clinical applications.

However, linear models like logistic regression come with limitations. The assumption of a linear relationship between predictors and the log-odds of the outcome may not fully capture the non-linear interactions inherent in complex biological processes such as epigenetic regulation. While more advanced models, like XGBoost, may capture these non-linearities, they often lack transparency and interpretability, which are essential for clinical decision-making[243,244].

4.1 Clinical interpretability of ridge-regularized logistic regression

When applying ridge-regularized logistic regression in clinical settings, it is important to consider how regularization impacts the interpretability of the model. In standard logistic regression, coefficients directly represent the log-odds of the outcome and can be easily transformed into odds ratios. These odds ratios quantify the effect of each predictor, providing valuable insights into the strength and direction of associations. Confidence intervals surrounding these odds ratios allow clinicians to assess the precision and reliability of these estimates, offering a clearer understanding of the statistical certainty.

However, ridge regularization introduces a penalty term to reduce overfitting, which is particularly beneficial when dealing with small sample sizes and multicollinearity. This penalty shrinks the coefficients, introducing bias and meaning that the coefficients no

longer represent maximum likelihood estimates as they do in standard logistic regression. As a result, traditional odds ratios and confidence intervals cannot be calculated or interpreted in the same manner, making it harder to directly connect predictors to clinical outcomes. This can present challenges for clinicians, who often rely on these direct relationships for decision-making.

Despite this reduction in direct interpretability, ridge regularization greatly enhances the model's generalizability and stability, critical factors for clinical use. To address the interpretability limitations, we conducted additional analyses on the CpG sites included in EPICART, identifying six loci individually associated with improved event-free and overall survival. These survival analyses offered clinically relevant insights into the prognostic value of the methylation status of these loci, helping to simplify the clinical application of EPICART. This approach enables practitioners to pinpoint individual key methylation sites most relevant to patient outcomes, making integration into routine care more feasible.

Understanding the trade-offs involved in using ridge-regularized logistic regression is vital when communicating results in clinical contexts. While the direct interpretation of coefficients as odds ratios is compromised, the enhanced predictive accuracy and generalizability provided by regularization offer significant advantages for real-world clinical applications.

4.2 Adoption of machine learning algorithms in clinical settings

In clinical practice, the need for clear and actionable decisions often outweighs marginal improvements in predictive accuracy. ML models like ridge-regularized logistic regression, which produce probabilistic scores, offer a flexible approach to clinical decision-making by converting probabilities into binary outcomes. For example, in EPICART, a threshold of 0.5 was used to classify patients as likely complete responders or non-complete responders to CAR T-cell therapy. In EPIMISC, a threshold of 0.3 was deliberately chosen to balance sensitivity and specificity. Given that EPIMISC is highly specific, this threshold allowed the model to become more sensitive in identifying true MIS-C cases while minimizing false positives. This tailored thresholding helps align the ML models with practical clinical needs[245].

While providing binary outcomes based on these thresholds can fit well within clinical decision-making processes, offering clinicians clear and user-friendly classifications, it is important to recognize that valuable probabilistic information may be lost. Probabilistic outputs could provide more nuanced insights into patient treatment response or risk, which might be useful in certain contexts. However, binary outcomes facilitate application in real-world medical practices where clear decisions are often required[246].

Logistic regression strikes a balance between model complexity and clinical applicability. While more complex models may capture non-linear patterns, their lack of interpretability and increased computational demands can hinder their integration into clinical workflows[247]. By contrast, logistic regression offers a pragmatic solution by providing reliable predictive performance coupled with transparency and efficiency.

The straightforward nature of logistic regression aids in regulatory approval and clinical integration. Models with high transparency are more likely to be accepted by regulatory bodies and adopted by healthcare providers, especially for high-stakes medical decisions where the implications of error are substantial[248]. In contexts such as diagnosing MIS-C or evaluating CAR T-cell therapy response, where timely and precise decision-making is crucial, straightforward and interpretable models are invaluable. This aligns with existing recommendations on Artificial Intelligence (AI) in healthcare, where interpretability is often prioritized to build trust and ensure clinical utility[248]. However, as regulatory frameworks evolve and explainable AI methods advance, the use of more complex ML models, such as deep neural networks, may increase without compromising clinical trust[249].

4.3 Limitations of EPICART and EPIMISC

While ridge-regularized logistic regression has been effective in developing EPICART and EPIMISC, it is crucial to recognize the model's inherent limitations. The linear assumptions of the algorithm may not fully account for the underlying biological complexities and interactions present in epigenetic and clinical data. Moreover, the performance of these models can be limited by small sample sizes, which may not represent the full diversity and variability observed in broader populations.

Another significant limitation is the need for extensive validation in external, independent cohorts that reflect real-world clinical settings. While EPICART and EPIMISC have shown promising results in our studies, testing their classification performance and generalizability across diverse patient populations and clinical contexts is an essential next step. Such validation is critical to ensuring that the models are robust, reliable, and applicable across varied demographics and disease presentations. Expanding validation efforts will be crucial for integrating these models into routine clinical practice and enhancing their clinical utility in personalized medicine.

5. Spatial transcriptomics sheds light on diffuse alveolar damage progression in fatal COVID-19

In Chapter IV, we utilized ST to investigate the molecular and cellular mechanisms driving DAD in fatal COVID-19. Our approach combined both local and global spatially-informed bivariate metrics to pinpoint co-expressed ligand-receptor pairs, along with a multi-view modeling approach that integrated spatial information from features such as cell type abundance, LRIs, and TF activity[224]. This approach was instrumental in mapping the intercellular communication networks involved in lung tissue remodeling and fibrosis. This comprehensive analysis provided a more holistic view by integrating multiple data layers, revealing complex spatial relationships within the tissue[224]. Through this, we identified key cellular players, including peribronchial fibroblasts, and highlighted signaling pathways, such as the TGF-β/SMAD3 axis, which were significantly enriched and active in fibrotic regions.

Unlike many ST studies that focus solely on individual sample analyses, our approach took a more integrative perspective. By collectively analyzing all samples and applying NMF, we identified coordinated intercellular communication programs across the various phases of DAD, from acute damage to proliferative fibrosis. This unsupervised, hypothesis-generating approach enabled us to uncover global communication drivers while also linking localized LRIs to broader tissue dynamics[224]. Furthermore, these intercellular communication programs were linked to downstream TF activities, providing a comprehensive view of both extracellular signaling and the corresponding intracellular responses[224]. It is important to emphasize, however, that while these insights

into intracellular signaling are promising, they remain preliminary and require further experimental validation.

Our use of multi-view modeling allowed us to simultaneously examine spatial relationships between diverse biological features, such as cell type composition, LRIs, and TF activities[224]. This approach uncovered complex spatial interactions, revealing how these variables influence one another within the same tissue environment. Although the explained variance in cell type abundance differed across populations, it was particularly insightful for key cell populations like peribronchial fibroblasts, highlighting their significant role in tissue remodeling. While some of the unexplained variance may stem from intra-spot composition or unmodeled variables, our goal was to characterize the intercellular interactions within spatial contexts that contribute to DAD progression. By examining these spatial dependencies, we identified pivotal interactions and signaling pathways that hold potential as therapeutic targets to mitigate lung damage and fibrosis.

Our study highlighted the significant role of aberrant myeloid cell activation in DAD pathogenesis, driven by ligands such as S100A8 and S100A9 and regulated by the TF MYB. Key LRIs, particularly TIMP1-CD63, were identified as central contributors to DAD progression, emphasizing the therapeutic potential of targeting these interactions[250]. Additionally, spatiotemporal trajectory analysis[251] allowed us to characterize the epithelial regeneration process, particularly the differentiation of alveolar type 2 cells into alveolar type 1 cells, a key marker of lung tissue repair. By tracking these trajectories, we identified markers associated with various stages of differentiation, providing valuable insights into how epithelial cells contribute to tissue repair following injury. These mechanistic insights could open avenues for therapeutic interventions aimed at preventing lung damage and promoting tissue regeneration in severe COVID-19 cases.

Both global and local spatial metrics were instrumental in summarizing interactions across the tissue slide and pinpointing specific communication sites. This multi-scale analysis allowed us to capture intricate spatial relationships within the tissue architecture. By combining NMF with spatially informed local LRIs, we uncovered key drivers of fibrosis in DAD and identified a potential causal signaling pathway in peribronchial fibroblasts, linking deregulated LRIs to downstream TGF-β-associated SMAD3 and

SMAD7 TFs. While these insights require further validation, they represent a significant step toward understanding the molecular mechanisms driving DAD.

Validation is crucial to ensure the robustness of our findings, particularly since all downstream analyses rely on the accuracy of cell type annotation and mapping. To assess the reliability of cell type deconvolution, we employed immunohistochemistry techniques with lineage-specific markers. This approach confirmed that cell types were accurately mapped to their expected locations within the tissue, such as epithelial cells lining the small airways and stromal cells along blood vessel walls. Accurate cell type mapping is foundational for subsequent analyses, making this validation step critical to ensure that the identified interactions and pathways are properly aligned with the spatial architecture of lung tissue.

5.1 Limitations and challenges in spatial transcriptomics data analysis

While ST has significantly enhanced our ability to investigate intricate cellular communication processes underlying tissue architecture disruption, several limitations and computational challenges remain that must be addressed to fully realize its potential. One important aspect to consider in our study was the reliance on linear models to infer relationships between variables. These models offer valuable interpretability and computational efficiency, allowing us to compute coefficients' t-values using Ordinary Least Squares to assess the role of specific LRIs or TFs in determining cell type abundance and tissue remodeling[224]. The t-values, which are signed and directly comparable, indicate the direction and magnitude of relationships and serve as a measure of feature importance. However, biological systems often exhibit non-linear relationships, and linear models may not fully capture the complexity of dynamics involved in cellular communication. To address this, future studies should explore advanced non-linear or spatially informed factorization models[252] that can provide deeper insights into cellular communication networks and tissue remodeling processes.

Another limitation stems from our dependence on curated databases like Omnipath[253], which primarily focus on protein-mediated interactions. While these databases provide a solid foundation for studying cell-cell communication, they often lack coverage of other modes of communication, such as those mediated by small molecules or metabolites. Expanding curated resources to include these additional communication modes will be

crucial for gaining a more comprehensive understanding of cell-cell communication processes.

Additionally, adapting prior knowledge to the specific cell types, tissues, or diseases under study is essential to reduce erroneous predictions. In our study, we utilized causal inference methods, such as LIANA+[224], to infer signaling networks while distinguishing between the activation and inhibition of molecular interactions. This distinction is vital for accurately modeling biological processes. However, the success of these approaches heavily depends on the accuracy and completeness of prior knowledge, highlighting the ongoing need to expand and tailor context-specific databases to ensure reliable and meaningful insights.

Validation remains a significant challenge for ST data analysis, especially in hypothesis-generating approaches. While our findings point to potentially significant cell-cell interactions and intracellular signaling pathways, experimental validation is necessary to confirm their biological relevance. In our study, we validated cell type annotations using immunohistochemistry markers, ensuring accurate mapping of cells to expected tissue locations. Although this validation step was essential, more advanced techniques capable of directly capturing genuine cell-cell communication events[254] are needed to confirm inferred interactions and to further support the biological and clinical relevance of our results.

Lastly, while cell-cell communication inference methods are powerful tools for hypothesis generation, translating ST findings into clinical practice presents additional challenges. The integration of multidimensional data, the development of more sophisticated computational tools, and the establishment of standardized workflows will be critical to making ST a reliable method in clinical diagnostics and therapeutic development. Overcoming these challenges will require continued refinement of bioinformatics tools[221], expansion of prior knowledge resources[253], and broader validation efforts to ensure that ST-derived insights can be translated into actionable clinical interventions.

6. Outlook to truly advance precision medicine

Precision medicine is set to revolutionize how we diagnose and treat diseases, offering highly personalized therapeutic options based on both molecular and clinical characteristics. Throughout this thesis, I have discussed the pivotal role of bioinformatics in driving these advancements. By leveraging diverse datasets, including genetic, epigenetic, spatially resolved, and clinical information, we aim to bridge molecular mechanisms with practical clinical applications, generating actionable insights that can reshape clinical practice. However, fully incorporating precision medicine into routine clinical settings remains challenging and will require collaboration across multiple fronts, from bioinformaticians and clinicians to regulators.

One clear example of how precision medicine is expected to evolve in the coming years is the shift toward tumor-agnostic approaches in cancer treatment[255]. Recent advances, such as the approval of pembrolizumab for tumors characterized by high MSI or TMB[93], demonstrate how treatment decisions are increasingly being driven by molecular alterations rather than the tumor's tissue of origin. This paradigm shift has broad implications, not only for how we classify cancers but also for how we develop targeted therapies across malignancies. It reflects a move away from organ-based cancer classification, focusing instead on molecular-targeted therapies that have the potential to revolutionize oncology practice.

Regulatory agencies must refine their approval processes to accommodate these organ-agnostic therapies, necessitating the development of novel clinical trial designs like basket trials[93]. These trials test therapies based on mutational status rather than cancer origin, challenging the traditional approach of approving drugs for each cancer type sequentially. Additionally, healthcare providers, including clinicians, medical students, and institutions, must adapt their education and clinical practices to focus on the molecular mechanisms driving cancer, while shifting clinical workflows and mindsets toward these new treatment paradigms[255]. Institutions must provide thorough training to ensure they are equipped to implement these new approaches effectively.

In this context, bioinformatics plays an indispensable role in guiding personalized treatment strategies. As demonstrated in our findings, variant calling pipelines, which

help identify cancer drivers and CAVs, are crucial tools for informing therapeutic decisions. Despite advances in improving the accuracy and consistency of these resources, effectively communicating and reporting insights from databases cataloging CAVs to point-of-care clinicians remains a significant challenge[93]. Equally important for clinical adoption is the interpretability of ML models, such as EPICART or EPIMISC. These models must deliver clear and actionable insights that can be easily understood by clinicians, not just data scientists. This further underscores the importance of fostering collaboration between bioinformaticians, clinicians, and other stakeholders to bridge the gap between computational models and practical, real-world applications.

The COVID-19 pandemic demonstrated the importance of collaboration across sectors, with the rapid development and deployment of vaccines showing how governmental agencies, academia, and pharmaceutical companies can work together to address global challenges[201]. Similarly, advancing precision medicine globally will require cooperation between regulators, academia, biotech companies, and clinical teams. Expanding access to molecular testing and advanced therapies in low- and middle-income countries is crucial to ensure that precision medicine is not limited to high-income regions.

The promise of precision medicine extends beyond cancer and infectious diseases like COVID-19. By integrating multi-omics data, including genomic, epigenomic, proteomic, transcriptomic, spatially resolved, and clinical data, we can achieve a more comprehensive understanding of disease mechanisms. Bioinformatics will play a key role in developing accessible, user-friendly tools to manage and interpret these complex datasets, enabling the discovery of novel biomarkers, a deeper understanding of molecular pathways, and ultimately, the ability to tailor treatments to individual patients.

In conclusion, the advancement of precision medicine requires not only technological and scientific innovation but also a collaborative ecosystem involving clinicians, bioinformaticians, regulators, and policymakers. As the title of this thesis suggests, this is a *multifaceted affair*. Success will depend on overcoming computational, clinical, and regulatory barriers alike, while ensuring global access to these innovations. By adopting an integrated and collaborative approach, we can unlock the full potential of precision medicine and make it a reality for patients worldwide.

Conclusions

Conclusions

1. The use of different variant calling tools leads to significant differences in the detection of cancer driver genes and clinically actionable variants, while having no notable impact on the quantification of mutational signatures.

2. A tailored variant calling strategy is required for each cancer type, as no single tool is universally optimal. The union of mutations from all variant callers outperforms more conservative strategies in detecting cancer driver genes. Among individual variant callers, MuTect2 identified more subclonal mutations and clinically actionable variants linked to therapeutic outcomes.

3. The DNA methylation profile of pre-infusion CD19-targeted CAR T-cells influences therapeutic outcomes in relapsed or refractory B-cell malignancies. The EPICART signature, a classification model based on DNA methylation markers, successfully predicted complete clinical response and was associated with improved overall survival in a cohort of 114 patients across three academic trials.

4. CAR T-cell products classified as being EPICART-positive contained higher proportions of naïve and central memory T-cells, leading to improved therapeutic outcomes compared to EPICART-negative products, which contained higher proportions of effector memory and terminally differentiated effector memory T-cells.

5. The EPIMISC signature, a DNA methylation-based classification model, successfully differentiated MIS-C patients from pediatric COVID-19 cases without MIS-C and from healthy children. EPIMISC supports the role of DNA methylation changes in driving hyperactivated immune responses in MIS-C. Its presence in Kawasaki disease suggests that shared immune mechanisms, likely triggered by viral infections like SARS-CoV-2 in MIS-C, contribute to the pathogenesis of both conditions.

6. Spatial transcriptomics provided a comprehensive understanding of the molecular and cellular mechanisms driving diffuse alveolar damage progression in fatal COVID-19. Global and local analyses of ligand-receptor interactions and transcription factor activity revealed communication programs that drive lung tissue remodeling and fibrosis, with

aberrant myeloid activation, peribronchial fibroblasts, and the TGF-β/SMAD3 pathway as major contributors.

7. EPICART exemplifies how multisector collaboration aims to drive precision medicine forward. By licensing it to a pharmaceutical company, we have paved the way for its validation across diverse patient cohorts, bringing it closer to potential clinical application. This highlights the critical role of partnerships in translating bioinformatic discoveries into real-world impact.

References

References

1. Hardy, A. ‘Death is the Cure of All Diseases’: Using the General Register Office Cause of Death Statistics for 1837–1920. *Social History of Medicine* **7**, 472–492 (1994).
2. Condrau, F. & Worboys, M. Second Opinions: Epidemics and Infections in Nineteenth-Century Britain. *Social History of Medicine* **20**, 147–158 (2007).
3. Armstrong, G. L., Conn, L. A. & Pinner, R. W. Trends in Infectious Disease Mortality in the United States During the 20th Century. *JAMA* **281**, 61–66 (1999).
4. Pasteur, L. *Studies on Fermentation: The Diseases of Beer, Their Causes, and the Means of Preventing Them*. (Macmillan & Company, 1879).
5. Ehrlich, P. & Bertheim, A. Über das salzsaure 3.3′-Diamino-4.4′-dioxy-arsenobenzol und seine nächsten Verwandten. *Berichte der deutschen chemischen Gesellschaft* **45**, 756–766 (1912).
6. Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol* **10**, 226–236 (1929).
7. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians* **74**, 12–49 (2024).
8. Halsted, W. S. I. The Results of Operations for the Cure of Cancer of the Breast Performed at the Johns Hopkins Hospital from June, 1889, to January, 1894. *Ann Surg* **20**, 497–555 (1894).
9. Connell, P. P. & Hellman, S. Advances in Radiotherapy and Implications for the Next Century: A Historical Perspective. *Cancer Research* **69**, 383–392 (2009).
10. Gilman, A. The initial clinical trial of nitrogen mustard. *The American Journal of Surgery* **105**, 574–578 (1963).
11. Chabner, B. A. & Roberts, T. G. Chemotherapy and the war on cancer. *Nat Rev Cancer* **5**, 65–72 (2005).
12. Farber, S., Diamond, L. K., Mercer, R. D., Sylvester, R. F. & Wolff, J. A. Temporary Remissions in Acute Leukemia in Children Produced by Folic Acid Antagonist, 4-Aminopteroyl-Glutamic Acid (Aminopterin). *New England Journal of Medicine* **238**, 787–793 (1948).
13. Huebner, R. J. & Todaro, G. J. Oncogenes of rna tumor viruses as determinants of cancer. *Proceedings of the National Academy of Sciences* **64**, 1087–1094 (1969).

14. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555–572 (2020).
15. Druker, B. J. *et al.* Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *New England Journal of Medicine* **344**, 1031–1037 (2001).
16. Mardis, E. R. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics* **9**, 387–402 (2008).
17. Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).
18. Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene Activity During Development. *Science* **187**, 226–232 (1975).
19. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3**, 415–428 (2002).
20. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
21. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
22. Ewing, J. *Neoplastic Diseases: A Treatise on Tumors*. (W.B. Saunders, 1928).
23. Brown, J. S. *et al.* Updating the Definition of Cancer. *Molecular Cancer Research* **21**, 1142–1147 (2023).
24. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231–1251 (2021).
25. Rosai, J. & Ackerman, L. V. The pathology of tumors, part III: grading, staging & classification. *CA Cancer J Clin* **29**, 66–77 (1979).
26. O'Sullivan, B. *et al.* The TNM classification of malignant tumours—towards common understanding and reasonable expectations. *Lancet Oncol* **18**, 849–851 (2017).
27. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
28. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
29. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229–263 (2024).

30. Jokhadze, N., Das, A. & Dizon, D. S. Global cancer statistics: A healthy population relies on population health. *CA Cancer J Clin* **74**, 224–226 (2024).
31. Williams, P. A., Zaidi, S. K. & Sengupta, R. AACR Cancer Progress Report 2024: Inspiring Science—Fueling Progress—Revolutionizing Care. *Clinical Cancer Research* **30**, 4296–4298 (2024).
32. Kipling, M. D. & Waldron, H. A. Percivall Pott and cancer scroti. *Occupational and Environmental Medicine* **32**, 244–246 (1975).
33. van de Vijver, M. J. The pathology of familial breast cancer: The pre-BRCA1/BRCA2 era - historical perspectives. *Breast Cancer Res* **1**, 27–30 (1999).
34. Boveri, T. *Zur Frage der Entstehung maligner Tumoren*. (Gustav Fischer, Jena, 1914).
35. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149–152 (1982).
36. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971).
37. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646 (1986).
38. Colby, W. W., Chen, E. Y., Smith, D. H. & Levinson, A. D. Identification and nucleotide sequence of a human locus homologous to the v-myc oncogene of avian myelocytomatosis virus MC29. *Nature* **301**, 722–725 (1983).
39. Velu, T. J. *et al.* Epidermal-Growth-Factor-Dependent Transformation by a Human EGF Receptor Proto-Oncogene. *Science* **238**, 1408–1410 (1987).
40. Nigro, J. M. *et al.* Mutations in the p53 gene occur in diverse human tumour types. *Nature* **342**, 705–708 (1989).
41. Li, J. *et al.* PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. *Science* **275**, 1943–1947 (1997).
42. Miki, Y. *et al.* A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science* **266**, 66–71 (1994).
43. Wooster, R. *et al.* Localization of a Breast Cancer Susceptibility Gene, BRCA2, to Chromosome 13q12-13. *Science* **265**, 2088–2090 (1994).
44. Laken, S. J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17**, 79–83 (1997).
45. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**,

860–921 (2001).

46. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
47. Collins, F. S. & Barker, A. D. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* **296**, 50–57 (2007).
48. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
49. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
50. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
51. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
52. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).
54. Xiao, W. *et al.* Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat Biotechnol* **39**, 1141–1150 (2021).
55. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nat Rev Genet* **23**, 298–314 (2022).
56. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).
57. Fan, Y. *et al.* Accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling for sequencing data. *bioRxiv* 055467 (2016) doi:10.1101/055467.
58. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
59. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

60. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).
61. Ally, A. *et al.* Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341.e23 (2017).
62. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
63. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
64. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *cels* **6**, 271-281.e7 (2018).
65. Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
66. Wood, D. E. *et al.* A machine learning approach for somatic mutation discovery. *Science Translational Medicine* **10**, eaar7939 (2018).
67. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *The New England journal of medicine* **375**, 1109–1112 (2016).
68. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6**, 10001 (2015).
69. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci Rep* **7**, 43169 (2017).
70. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623–630 (2015).
71. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
72. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
73. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics* **49**, 1785–1788 (2017).
74. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology* **17**, 128 (2016).
75. Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nature Cancer* **1**, 122–

135 (2020).

76. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**, 3719–3731 (2016).
77. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019).
78. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nature Genetics* **52**, 208–218 (2020).
79. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183 (2004).
80. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods* **10**, 1081–1082 (2013).
81. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
82. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
83. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).
84. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
85. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
86. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402–1407 (2015).
87. Pfeifer, G. P. Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* **301**, 259–281 (2006).
88. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
89. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology* **17**, 31 (2016).
90. Färkkilä, A. *et al.* Immunogenomic profiling determines responses to combined PARP and PD-1 inhibition in ovarian cancer. *Nat Commun* **11**, 1459 (2020).

91. André, T. *et al.* Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *New England Journal of Medicine* **383**, 2207–2218 (2020).
92. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat Genet* **51**, 1732–1740 (2019).
93. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat Rev Genet* **22**, 483–501 (2021).
94. de Klein, A. *et al.* A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **300**, 765–767 (1982).
95. Kantarjian, H. *et al.* Improved survival in chronic myeloid leukemia since the introduction of imatinib therapy: a single-institution historical experience. *Blood* **119**, 1981–1987 (2012).
96. Maemondo, M. *et al.* Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* **362**, 2380–2388 (2010).
97. Rosell, R. *et al.* Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol* **13**, 239–246 (2012).
98. Chapman, P. B. *et al.* Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine* **364**, 2507–2516 (2011).
99. Mansfield, E. A. FDA perspective on companion diagnostics: an evolving paradigm. *Clin Cancer Res* **20**, 1453–1457 (2014).
100. Redig, A. J. & Jänne, P. A. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J Clin Oncol* **33**, 975–977 (2015).
101. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine* **372**, 2509–2520 (2015).
102. Lemery, S., Keegan, P. & Pazdur, R. First FDA Approval Agnostic of Cancer Site — When a Biomarker Defines the Indication. *New England Journal of Medicine* **377**, 1409–1412 (2017).
103. Marabelle, A. *et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol* **21**, 1353–1365 (2020).
104. Drilon, A. *et al.* Efficacy of Larotrectinib in TRK Fusion–Positive Cancers in Adults and Children. *New England Journal of Medicine* **378**, 731–739 (2018).

105. Hyman, D. M. *et al.* Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *N Engl J Med* **373**, 726–736 (2015).
106. Kopetz, S. *et al.* Encorafenib, Binimetinib, and Cetuximab in BRAF V600E–Mutated Colorectal Cancer. *New England Journal of Medicine* **381**, 1632–1643 (2019).
107. Reardon, B. *et al.* Integrating molecular profiles into clinical frameworks through the Molecular Oncology Almanac to prospectively guide precision oncology. *Nat Cancer* **2**, 1102–1112 (2021).
108. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine* **10**, 25 (2018).
109. Jiménez-Santos, M. J. *et al.* PanDrugs2: prioritizing cancer therapies using integrated individual multi-omics data. *Nucleic Acids Research* **51**, W411–W418 (2023).
110. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, PO.17.00011 (2017).
111. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
112. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
113. Ostrand-Rosenberg, S. & Sinha, P. Myeloid-derived suppressor cells: linking inflammation and cancer. *J Immunol* **182**, 4499–4506 (2009).
114. Qian, B.-Z. & Pollard, J. W. Macrophage diversity enhances tumor progression and metastasis. *Cell* **141**, 39–51 (2010).
115. Mougiakakos, D., Choudhury, A., Lladser, A., Kiessling, R. & Johansson, C. C. Regulatory T cells in cancer. *Adv Cancer Res* **107**, 57–117 (2010).
116. Yang, L., Pang, Y. & Moses, H. L. TGF-beta and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends Immunol* **31**, 220–227 (2010).
117. Esteller, M. Epigenetics in Cancer. *New England Journal of Medicine* **358**, 1148–1159 (2008).
118. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
119. Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discov* **9**, 837–851 (2019).
120. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
121. Darwiche, N. Epigenetic mechanisms and the hallmarks of cancer: an intimate affair.

Am J Cancer Res **10**, 1954–1978 (2020).

122. Thienpont, B., Van Dyck, L. & Lambrechts, D. Tumors smother their epigenome. *Mol Cell Oncol* **3**, e1240549 (2016).
123. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
124. Lindner, P. *et al.* EMT transcription factor ZEB1 alters the epigenetic landscape of colorectal cancer cells. *Cell Death Dis* **11**, 147 (2020).
125. He, S. & Sharpless, N. E. Senescence in Health and Disease. *Cell* **169**, 1000–1011 (2017).
126. Faget, D. V., Ren, Q. & Stewart, S. A. Unmasking senescence: context-dependent effects of SASP in cancer. *Nat Rev Cancer* **19**, 439–453 (2019).
127. Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nat Med* **25**, 377–388 (2019).
128. Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A. & Wargo, J. A. The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell* **33**, 570–580 (2018).
129. Esteller, M. *et al.* The Epigenetic Hallmarks of Cancer. *Cancer Discovery* **14**, 1783–1809 (2024).
130. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
131. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).
132. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet* **37**, 1012–1027 (2021).
133. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320–331 (2010).
134. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
135. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590–607 (2019).
136. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **8**, a019505 (2016).
137. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation

signatures in human cells. *Nat Biotechnol* **27**, 361–368 (2009).

138. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: a landscape takes shape. *Cell* **128**, 635–638 (2007).
139. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
140. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399 (2016).
141. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827–1831 (1992).
142. Tost, J. & Gut, I. G. DNA methylation analysis by pyrosequencing. *Nat Protoc* **2**, 2265–2275 (2007).
143. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
144. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**, 468–481 (2011).
145. Karemaker, I. D. & Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol* **36**, 952–965 (2018).
146. Smith, Z. D., Hetzel, S. & Meissner, A. DNA methylation in mammalian development and disease. *Nat Rev Genet* 1–24 (2024) doi:10.1038/s41576-024-00760-8.
147. Fenaux, P. *et al.* Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. *Lancet Oncol* **10**, 223–232 (2009).
148. Kantarjian, H. *et al.* Decitabine improves patient outcomes in myelodysplastic syndromes: results of a phase III randomized study. *Cancer* **106**, 1794–1803 (2006).
149. Marks, P. A. & Breslow, R. Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. *Nat Biotechnol* **25**, 84–90 (2007).
150. Duvic, M. *et al.* Phase 2 trial of oral vorinostat (suberoylanilide hydroxamic acid, SAHA) for refractory cutaneous T-cell lymphoma (CTCL). *Blood* **109**, 31–39 (2007).
151. Whittaker, S. J. *et al.* Final results from a multicenter, international, pivotal study of romidepsin in refractory cutaneous T-cell lymphoma. *J Clin Oncol* **28**, 4485–4491 (2010).

152. Pagliaro, L. *et al.* Acute lymphoblastic leukaemia. *Nat Rev Dis Primers* **10**, 41 (2024).
153. Silkenstedt, E., Salles, G., Campo, E. & Dreyling, M. B-cell non-Hodgkin lymphomas. *Lancet* **403**, 1791–1807 (2024).
154. Ward, E., DeSantis, C., Robbins, A., Kohler, B. & Jemal, A. Childhood and adolescent cancer statistics, 2014. *CA Cancer J Clin* **64**, 83–103 (2014).
155. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. *New England Journal of Medicine* **373**, 1541–1552 (2015).
156. Pölönen, P. *et al.* The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature* **632**, 1082–1091 (2024).
157. Hetzel, S. *et al.* Acute lymphoblastic leukemia displays a distinct highly methylated genome. *Nat Cancer* **3**, 768–782 (2022).
158. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* **50**, 591–602 (2018).
159. Bensberg, M. *et al.* TET2 as a tumor suppressor and therapeutic target in T-cell acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **118**, e2110758118 (2021).
160. Maurer, M. J. *et al.* Event-Free Survival at 24 Months Is a Robust End Point for Disease-Related Outcome in Diffuse Large B-Cell Lymphoma Treated With Immunochemotherapy. *JCO* **32**, 1066–1073 (2014).
161. Maury, S. *et al.* Rituximab in B-Lineage Adult Acute Lymphoblastic Leukemia. *New England Journal of Medicine* **375**, 1044–1053 (2016).
162. Minard-Colin, V. *et al.* Rituximab for High-Risk, Mature B-Cell Non-Hodgkin's Lymphoma in Children. *N Engl J Med* **382**, 2207–2219 (2020).
163. Wynne, J., Wright, D. & Stock, W. Inotuzumab: from preclinical development to success in B-cell acute lymphoblastic leukemia. *Blood Adv* **3**, 96–104 (2019).
164. Caimi, P. F. *et al.* Loncastuximab tesirine in relapsed/refractory diffuse large B-cell lymphoma: long-term efficacy and safety from the phase II LOTIS-2 study. *Haematologica* **109**, 1184–1193 (2024).
165. Kantarjian, H. *et al.* Blinatumomab versus Chemotherapy for Advanced Acute Lymphoblastic Leukemia. *N Engl J Med* **376**, 836–847 (2017).
166. Hutchings, M. *et al.* Glofitamab, a Novel, Bivalent CD20-Targeting T-Cell-Engaging Bispecific Antibody, Induces Durable Complete Remissions in Relapsed or Refractory B-Cell Lymphoma: A Phase I Trial. *J Clin Oncol* **39**, 1959–1970 (2021).
167. Neelapu, S. S. *et al.* Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma. *N Engl J Med* **377**, 2531–2544 (2017).

168. Maude, S. L. *et al.* Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia. *N Engl J Med* **378**, 439–448 (2018).
169. Cappell, K. M. & Kochenderfer, J. N. Long-term outcomes following CAR T cell therapy: what we know so far. *Nat Rev Clin Oncol* **20**, 359–371 (2023).
170. June, C. H. & Sadelain, M. Chimeric Antigen Receptor Therapy. *N Engl J Med* **379**, 64–73 (2018).
171. Dias, J., Garcia, J., Agliardi, G. & Roddie, C. CAR-T cell manufacturing landscape-Lessons from the past decade and considerations for early clinical development. *Mol Ther Methods Clin Dev* **32**, 101250 (2024).
172. Ayala Ceja, M., Khericha, M., Harris, C. M., Puig-Saus, C. & Chen, Y. Y. CAR-T cell manufacturing: Major process parameters and next-generation strategies. *Journal of Experimental Medicine* **221**, e20230903 (2024).
173. Kochenderfer, J. N. *et al.* Eradication of B-lineage cells and regression of lymphoma in a patient treated with autologous T cells genetically engineered to recognize CD19. *Blood* **116**, 4099–4102 (2010).
174. Brentjens, R. J. *et al.* CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Sci Transl Med* **5**, 177ra38 (2013).
175. Schuster, S. J. *et al.* Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *N Engl J Med* **380**, 45–56 (2019).
176. Abramson, J. S. *et al.* Lisocabtagene maraleucel for patients with relapsed or refractory large B-cell lymphomas (TRANSCEND NHL 001): a multicentre seamless design study. *Lancet* **396**, 839–852 (2020).
177. Shah, B. D. *et al.* KTE-X19 for relapsed or refractory adult B-cell acute lymphoblastic leukaemia: phase 2 results of the single-arm, open-label, multicentre ZUMA-3 study. *Lancet* **398**, 491–502 (2021).
178. Brudno, J. N. & Kochenderfer, J. N. Recent advances in CAR T-cell toxicity: Mechanisms, manifestations and management. *Blood Rev* **34**, 45–55 (2019).
179. Kennedy, L. B. & Salama, A. K. S. A review of cancer immunotherapy toxicity. *CA Cancer J Clin* **70**, 86–104 (2020).
180. Morris, E. C., Neelapu, S. S., Giavridis, T. & Sadelain, M. Cytokine release syndrome and associated neurotoxicity in cancer immunotherapy. *Nat Rev Immunol* **22**, 85–96 (2022).
181. Chong, E. A., Ruella, M., Schuster, S. J., & Lymphoma Program Investigators at the

University of Pennsylvania. Five-Year Outcomes for Refractory B-Cell Lymphomas with CAR T-Cell Therapy. *N Engl J Med* **384**, 673–674 (2021).

182. Laetsch, T. W. *et al.* Three-Year Update of Tisagenlecleucel in Pediatric and Young Adult Patients With Relapsed/Refractory Acute Lymphoblastic Leukemia in the ELIANA Trial. *J Clin Oncol* **41**, 1664–1669 (2023).
183. Shah, B. D. *et al.* KTE-X19 anti-CD19 CAR T-cell therapy in adult relapsed/refractory acute lymphoblastic leukemia: ZUMA-3 phase 1 results. *Blood* **138**, 11–22 (2021).
184. Yang, J., Chen, Y., Jing, Y., Green, M. R. & Han, L. Advancing CAR T cell therapy through the use of multidimensional omics data. *Nat Rev Clin Oncol* **20**, 211–228 (2023).
185. Chi, H., Pepper, M. & Thomas, P. G. Principles and therapeutic applications of adaptive immunity. *Cell* **187**, 2052–2078 (2024).
186. Abdelsamed, H. A. *et al.* Human memory CD8 T cell effector potential is epigenetically preserved during in vivo homeostasis. *J Exp Med* **214**, 1593–1606 (2017).
187. Akondy, R. S. *et al.* Origin and differentiation of human memory CD8 T cells after vaccination. *Nature* **552**, 362–367 (2017).
188. Youngblood, B. *et al.* Effector CD8 T cells dedifferentiate into long-lived memory cells. *Nature* **552**, 404–409 (2017).
189. Ghoneim, H. E. *et al.* De Novo Epigenetic Programs Inhibit PD-1 Blockade-Mediated T Cell Rejuvenation. *Cell* **170**, 142-157.e19 (2017).
190. Henning, A. N., Roychoudhuri, R. & Restifo, N. P. Epigenetic control of CD8+ T cell differentiation. *Nat Rev Immunol* **18**, 340–356 (2018).
191. Guo, A. *et al.* cBAF complex components and MYC cooperate early in CD8+ T cell fate. *Nature* **607**, 135–141 (2022).
192. Zebley, C. C. *et al.* CD19-CAR T cells undergo exhaustion DNA methylation programming in patients with acute lymphoblastic leukemia. *Cell Reports* **37**, (2021).
193. Fraietta, J. A. *et al.* Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia. *Nat Med* **24**, 563–571 (2018).
194. Deng, Q. *et al.* Characteristics of anti-CD19 CAR T cell infusion products associated with efficacy and toxicity in patients with large B cell lymphomas. *Nat Med* **26**, 1878–1887 (2020).

195. Fraietta, J. A. *et al.* Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. *Nature* **558**, 307–312 (2018).
196. Carty, S. A. *et al.* The Loss of TET2 Promotes CD8+ T Cell Memory Differentiation. *J Immunol* **200**, 82–91 (2018).
197. Nowicki, T. S. *et al.* Epigenetic Suppression of Transgenic T-cell Receptor Expression via Gamma-Retroviral Vector Methylation in Adoptive Cell Transfer Therapy. *Cancer Discov* **10**, 1645–1653 (2020).
198. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727–733 (2020).
199. GBD 2021 Demographics Collaborators. Global age-sex-specific mortality, life expectancy, and population estimates in 204 countries and territories and 811 subnational locations, 1950-2021, and the impact of the COVID-19 pandemic: a comprehensive demographic analysis for the Global Burden of Disease Study 2021. *Lancet* **403**, 1989–2056 (2024).
200. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat Rev Microbiol* **21**, 133–146 (2023).
201. Collins, F. *et al.* The NIH-led research response to COVID-19. *Science* **379**, 441–444 (2023).
202. Del Rio, C. & Malani, P. N. COVID-19 in the Fall of 2023-Forgotten but Not Gone. *JAMA* **330**, 1517–1518 (2023).
203. Slaoui, M. & Hepburn, M. Developing Safe and Effective Covid Vaccines — Operation Warp Speed's Strategy and Approach. *New England Journal of Medicine* **383**, 1701–1703 (2020).
204. Lamers, M. M. & Haagmans, B. L. SARS-CoV-2 pathogenesis. *Nat Rev Microbiol* **20**, 270–284 (2022).
205. Pierce, C. A. *et al.* COVID-19 and children. *Science* **377**, 1144–1149 (2022).
206. Yoshida, M. *et al.* Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* **602**, 321–327 (2022).
207. Lu, X. *et al.* SARS-CoV-2 Infection in Children. *N Engl J Med* **382**, 1663–1665 (2020).
208. Belot, A., Levy-Bruhl, D., & French Covid-19 Pediatric Inflammation Consortium. Multisystem Inflammatory Syndrome in Children in the United States. *N Engl J Med* **383**, 1793–1794 (2020).
209. Levin, M. Childhood Multisystem Inflammatory Syndrome — A New Challenge in

the Pandemic. *New England Journal of Medicine* **383**, 393–395 (2020).

210. Sharma, C. *et al.* Multisystem inflammatory syndrome in children and Kawasaki disease: a critical comparison. *Nat Rev Rheumatol* **17**, 731–748 (2021).
211. Yousaf, A. R. *et al.* Notes from the Field: Surveillance for Multisystem Inflammatory Syndrome in Children - United States, 2023. *MMWR Morb Mortal Wkly Rep* **73**, 225–228 (2024).
212. Chou, J. *et al.* Mechanisms underlying genetic susceptibility to multisystem inflammatory syndrome in children (MIS-C). *J Allergy Clin Immunol* **148**, 732-738.e1 (2021).
213. Castro de Moura, M. *et al.* Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* **66**, 103339 (2021).
214. Moreews, M. *et al.* Polyclonal expansion of TCR Vbeta 21.3+ CD4+ and CD8+ T cells is a hallmark of Multisystem Inflammatory Syndrome in Children. *Sci Immunol* **6**, eabh1516 (2021).
215. Bodansky, A. *et al.* Molecular mimicry in multisystem inflammatory syndrome in children. *Nature* **632**, 622–629 (2024).
216. D'Agnillo, F. *et al.* Lung epithelial and endothelial damage, loss of tissue repair, inhibition of fibrinolysis, and cellular senescence in fatal COVID-19. *Sci Transl Med* **13**, eabj7790 (2021).
217. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114–119 (2021).
218. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med* **14**, 68 (2022).
219. Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat Rev Genet* **23**, 741–759 (2022).
220. Milross, L. *et al.* Post-mortem lung tissue: the fossil record of the pathophysiology and immunopathology of severe COVID-19. *Lancet Respir Med* **10**, 95–106 (2022).
221. Yue, L. *et al.* A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. *Comput Struct Biotechnol J* **21**, 940–955 (2023).
222. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* **40**, 661–671 (2022).
223. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
224. Dimitrov, D. *et al.* LIANA+ provides an all-in-one framework for cell–cell

communication inference. *Nat Cell Biol* **26**, 1613–1622 (2024).

225. Li, Z., Wang, T., Liu, P. & Huang, Y. SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nat Commun* **14**, 3995 (2023).
226. Tanevski, J., Flores, R. O. R., Gabor, A., Schapiro, D. & Saez-Rodriguez, J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol* **23**, 97 (2022).
227. Tomkova, M. *et al.* Human DNA polymerase ε is a source of C>T mutations at CpG dinucleotides. *Nat Genet* 1–11 (2024) doi:10.1038/s41588-024-01945-x.
228. McCutcheon, S. R., Rohm, D., Iglesias, N. & Gersbach, C. A. Epigenome editing technologies for discovery and medicine. *Nat Biotechnol* **42**, 1199–1217 (2024).
229. Salz, L. *et al.* Culture expansion of CAR T cells results in aberrant DNA methylation that is associated with adverse clinical outcome. *Leukemia* **37**, 1868–1878 (2023).
230. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* **22**, 276–282 (2012).
231. Mittelbrunn, M. & Kroemer, G. Hallmarks of T cell aging. *Nat Immunol* **22**, 687–698 (2021).
232. Durek, P. *et al.* Epigenomic Profiling of Human CD4+ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity* **45**, 1148–1161 (2016).
233. Passaro, A. *et al.* Cancer biomarkers: Emerging trends and clinical implications for personalized treatment. *Cell* **187**, 1617–1635 (2024).
234. Vokinger, K. N., Avorn, J. & Kesselheim, A. S. Sources of Innovation in Gene Therapies — Approaches to Achieving Affordable Prices. *New England Journal of Medicine* **388**, 292–295 (2023).
235. Juan, M., Delgado, J., Calvo, G., Trias, E. & Urbano-Ispizua, Á. Is Hospital Exemption an Alternative or a Bridge to European Medicines Agency for Developing Academic Chimeric Antigen Receptor T-Cell in Europe? Our Experience with ARI-0001. *Human Gene Therapy* **32**, 1004–1007 (2021).
236. Scott, C. L. & Omilusik, K. D. ZEBs: Novel Players in Immune Cell Development and Function. *Trends Immunol* **40**, 431–446 (2019).
237. Dominguez, C. X. *et al.* The transcription factors ZEB2 and T-bet cooperate to program cytotoxic T cell terminal differentiation in response to LCMV viral infection. *J Exp Med* **212**, 2041–2056 (2015).

238. Sharma, K. *et al.* Epigenetics in Kawasaki Disease. *Front Pediatr* **9**, 673294 (2021).
239. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* **110**, 12–22 (2019).
240. Vittinghoff, E. & McCulloch, C. E. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology* **165**, 710–718 (2007).
241. Arashi, M., Roozbeh, M., Hamzah, N. A. & Gasparini, M. Ridge regression and its applications in genetic studies. *PLOS ONE* **16**, e0245376 (2021).
242. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320 (2005).
243. Obermeyer, Z. & Emanuel, E. J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine* **375**, 1216–1219 (2016).
244. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **116**, 22071–22080 (2019).
245. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* **35**, 1925–1931 (2014).
246. Staffa, S. J., Stanek, K., Nasr, V. G. & Zurakowski, D. Five steps in performing machine learning for binary outcomes. *J Thorac Cardiovasc Surg* S0022-5223(24)00782–7 (2024) doi:10.1016/j.jtcvs.2024.08.048.
247. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206–215 (2019).
248. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56 (2019).
249. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nat Biotechnol* **36**, 829–838 (2018).
250. Almuntashiri, S. *et al.* TIMP-1 and its potential diagnostic and prognostic value in pulmonary diseases. *Chin Med J Pulm Crit Care Med* **1**, 67–76 (2023).
251. Pham, D. *et al.* Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat Commun* **14**, 7739 (2023).

252. Townes, F. W. & Engelhardt, B. E. Nonnegative spatial factorization applied to spatial genomics. *Nat Methods* **20**, 229–238 (2023).
253. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology* **17**, e9923 (2021).
254. Zhang, S. *et al.* Monitoring of cell-cell communication and contact history in mammals. *Science* **378**, eabo5503 (2022).
255. André, F., Rassy, E., Marabelle, A., Michiels, S. & Besse, B. Forget lung, breast or prostate cancer: why tumour naming needs to change. *Nature* **626**, 26–29 (2024).

List of abbreviations

List of abbreviations

AI	Artificial Intelligence.
ALL	acute lymphoblastic leukemia.
ARDS	acute respiratory distress syndrome.
B-ALL	B-cell acute lymphoblastic leukemia.
B-NHL	B-cell non-Hodgkin lymphoma.
CAR	chimeric antigen receptor.
CAVs	clinically actionable variants.
CGC	Cancer Gene Census.
COSMIC	Catalogue Of Somatic Mutations In Cancer.
COVID-19	coronavirus disease 2019.
CpG	cytosine followed by guanine dinucleotide.
CR	complete response.
CRISPR	clustered regularly interspaced short palindromic repeats.
CRS	cytokine release syndrome.
DAD	diffuse alveolar damage.
DLBCL	Diffuse large B-cell lymphoma.
dMMR	deficient mismatch repair.
DNA	Deoxyribonucleic acid.
DNMTs	DNA methyltransferases.
EMT	epithelial-to-mesenchymal transition.
FDA	Food and Drug Administration.
H3K27me3	trimethylation of lysine 27 on histone H3.
H3K36me3	trimethylation of lysine 36 on histone H3.
ICANS	immune effector cell-associated neurotoxicity syndrome.
ICGC	International Cancer Genome Consortium.
Indels	insertions and deletions.
KD	Kawasaki disease.
LRIs	ligand-receptor interactions.
MC3	Multi-Center Mutation Calling in Multiple Cancers.
MIS-C	multisystem inflammatory syndrome in children.
ML	machine learning.
MOAlmanac	Molecular Oncology Almanac.
mRNA	messenger RNA.
MRD	measurable residual disease.
MSI	microsatellite instability.
NGS	next-generation sequencing.
NHL	non-Hodgkin lymphoma.
NMF	non-negative matrix factorization.
PARP	poly (ADP-ribose) polymerase.

PCAWG	Pan-Cancer Analysis of Whole Genomes.
PD-1	programmed cell death protein 1.
pDCs	plasmacytoid dendritic cells.
qRT-PCR	quantitative reverse transcription–polymerase chain reaction.
ROC	receiver operating characteristic.
R/R	relapsed or refractory.
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2.
SBS	single base substitutions.
SNVs	single nucleotide variants.
ST	spatial transcriptomics.
T-ALL	T-cell acute lymphoblastic leukemia.
TCGA	The Cancer Genome Atlas.
TCR	T-cell receptor.
TF	transcription factor.
TMB	tumor mutational burden.
TME	tumor microenvironment.
UV	ultraviolet.
VAF	variant allele frequency.
WES	whole-exome sequencing.
WGS	whole-genome sequencing.
WHO	World Health Organization.
XGBoost	Extreme Gradient Boosting.

Annex

Annex

This section presents the papers I have published as first or co-first author during my doctoral program. While these publications are not included within the main scope of this thesis, they significantly contribute to my overall research portfolio:

1. Garcia-Prieto, C. A. *et al.* Validation of a DNA methylation microarray for 285,000 CpG sites in the mouse genome. *Epigenetics* **17**, 1677–1685 (2022).
2. Joshi, R. S. *et al.* Look-alike humans identified by facial recognition algorithms show genetic similarities. *Cell Reports* **40**, 111257 (2022).

Additionally, this annex includes a list of co-authored publications that reflect my contributions to various collaborative research projects, further underscoring my involvement in diverse scientific initiatives throughout my doctoral program.

EPIGENETICS
2022, VOL. 17, NO. 12, 1677–1685
https://doi.org/10.1080/15592294.2022.2053816

Taylor & Francis
Taylor & Francis Group

BRIEF REPORT

Validation of a DNA methylation microarray for 285,000 CpG sites in the mouse genome

Carlos A. Garcia-Prieto[a,b], Damiana Álvarez-Errico[a], Eva Musulen[a,c], Alberto Bueno-Costa[a], Berta N. Vazquez[d], Alejandro Vaquero[d], and Manel Esteller [a,e,f,g]

[a]Cancer Epigenetics Group, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Spain; [b]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain; [c]Department of Pathology, Hospital Universitari General de Catalunya Grupo-QuirónSalud, Barcelona, Spain; [d]Chromatin Biology Group, Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Spain; [e]Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain; [f]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain; [g]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Spain

ABSTRACT

Mouse has been extensively used as a model organism in many studies to characterize biological pathways and drug effects and to mimic human diseases. Similar DNA sequences between both species facilitate these types of experiments. However, much less is known about the mouse epigenome, particularly for DNA methylation. Progress in delivering mouse DNA methylomes has been slow due to the currently available time-consuming and expensive methodologies. Following the great acceptance of the human DNA methylation microarrays, we have herein validated a newly developed DNA methylation microarray (Infinium Mouse Methylation BeadChip) that interrogates 280,754 unique CpG sites within the mouse genome. The CpGs included in the platform cover CpG Islands, shores, shelves and open sea sequences, and loci surrounding transcription start sites and gene bodies. From a functional standpoint, mouse ENCODE representative DNase hypersensitivity sites (rDHSs) and candidate cis-Regulatory Elements (cCREs) are also included. Herein, we show that the profiled mouse DNA methylation microarray provides reliable values among technical replicates; matched results from fresh frozen versus formalin-fixed samples; detects hemimethylated X-chromosome and imprinted CpG sites; and is able to determine CpG methylation changes in mouse cell lines treated with a DNA demethylating agent or upon genetic disruption of a DNA methyltransferase. Most important, using unsupervised hierarchical clustering and t-SNE approaches, the platform is able to classify all types of normal mouse tissues and organs. These data underscore the great features of the assessed microarray to obtain comprehensive DNA methylation profiles of the mouse genome.

ARTICLE HISTORY
Received 18 February 2022
Revised 8 March 2022
Accepted 9 March 2022

KEYWORDS
Mouse; DNA methylation; microarray; epigenetics; CpG sites; validation

Background

Mice (*Mus musculus*) have been widely used as animal models in the biomedical field to interrogate different physiological pathways and to recapitulate human pathologies [1–3]. Many motives can be claimed for their utilization in the aforementioned studies, among them the overall low cost, efficient reproduction in a short time, easy manipulation, actionability to genetic engineering interventions, and the biological and structural commonalities to the *Homo sapiens*. In this regard, the less problematic generation of embryonic stem cells from mice, the controlled experimental environment, and the close similarity between the human and the rodent genome have further fostered the extensive application of mice models in many fields of life sciences, particularly in the translation to potential clinical applications [1–3]. Related to this last issue, most human clinical trials for new drugs have been preceded by comprehensive preclinical mouse studies to enlighten us about efficacy and toxicity of the new pharmacological compound [1–3]. Thus, despite the need to support and promote the use of non-animal approaches to validate mechanisms of actions in humans, mouse models continue to play a central role in many stages of biomedical research, including the understanding and development of new drugs for such devastating conditions, such as cancer or neurodegenerative diseases.

CONTACT Manel Esteller mesteller@carrerasresearch.org Josep Carreras Leukaemia Research Institute Carretera de Can Ruti, Camí de les Escoles s/n, Badalona, Barcelona 08916 Spain

Importantly, even though the mouse genome has been studied in large detail, we know little about the DNA methylation landscape of the mouse in comparison to humans. One apparent reason for the scarcity of mouse cancer epigenetic data is the unavailability of a reliable, versatile, and exchangeable tool between researchers around the world that allows the study of hundreds of samples in an objective precise manner and is comparable to the human DNA methylation microarrays genomic platform, The Cancer Genome Atlas (TCGA) Program, which has molecularly dissected most human tumour types (https://portal.gdc.cancer.gov/) [4,5]. Herein, we introduce a comprehensive mouse DNA methylation microarray, the Infinium Mouse Methylation BeadChip (Illumina Inc., CA, USA), that we have experimentally validated at both technical and biological levels and used to interrogate the epigenetic profile of normal murine tissues. Our work provides the first necessary demonstration of the great value of this platform to obtain an extensive view of the mouse DNA methylome that will open its general use to study mouse models of a diverse range of human diseases.

Results and Discussion

Genomic and functional classification of the over 285,000 probes in the mouse DNA methylation microarray

Bisulphite genomic sequencing provides a digital read of the CpG methylation status of a DNA sequence. This approach, associated with deep-sequencing chemistry, has made possible the obtention of a set of whole-genome bisulphite sequences for the human and mouse genomes [6–10]. However, delivering a full organism DNA methylome needs to take into account the import budgetary cost, time-consumption, and the need for complex bioinformatic analyses [11]. User-friendly DNA methylation microarrays have been developed following the example of the carefully annotated DNA microarrays with great genomic coverage used to detect SNPs in genome-wide association studies (GWAS). Those more commonly used were the Illumina Infinium HumanMethylation450 BeadChip 450,000 CpG site platform (450 K; Illumina Inc., CA, USA) [4] and its current updated version, the Methylation EPIC BeadChip (Infinium) microarray, which covers over 850,000 CpG methylation sites (850 K) [5]. These DNA methylation microarrays are the platforms selected for The Cancer Genome Atlas (TCGA) studies (https://portal.gdc.cancer.gov/) and the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/) but also for hundreds of other studies wondering about the DNA methylation profiles of human cells in distinct physiological and pathological conditions. The versatility of the described platforms has also been demonstrated by its use to obtain 5-methylcytosine DNA profiles from formalin-fixed paraffin-embedded samples [12]. These useful tools to study human DNA methylation, which in addition allow the easy exchange of data from scientists around the world and post-publication further data mining, did not exist until now for the mouse DNA methylome. Herein, we have validated from a biological and technical standpoint a comprehensive mouse DNA methylation microarray, termed Infinium Mouse Methylation BeadChip (Illumina Inc., CA, USA), as a new robust genomic platform that is available for the epigenetics community to characterize the mouse DNA methylome.

The recently developed Infinium Mouse Methylation BeadChip microarray interrogates the DNA methylation status of 280,754 unique CpG sites covering all chromosomes of the mouse genome (Figure 1a). In addition, it contains 642 control probes for quality control, 1352 genotyping probes for mouse strain, and 938 CpH probes ('H' meaning any nucleotide, except guanine). Most of the probes (78.3%) are Infinium II Probe Design that use only one probe per locus (one bead type for both alleles), whereas 21.7% of the probes were Infinium I Probe Design that utilizes two probes per SNP to assess the relative intensity ratio of the two possible target alleles for that locus (two bead types, one for each allele) (Figure 1a). According to the CpG content of the DNA region, 10.7% CpG sites were located in CpG Islands, 11.3% in CpG shores, 5% in CpG shelves, and 73% were placed in mouse genome sequences with very low CpG density (open sea) (Figure 1a). From a functional standpoint, 70.4% of the CpG

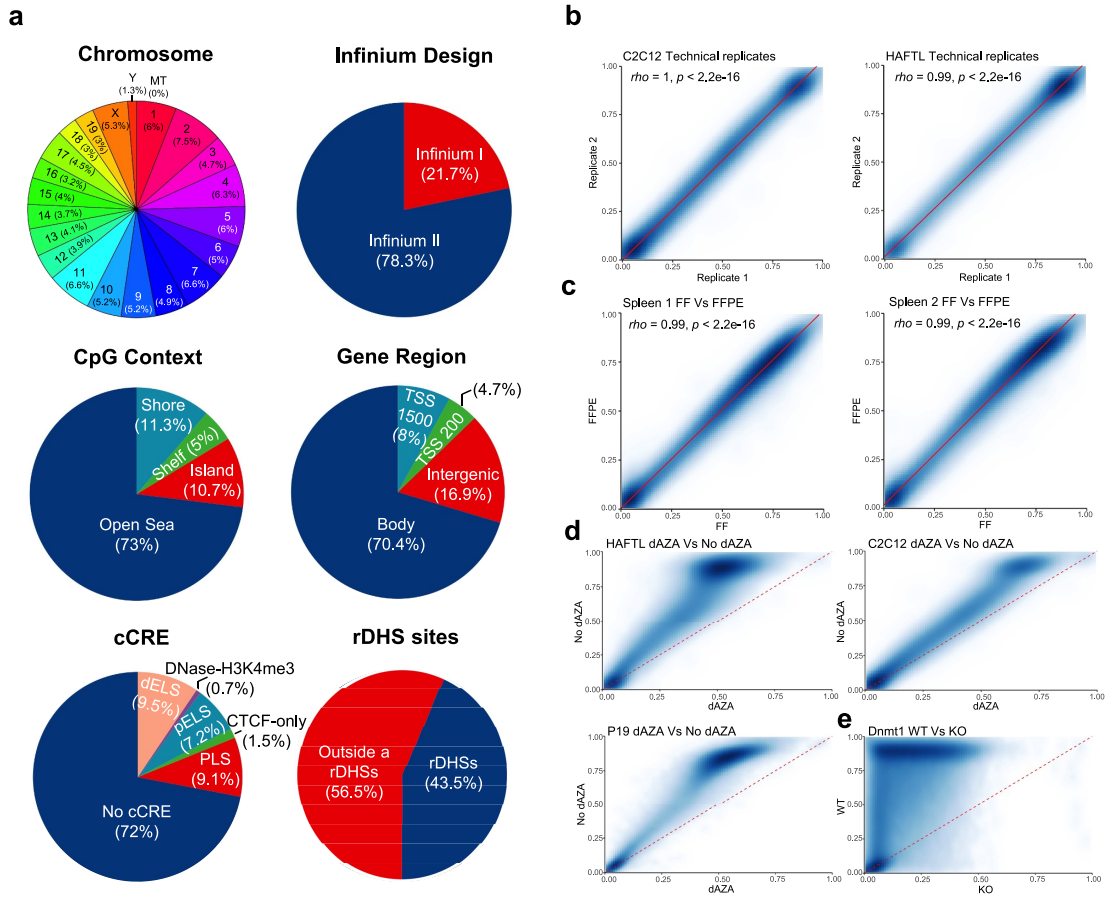


Figure 1. Description and technical and biological validation of the 285,000 CpG sites mouse DNA methylation microarray. (a) Genomic and functional context of the 280,754 CpG sites contained in the Infinium Mouse Methylation BeadChip microarray: Chromosome location; Infinium design chemistry (Infinium I or II) of the probes; CpG content and neighbourhood context classified in CpG Island, shore, shelf, and other (open sea); functional genomic distribution of the CpG sites classified in gene body, TSS200, TSS1500, and intergenic; distribution among ENCODE candidate cis-Regulatory Elements (cCREs) with promoter-like signature (PLS), with enhancer-like signature (ELS), with distal enhancer location (dELS), with high H3K4me3 and low H3K27ac signal (DNase-H3K4me3), and with CTCF-only elements; and association with an ENCODE representative DNase hypersensitive site (rDHS). (b) Correlation plot of the CpG methylation values to show assay reproducibility of the measurements when using technical replicates on the mouse cell lines C2C12 and HAFTL. (c) Spearman's correlation plot of the CpG methylation values obtained from two spleen fresh frozen (FF) samples when compared with their consecutive sections that were preserved as formalin-fixed paraffin-embedded (FFPE). (d) Correlation plot of the CpG methylation values to show DNA hypomethylation events in the mouse cell lines C2C12, HAFTL, and P19 upon the use of the demethylating agent 5-aza-2′-deoxycytidine (dAZA). (e) Correlation plot of the CpG methylation values to show DNA hypomethylation events in the mouse embryonic stem cells upon genetic knock-out of the maintenance DNA methyltransferase *Dnmt1*.

sites were located on gene bodies and 16.9% in intergenic regions, whereas those located in more classical 5-end regulatory regions such as Transcription Start Site 200 bp and Transcription Start Site 1,500 bp were 4.7% and 8%, respectively (Figure 1a). According to the mouse the Encyclopaedia of DNA Elements (ENCODE) project (http://www.mouseencode.org/), 56.5% of CpGs were located outside of representative DNase hypersensitivity sites (rDHSs), whereas 43.5% were in rDHSs regions (Figure 1a). Importantly, among the above-described CpG probes, there are annotated sites according to the ENCODE Registry of candidate cis-Regulatory

Elements (cCREs) corresponding to distal enhancer-like signature (dELS, 9.5%), promoter-like signature (PLS, 9.1%), proximal enhancer-like signature (pELS, 7.2%), CTCF-only elements (1.5%), and DNase-H3K4me3 elements (0.7%) that are those with promoter-like biochemical signature that are not within 200 bp of an annotated TSS (Figure 1a). Seventy-two per cent of the CpG sites were placed outside cCREs (Figure 1a). The genomic location along with structural and functional context for each CpG dinucleotide among the 280,754 CpG sites can be found at the manifest of the Infinium Mouse Methylation BeadChip (https://support.illumina.com/downloads/infinium-mouse-methylation-manifest-file.html).

Technical and biological validation of the mouse methylation BeadChip

Although the reproducibility of the Infinium Mouse Methylation BeadChip is mentioned on the manufacturer site (https://www.illumina.com/products/by-type/microarray-kits/infinium-mouse-methylation.html), we have herein confirmed its robustness and reliability using a comprehensive set of different technical, experimental, and biological models.

To demonstrate the capability of the Mouse Methylation BeadChip for the analysis of DNA methylation, we have developed several distinct methodological approaches. First, we obtained a technical validation of the mouse DNA methylation microarray data by performing replication experiments, where we hybridized the same samples twice, the mouse cell lines C2C12 (immortalized myoblasts) and HAFTL (pre-B cells), to the Mouse Methylation BeadChip. We observed that the methylation levels detected at CpG sites derived from each experiment were highly correlated and interchangeable (Figure 1b). Second, given the optimal performance of the human Infinium HumanMethylation450 BeadChip and MethylationEPIC BeadChip microarrays for formalin-fixed paraffin-embedded (FFPE) samples [12], we wondered about the robustness of the mouse microarray to determine the DNA methylation in this type of archival material. To address this point, we hybridized to the platform the same DNA samples from two normal mouse spleen samples obtained from consecutive fresh or FFPE sections from the same specimen. We found that the methylation levels assessed at each CpG site from each sample source were highly correlated (Figure 1c).

We then analysed the reliability of the Mouse Methylation BeadChip to detect CpG methylation changes using both drug and genetic approaches. For the pharmacological strategy, we treated the mouse cell lines C2C12 and HAFTL (both described above) and P19 (derived from an embryonal carcinoma induced in a C3H/He strain mouse) with the well-known inhibitor of DNA methylation 5-aza-2′-deoxycytidine. We observed that the use of the demethylating agent provoked widespread hypomethylation events in the described mouse cell lines (Figure 1d). Finally, we took advantage of the existence of mouse cells with deletion of the maintenance DNA methyltransferase *Dnmt1* [13] to further assess the capacity of the new mouse microarray to detect CpG methylation changes. We observed a profound hypomethylation landscape in the *Dnmt1* deficient cells in comparison to the wild-type (Figure 1e), as it has been previously reported [13]. All the above-described data demonstrate the idoneity of the studied DNA methylation microarray as a reliable epigenomic tool for biological and pathological studies that use mouse models.

A DNA methylation draft of mouse normal tissues

First, we interrogated the DNA methylation profiles for 56 samples corresponding to 11 normal mouse tissues or organs: lung, brain, prostate, breast, bone marrow, spleen, skin, colon, thymus, liver, and pancreas. Significantly distinct DNA methylation profiles were discovered between the different normal samples for all 226,000 CpG dinucleotides (after removal of erratic probe signals, X-chromosome sites, and genotyping probes) using multiscale bootstrap resampling (approximately unbiased p-value and bootstrap probability of 100% for all tissue type-specific clusters), which enabled their distinction on the basis of tissue type by the use of an unsupervised hierarchical clustering approach (Figure 2a). The above-described tissue type-specific DNA methylation classification

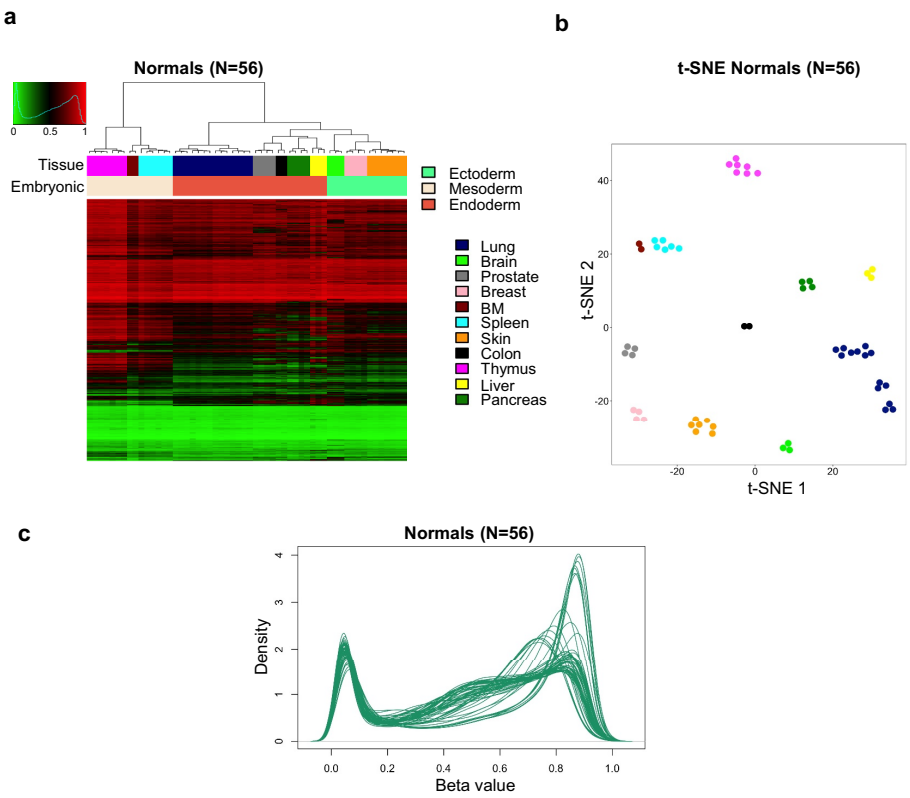


Figure 2. DNA methylation atlas for mouse normal tissues. (a) Unsupervised hierarchical clustering and heatmap for 56 normal primary samples from 11 distinct source types. Tissue type and development layers are shown in the distinct colours as described in the figure legends. Methylation values are displayed from 0 (green) to 1 (red). (b) DNA methylation variances between mouse normal tissues and organs are displayed as t-distributed stochastic neighbour embedding (t-SNE) of Beta values. (c) Density plot of methylation Beta values showing their distribution from 56 normal tissue samples for all 226,000 CpG dinucleotides that remain after removal of erratic probe signals, XY chromosomes probes, and genotyping probes.

also matched the developmental layers from which the tissues are derived (ectoderm, mesoderm, or endoderm) (Figure 2a), related to the presence of germ-layer-specific DNA methylation [14]. Dimensionality reduction analysis by t-SNE again yielded identical results clustering each mouse normal tissue and organ according to its DNA methylation profile (Figure 2b). Overall, the representation of the methylation content according to Beta value of the 226,000 CpG sites mostly shows a bimodal distribution with dinucleotides heavily methylated or largely hypomethylated (Figure 2c).

Significantly distinct DNA methylation profiles were discovered between male and female samples for all the CpG dinucleotides located at the X-chromosome (after removal of erratic probe signals) using multiscale bootstrap resampling (approximately unbiased p-value and bootstrap probability of 100% for all biological sex-specific clusters), which enabled their distinction on the basis of biological sex by the use of an unsupervised hierarchical clustering approach (Figure 3a). Dimensionality reduction analysis by t-SNE again produced similar results clustering each mouse's gender according to its DNA methylation profile (Figure 3b). As expected, the CpG sites of the microarray located in the X-chromosome exhibited around a 50% methylation content in the females (Figure 3c) due to well-known DNA methylation-dependent X-chromosome inactivation in that biological sex [15]. Importantly, the other CpG sites that displayed a 50% methylation content in normal tissues were those located in the differentially methylated regions (DMRs) of

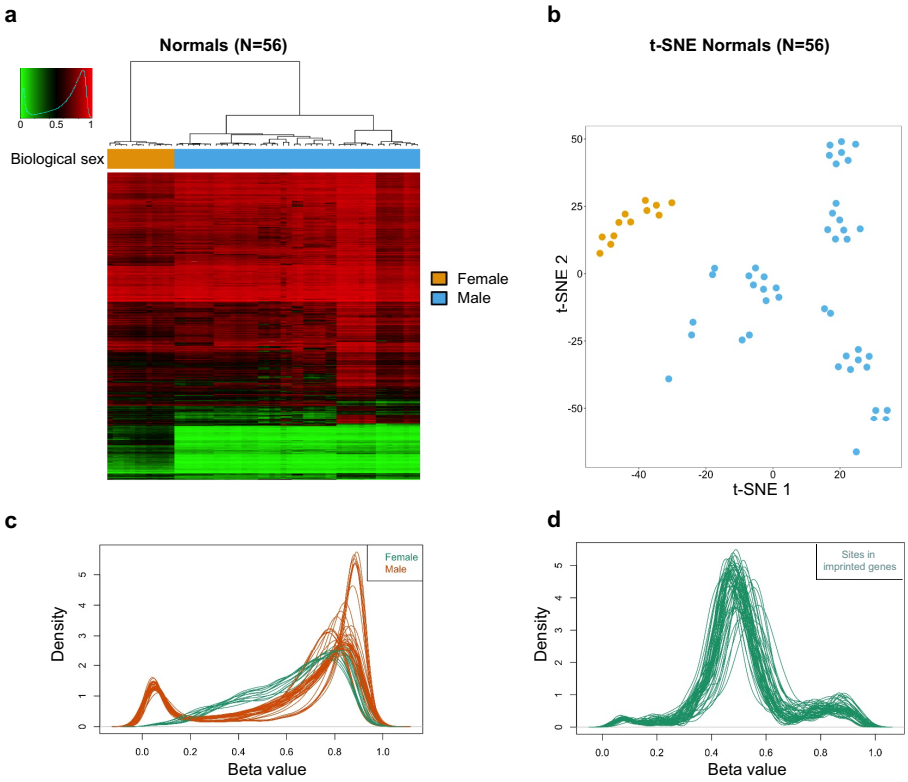


Figure 3. Mouse DNA methylation mapping according to biological sex, X-chromosome, and imprinted CpG sites. (a) Unsupervised hierarchical clustering and heatmap for 56 normal primary samples from 11 distinct source types according to CpG sites located in the X-chromosome. Methylation values are displayed from 0 (green) to 1 (red). (b) Biological sex type is shown in distinct colours as described in the figure legends. DNA methylation variances between female and male mouse samples are displayed as t-distributed stochastic neighbour embedding (t-SNE) of Beta values. (c) Density plot of methylation Beta values showing their distribution from 56 normal tissue samples for the CpG dinucleotides located at the X-chromosome. (d) Density plot of methylation Beta values showing their distribution from 56 normal tissue samples for the CpG dinucleotides located at imprinted genes.

mouse imprinted genes (Figure 3d), related to parentally determined monoallelic expression [16,17].

We also validated that the mouse genotyping probes (n = 1352) included in the microarray for different *Mus musculus* strains were indeed specific and informative. In this regard, significant distinct SNP profiles were discovered between the C57BL/6J, C57BL/6 × 129/Sv, FVB, and C57BL/6 × FVB strains using multiscale bootstrap resampling (approximately unbiased p-value and bootstrap probability of 100% for all strain-specific clusters), which allowed their classification on the basis of mouse strain by the use of an unsupervised hierarchical clustering approach (Figure 4a). Dimensionality reduction analysis by t-SNE revealed identical results clustering each mouse strain according to its SNP profile (Figure 4b).

Finally, and most importantly, we have deposited all the obtained mouse DNA methylation data in the open Gene Expression Omnibus (GEO) repository (accession GSE196902) to help fellow scientists in their ongoing and future studies to characterize the mouse DNA methylome in health and disease.

Conclusions

Herein, we have technically and biologically validated a comprehensive mouse DNA methylation microarray that we have used to interrogate the methylation status of 280,754 CpG sites in murine

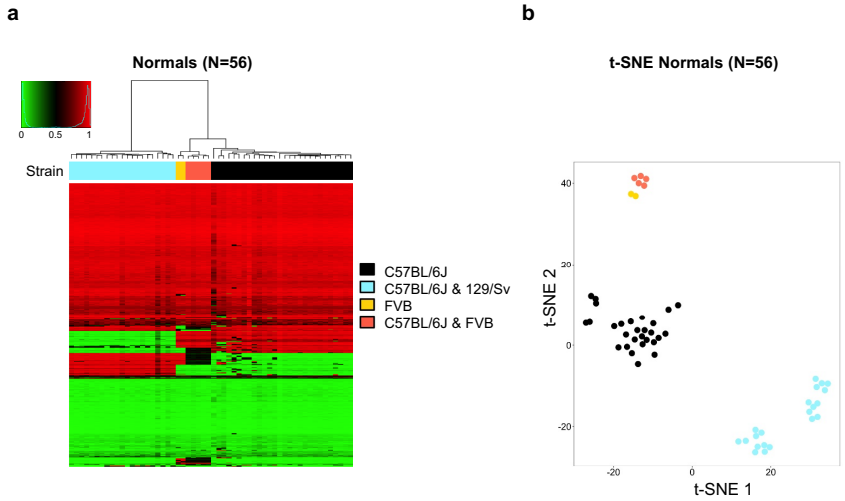


Figure 4. Mouse strains according to the genotyping probes of the DNA methylation microarray. (a) Unsupervised hierarchical clustering and heatmap for 56 normal primary samples from four different mouse strains according to the SNP sites included in the microarray. Mouse strain is shown in different colours as described in the figure legends. (b) SNP genotyping among the mouse strains is displayed as t-distributed stochastic neighbour embedding (t-SNE) of Beta values.

samples from primary samples and cell lines corresponding to eleven tissues and organs. This study represents one of the most extensive investigations into DNA methylation profiles within the mouse setting. The analysed platform has demonstrated its robustness and reliability in assessing DNA methylation patterns among replicates and in paraffin-embedded (FFPE), also being able to detect hypomethylation events caused by pharmacological and genetic interventions such as the use of the DNA demethylating agent 5-aza-2′-deoxycytidine and the analysis of *Dnmt1* deficient cells, respectively. Finally, the obtained DNA methylation patterns in the normal samples enable their clustering according to tissue type, organ, and germ layer.

Individual laboratory initiatives and the colossal effort of the ENCODE project have produced detailed mouse DNA methylomes for selected samples, particularly in the context of embryonic stem cells, foetal development, and adult normal tissues [9,10,18–22]. These landmark discoveries have provided reference mouse DNA methylomes by using Whole-Genome Bisulphite Sequencing (WGBS) that yields single-nucleotide resolution. WGBS is a very informative approach, but it is expensive, time-consuming, and requires a sophisticated bioinformatic pipeline. Thus, it is difficult to apply to the study of many samples in a user-friendly manner. In the human scenario, this has been solved by the introduction of DNA methylation microarrays where in its last inception, more than 850,000 functionally well defined and annotated CpG sites are included [5]. This methodology has been immensely popular due to its affordability and the easiness of the associated bioinformatic tools, making possible the study of the DNA methylation fingerprints of all types of tissues among different stages of differentiation, pathological samples from the cancer arena to the neurodegenerative field, ultimately opening the door to Epigenome Wide Association Studies (EWAS) that can include hundreds or thousands of samples for many human disorders, including COVID-19 [23]. This versatile tool to address all the above-described biological and disease-oriented projects was missing for the mouse species. The herein characterized DNA methylation microarray fills this void and most probably would be a '*trampoline*' for many studies in biology and

medical sciences focused on the mouse epigenome and its translation to the human context.

Methods

DNA isolation and DNA methylation profiling from mouse samples using universal bead arrays

DNA was isolated with the DNAeasy blood and tissue kit (Qiagen GmbH, Hilden, Germany) and ReliaPrep™ FFPE gDNA Miniprep System (Promega, Wisconsin, USA) for fresh frozen and formalin-fixed paraffin-embedded samples, respectively. C2C12, HAFTL, and P19 cell lines were cultured in 10 mL RPMI-1640 GlutaMAX, 10% FBS, and 1X Penicillin/Streptomycin and treated with 5-aza-2′-deoxycytidine (1 μM). Cells were plated in 25 cm^2 flasks, incubated at 5% CO_2 at 37°C, and harvested after 72 hours of culture. DNA from frozen pellets was purified using DNeasy Blood and Tissue Kit (Qiagen GmbH, Hilden, Germany). Purified genomic DNA was quantified with Qubit (Invitrogen, Carlsbad, CA, USA) according to manufacturer's instructions. Infinium Mouse Methylation BeadChip (Illumina, Inc., San Diego, CA, USA) arrays were used to analyse DNA methylation. This platform allows over 285,000 methylation sites per sample to be interrogated at single-nucleotide resolution. The samples were bisulphite converted using EZ DNA Methylation-Gold™ Kit (Zymo Research, CA, USA) and were hybridized in the array following the manufacturer's instructions.

DNA methylation data and computational analyses

The DNA methylation status of the studied samples was obtained using the Infinium Mouse Methylation BeadChip Array (~285,000 methylation sites). Raw signal intensities were assessed and analysed with GenomeStudio Software 2011.1 (Illumina). DNA methylation beta values were obtained from raw IDAT files with GenomeStudio default normalization using control probes and background subtraction. A number of quality control steps were applied to minimize errors and remove erratic probe signals. This involved removal of failed probes (probes with detection P value > 0.01) and manufacturing flagged (MFG) probes. XY chromosomes probes and genotyping probes were also removed for the DNA methylation analyses where the beta values of these probes were not required. The genomic analysis presented in the study was performed using the mm10 mouse genome reference build, as described in the Illumina manifest file associated with the Infinium Mouse Methylation BeadChip.

Unsupervised hierarchical clustering with 100 bootstrap replications was performed with R function package pvclust (v2.2-0). The Canberra distance scores and Ward's minimum variance method were applied to attain hierarchical clustering represented as a heatmap using the gplots (v3.1.1) package in R. t-Distributed stochastic neighbour embedding (t-SNE) was performed using R package M3C (v1.12.0). Density plots were performed with minfi (v1.36.0) package in R. Correlation plots and pie charts were performed using ggplot2 (v3.3.3) R package Quality control, and downstream analyses were performed within the R statistical environment (v4.0.3).

Author's Contributions

CAGP and ME conceived and designed the approach, interpreted the results, and wrote and revised the manuscript. DAE, EM, ABC, BNV, and AV provided experimental support. All authors have read and approved the final manuscript.

Disclosure statement

ME is a consultant for Ferrer International and Quimatryx. The remanining authors declare that they have no conflict of interest.

Funding

We thank the CERCA Programme/Generalitat de Catalunya for institutional support. This work was supported by the Health Department PERIS-project no. SLT/002/16/00374 and AGAUR-project no. 2017SGR1080 of the Catalan Government (Generalitat de Catalunya); Ministerio de Ciencia e Innovación (MCI), Agencia Estatal de Investigación (AEI), and European Regional Development

Fund (ERDF) project no. RTI2018-094049-B-I00 and PID2020-117284RB-I00; the Cellex Foundation; Marie Sklodowska-Curie Fellowship no. 895979 from the European Commission (BNV); and 'la Caixa' Banking Foundation (LCF/PR/GN18/51140001).

Data availability

The complete DNA methylation data are freely available on the GEO repository under accession number GSE196902.

ORCID

Manel Esteller http://orcid.org/0000-0003-4490-6093

References

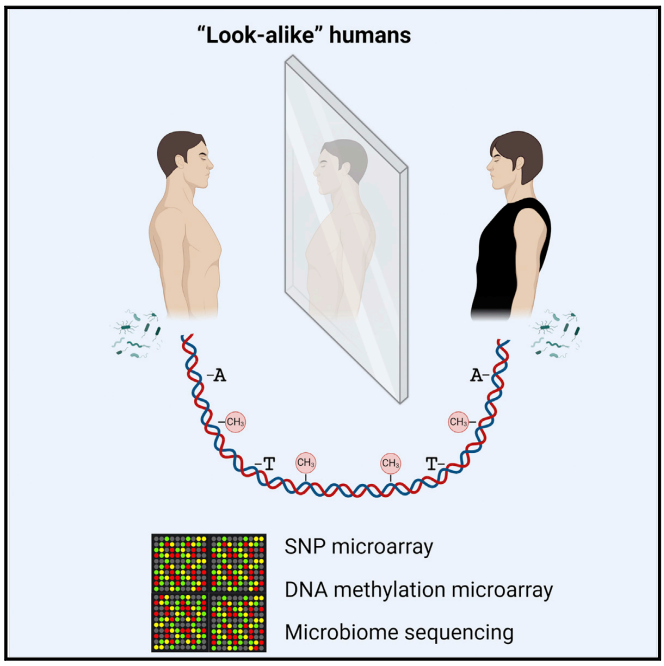
[1] Whitelaw CB, Sheets TP, Lillico SG, et al. Engineering large animal models of human disease. J Pathol. 2016;238:247–256.
[2] Gurumurthy CB, Lloyd KCK. Generating mouse models for biomedical research: technological advances. Dis Model Mech. 2019;12:dmm029462.
[3] Li H, Auwerx J. Mouse systems genetics as a prelude to precision medicine. Trends Genet. 2020;36:259–272.
[4] Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6:692–702.
[5] Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8:389–399.
[6] Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–322.
[7] Heyn H, Li N, Ferreira HJ, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A. 2012;109:10522–10527.
[8] Stunnenberg HG; Hirst M, Abrignani, S.; International Human Epigenome Consortium. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;167:1145–1149.
[9] He Y, Hariharan M, Gorkin DU, et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature. 2020;583:752–759.
[10] Moore JE, Purcaro MJ, Pratt, H E, et al.; ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.
[11] BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016;34:726–737.
[12] Moran S, Esteller M. Infinium DNA methylation microarrays on formalin-fixed, paraffin-embedded samples. Methods Mol Biol. 2018;1766:83–107.
[13] Jackson-Grusby L, Beard C, Possemato R, et al. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nat Genet. 2001;27:31–39.
[14] Sakamoto H, Suzuki M, Abe T, Hosoyama T, Himeno E, Tanaka S, Greally JM, Hattori N, Yagi S, Shiota K. Cell type-specific methylation profiles occurring disproportionately in CpG-less regions that delineate developmental similarity. Genes Cells. 2007;12:1123–1132.
[15] Escamilla-Del-Arenal M, da Rocha ST, Heard E. Evolutionary diversity and developmental regulation of X-chromosome inactivation. Hum Genet. 2011;130:307–327.
[16] Dindot SV, Person R, Strivens M, et al. Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. Genome Res. 2009;19:1374–1383.
[17] Monk D. Deciphering the cancer imprintome. Brief Funct Genomics. 2010;9:329–339.
[18] Stadler MB, Murr R, Burger L, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480:490–495.
[19] Hon GC, Rajagopal N, Shen Y, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013;45:1198–1206.
[20] Yue F, Cheng Y, Breschi A, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515:355–364.
[21] Bogdanović O, Smits AH, de la Calle Mustienes E, et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet. 2016;48:417–426.
[22] Ginno PA, Gaidatzis D, Feldmann A, et al. A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. Nat Commun. 2020;11:2680.
[23] Castro de Moura M, Davalos V, Planas-Serra L, et al. Epigenome-wide association study of COVID-19 severity with respiratory failure. EBioMedicine. 2021;66:10.

Report

Cell Reports

Look-alike humans identified by facial recognition algorithms show genetic similarities

Graphical abstract



Highlights

- Facial recognition algorithms identify "look-alike" humans for multiomics studies
- Intrapair look-alikes share common genetic sequences such as face trait variants
- DNA methylation and microbiome profiles only contribute modestly to human likeness
- The identified SNPs impact physical and behavioral phenotypes beyond facial features

Authors

Ricky S. Joshi, Maria Rigau, Carlos A. García-Prieto, ..., Xavier Binefa, Alfonso Valencia, Manel Esteller

Correspondence

mesteller@carrerasresearch.org

In brief

We recognize each other by relying on our face uniqueness. However, there are humans with uncanny resemblance. Joshi et al. reported that look-alike pairs identified by facial recognition algorithms share genotypes but not DNA methylomes and microbiomes. The identified SNPs also provide a readout of other anthropomorphic and behavioral characteristics.

Joshi et al., 2022, Cell Reports *40*, 111257
August 23, 2022
https://doi.org/10.1016/j.celrep.2022.111257

CellPress

Cell Reports

CellPress
OPEN ACCESS

Report

Look-alike humans identified by facial recognition algorithms show genetic similarities

Ricky S. Joshi,[1,10] Maria Rigau,[2,10] Carlos A. García-Prieto,[1,2,10] Manuel Castro de Moura,[1] David Piñeyro,[1,3] Sebastian Moran,[1] Veronica Davalos,[1] Pablo Carrión,[4] Manuel Ferrando-Bernal,[4] Iñigo Olalde,[4] Carles Lalueza-Fox,[4] Arcadi Navarro,[4,5,6] Carles Fernández-Tena,[7] Decky Aspandi,[8] Federico M. Sukno,[8] Xavier Binefa,[8] Alfonso Valencia,[2,6] and Manel Esteller[1,3,6,9,11,*]

[1]Josep Carreras Leukaemia Research Institute (IJC), Badalona, 08916 Barcelona, Spain
[2]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
[3]Centro de Investigacion Biomedica en Red Cancer (CIBERONC), 28029 Madrid, Spain
[4]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[5]Centre for Genomic Regulation (CNAG-CRG), 08003 Barcelona, Catalonia, Spain
[6]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain
[7]Herta Security, S.L., 08037 Barcelona, Spain
[8]Departament de Tecnologies de la Informació i les Comunicaciones (DTIC), Universitat Pompeu Fabra (UPF), 08018 Barcelona, Spain
[9]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), L'Hospitalet, 08907 Barcelona, Spain
[10]These authors contributed equally
[11]Lead contact
*Correspondence: mesteller@carrerasresearch.org
https://doi.org/10.1016/j.celrep.2022.111257

SUMMARY

The human face is one of the most visible features of our unique identity as individuals. Interestingly, monozygotic twins share almost identical facial traits and the same DNA sequence but could exhibit differences in other biometrical parameters. The expansion of the world wide web and the possibility to exchange pictures of humans across the planet has increased the number of people identified online as virtual twins or doubles that are not family related. Herein, we have characterized in detail a set of "look-alike" humans, defined by facial recognition algorithms, for their multiomics landscape. We report that these individuals share similar genotypes and differ in their DNA methylation and microbiome landscape. These results not only provide insights about the genetics that determine our face but also might have implications for the establishment of other human anthropometric properties and even personality characteristics.

INTRODUCTION

The discussion about the relevance of "nature versus nurture," or, in a similar manner, of "genotype versus phenotype," in human biology and medicine is a long-standing issue that still remains largely unsolved. Relevant studies in this area include our original observation that monozygotic twins show epigenetic differences (Fraga et al., 2005), understood as the chemical marks such as DNA methylation and histone modifications that regulate gene expression, that might explain different population traits and distinct penetrance of diseases in these people, a finding supported in later studies (Kaminsky et al., 2009), including The NASA Twins Study (Garrett-Bakelman et al., 2019). These questions can be more easily addressed in experimental models where the researcher can intervene, such as the Agouti mice (Wolff et al., 1998) and cloned animals (Rideout et al., 2001), whereas in humans, the investigator has a more passive role, waiting for the right sample to appear. In this regard, one of the most documented cases is the Dutch famine at the end of WWII that was associated with less DNA methylation of the imprinted *IGF2* gene compared with their unexposed, same-sex siblings (Heijmans et al., 2008).

Human individual identity also relates to biological properties and environment. In this regard, the way we initially recognize each other relies often on our unique face, and there is a sophisticated brain code to distinguish facial identities (Tsao et al., 2006; Chang and Tsao, 2017; Quian Quiroga, 2017). This explains why so commonly twins catch our attention and are used to understand how the balance between nature and nurture generates a phenotype. Here, we present a study that, on a molecular level, aims to characterize random human beings that objectively share facial features. This extraordinary set of individuals, characterized by their high likeliness, are what are called, in lay-language, look-alike humans, unknown twins, twin strangers, doubles, or doppelgänger, in German. This unique set of samples has allowed us to study how genomics,

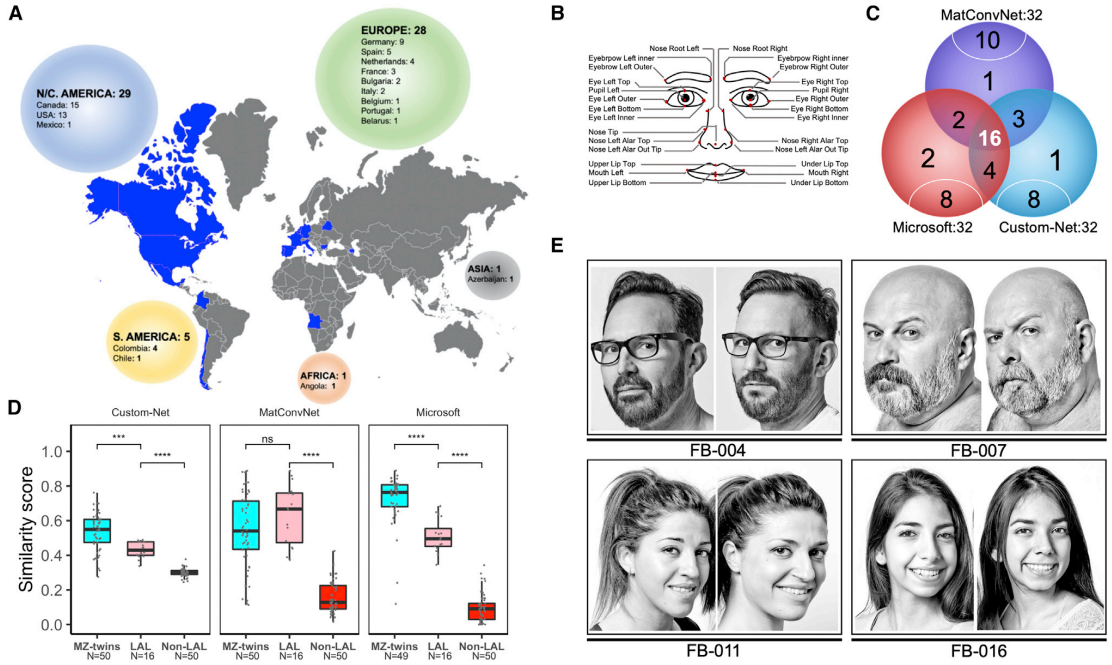


Figure 1. Recruitment and objective determination of look-alike human pairs

(A) Representation of the global worldwide distribution of 32 look-alike pairs (n = 64) in this study.

(B) 27 facial parameters by which the Microsoft Oxford Project face API (Microsoft) objectively performs face detection.

(C) Venn diagram showing the number of look-alike pairs discerned and jointly identified in the three facial recognition programs: MatConvNet, Custom-Net, and Microsoft. Numbers within the semi-circle present the pairs that did not cluster in each software.

(D) Boxplots showing unbiased quantitative similarity scores comparing each facial recognition software (MatConvNet, Custom-Net, Microsoft) for monozygotic twins (MZs; blue), look-alike pairs (LALs; rose), and random non-LALs (red). The x axis represents the different cohorts analyzed. The y axis exhibits similarity scores measured between 0 and 1. 1 represents identical facial image, and 0 represents two totally different photographic entities. "N" indicates the number of couples. Differences calculated using two-sided Mann-Whitney-Wilcoxon test: ****$p < 0.0001$; ***$p < 0.001$; ns, non-significant.

(E) Photographic examples of LALs used in this study.

epigenomics, and microbiomics can contribute to human resemblance. Our study provides a rare insight into human likeness by showing that people with extreme look-alike faces share common genotypes, whereas they are discordant at their epigenome and microbiome. Genomics clusters them together, and the rest set them apart. These findings do not only provide clues about the genetic setting associated with our facial aspect, and probably other traits of our body and personality, but also highlight how much of what we are, and what defines us, is really inherited or instead is acquired during our lifetime.

RESULTS

Facial recognition algorithms and multiomics approaches for look-alike humans

Human doubles were recruited from the photographic work of François Brunelle, a Canadian artist who has been obtaining worldwide pictures of look-alikes since 1999 (http://www.francoisbrunelle.com/webn/e-project.html). We obtained headshot pictures of thirty-two candidate look-alike couples. All participants completed a comprehensive biometric and lifestyle questionnaire in their native language (English, Spanish, and French) (Methods S1). Their geographic locations are shown in Figure 1A. We first determined an objective measure of "likeness" for the candidate double pairs. We used three different methods of facial recognition: the custom deep convolutional neural network Custom-Net, (www.hertasecurity.com), the MatConvNet algorithm (Vedaldi and Lenc 2015), and the Microsoft Oxford Project face API (https://azure.microsoft.com/es-es/services/cognitive-services/face/) (STAR Methods). We used three methods because each system can yield variable results, and we selected those systems to reflect the diversity of possible outcomes. MatConvNet was designed for facial classification, Custom-Net for surveillance, and Microsoft API for generalized facial analysis. These models have millions of learned parameters and have been trained with millions of facial images from thousands of subjects, in a variety of unconstrained situations: differences in pose, hairstyle, expression, age, and accessories within a subject. Thus, the impact of these attributes is likely minimal. Each software provides a facial similarity score between

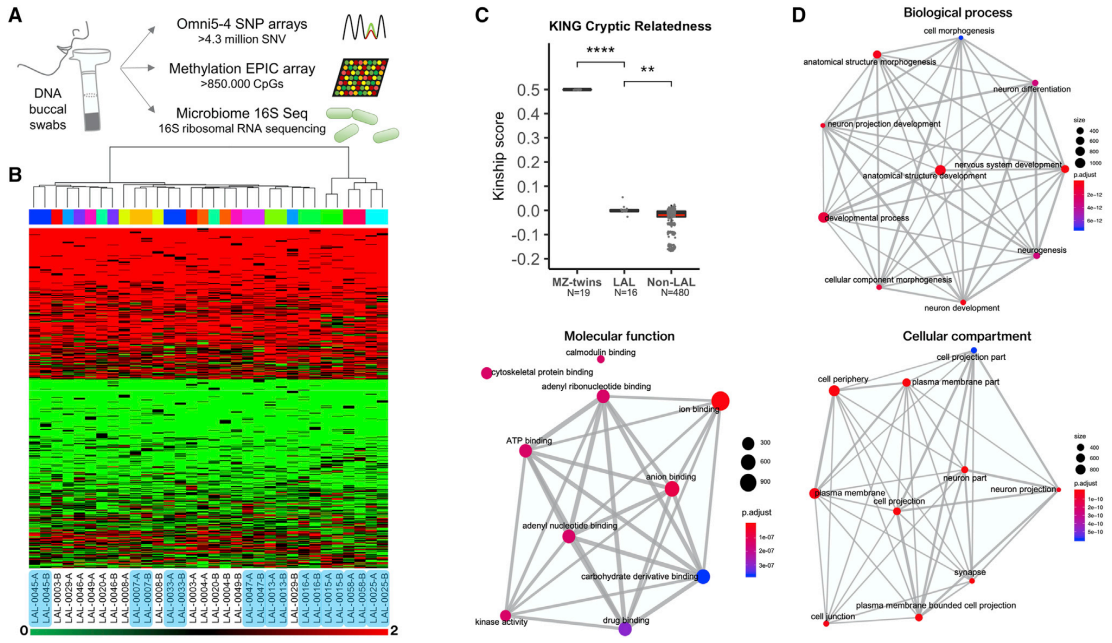


Figure 2. Genetic analysis of look-alike human pairs

(A) Saliva DNA was obtained from 32 LALs recruited to this study. DNA was subjected to genotyping (Omni5-4 SNP arrays Illumina), DNA methylation (Infinium MethylationEPIC arrays, Illumina), and microbiome analysis (16S Metagenomics sequencing, Illumina).

(B) Heatmap of hierarchical genetic clustering with bootstrap of genome-wide SNP genotyping arrays in the 16 LALs. Genotype clustering was performed using Euclidean distances and Ward.D2 cluster method. Blue rectangles represent 9 LALs that unbiasedly clustered. 0 = homozygous reference SNPs (green), 1 = heterozygous SNPs (black), and 2 = homozygous alternate SNPs (red).

(C) Boxplot showing Kinship scores between MZs, LALs, and random non-LALs. Kinship scores range between −0.2 (it represents two unrelated individuals) and 0.5 (it represents duplicated genotypes and MZs). "N" indicates the number of couples. Differences calculated using two-sided Student's t test: ****$p < 0.0001$; **$p < 0.01$.

(D) Gene Ontology (GO) analysis performed using all SNPs found to be shared in all LALs (19,277 SNPs in 3,730 genes). GO enrichments were ran using EnrichGO R package for the 3,730 genes, and the top 10 most significant hits are plotted in network graphs. GO terms are presented with circles. The size and color of each circle represents numbers of genes in each GO term and its statistical significance, respectively. The gray lines represent the interaction of genes, and the thickness is proportional to the number of genes interacting in each GO term. GO subcategories are presented: Biological Process, Cellular Component, and Molecular Function.

0 and 1, where 1 is the same facial image and 0 is two different entities. Comparisons are pairwise, with every image compared with every other image. As an example of the parameters computed, the 27 face landmarks of the Microsoft algorithm are shown in Figure 1B. The results obtained from the different combinations of each approach are shown in a Venn diagram in Figure 1C. Interestingly, the number of pairs that were considered to be correlated by at least two of the facial models was very high (25 out of total 32, >75%), closer to the human ability to recognize identical twins (Biswas et al., 2011). Most importantly, we found that 16 of the original 32 (50%) look-alike pairs were matched by all three facial recognition systems. As an internal positive control for high similarity score, we ran the three facial recognition software in monozygotic twin photograph images from the University of Notre Dame twins database 2009/2010 (https://cvrl.nd.edu/projects/data/). Importantly, similarity scores from the 16 look-alike couples were similar to those obtained from monozygotic twins according to MatConvNet and significantly higher than those observed in random non-look-alike pairs (Figure 1D). Thus, these highly look-alike humans were the focus of our further research. Illustrative examples of these "double" individuals are shown in Figure 1E.

Saliva DNA for these cases was analyzed by multiomics at three levels of biological information: genome, by means of an SNP microarray that interrogates 4,327,108 genetic variants selected from the International HapMap and 1,000 Genomes Projects, which target genetic variation down to 1% minor allele frequency (MAF) (Xing et al., 2016); epigenome, using a DNA methylation microarray that studies over 850,000 CpG sites (Moran et al., 2016); and microbiome, by ribosomal RNA direct sequencing (Klindworth et al., 2013) (Figure 2A; STAR Methods).

Genomic characterization of look-alike humans

Genomic analyses of these 16 couples provided a striking result: more than half (9 of 16, 56.2%) of these look-alike pairs clustered together in the unsupervised clustering heatmap with bootstrap

(Figure 2B). These nine couples were denominated as "ultra" look-alike. K-means algorithm represented by principal-component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) also showed that the look-alike couples that clustered by the unsupervised clustering heatmap analysis were in close proximity (Figure S1), indicating a likely genotyping resemblance of the studied pairs. In contrast, the 16 candidate look-alike cases that did not cluster by the three facial recognition (FR) networks (Figure 1C) showed that only one pair clustered together (1 of 16, 6.2%) (Figure S1).

We studied two possible confounding factors: population stratification (ancestry) and kinship. Using KING Relationship Inference (Manichaikul et al., 2010) to determine kinship scores, we discarded the possibility of unknown familial relationships (first and second degree) between look-alike pairs (Figure 2C). We observed that look-alike pairs were more similar to non-look-alike pairs than to monozygotic twins (Figure 2C); supporting that look-alike pairing in the SNP clustering is not related to familyhood genotype but instead to a distinct subset of genetic similarity. Using PLINK (Purcell et al., 2007) (STAR Methods), close kinship could be excluded in almost all cases: only one pair share SNPs in proportions that could be compatible with up to third-degree relatives and only one pair share a long (>10 cM) identity by descent (IBD) segment that could suggest co-ancestry in the last few hundreds of years. Interestingly, the latter is a French-Canadian pair, a population known to have experienced a dramatic founder effect in the 17th century. Importantly, when we conducted all the downstream analyses without this French-Canadian pair, the remaining eight ultra-look-alike pairs clustered together (Figure S2). The detailed kinship assessment data are provided in Table S1.

Related to population stratification, among the 16 look-alike pairs, 13 were of European ancestry, 1 Hispanic, 1 East Asian, and 1 Central-South Asian. Although background genetic ancestry is a principal determinant for genetic variance between human populations, we observed that of the 13 White look-alike pairs, 7 (54%) did not cluster genetically, suggesting alternative purposes for shared genetic variation between look-alike pairs. To further determine ancestry, genotyping of the 16 look-alike cases was performed using GenomeStudio v.2.0.5 to create PACKPED Plink files (STAR Methods). Their genomic data were merged with 1,980 West Eurasian, Asian, and Native American individuals genotyped in the Affymetrix Human Origins (HO) array (Lazaridis et al., 2014), where the remaining dataset held 175,469 common SNPs. PCA was generated with the HO individuals (Figure S3) and look-alike individuals (Figure S3B for West Eurasia and Figure S3C for West Eurasia, Asia, and America) (Price et al., 2006; Patterson et al., 2006) (STAR Methods). We observed that almost all the look-alike pairs cluster close to each other according to their countries of origin (or self-attributed ethnic background) (Figure S3). However, they are not more closely related than other pairs of individuals from the same populations taken at random. The detailed population stratification data are provided in Table S1.

Among the 9 couples of ultra-look-alikes, 19,277 SNP positions annotated for 3,730 genes (Table S2) were defined as SNPs with shared genotypes in each look-alike pair. These SNPs correspond to non-monomorphic positions in which every pair of ultra-look-alikes shared the genotype. For example, where one individual in a pair was heterozygous for a given SNP, the corresponding individual in the pair was also heterozygous. This genotype match must be consistent across all pairs for an SNP to be considered shared and therefore represented indicative SNPs relevant for look-alike resemblance. The number of shared SNP positions was significantly higher compared with random non-look-alike pairs in the studied population ($p < 2.2 \times 10^{-16}$, Pearson's chi-squared test). Taking into account ethnicity, shared SNP positions by the European ultra-look-alike pairs was significantly higher compared with random non-look-alike pairs in the studied population ($p = 0.03$, Pearson's chi-squared test). For the remaining three ethnicities, only one individual from each group was available in our dataset. Thus, we interrogated the individuals genotyped in the 1000 Genomes database (https://www.internationalgenome.org/). The number of shared SNP positions by the Hispanic ultra-look-alike pair was significantly higher compared with random individual pairs from the same ethnicity ($p < 2.2 \times 10^{-16}$, Pearson's chi-squared test). No significant enrichment was observed for the remaining two couples, one East Asian and one Central-South Asian. Importantly, only 16 variants of the 19,277 SNPs (0.08%) selected from the ultra-look-alikes presented a linkage disequilibrium detected by iterative pruning analysis (Weir et al., 2014).

The identified genetic variants might have a profound impact on the degree of similitude between the phenotype of humans. Using the clusterProfiler R package (Yu et al., 2012), we performed gene enrichment analyses using the list of look-alike SNPs compared with the background of all genes annotated in the SNP microarray. We observed an enrichment for Gene Ontology (GO) Biological Processes related to anatomical, developmental, and adhesion terms (Figure 2D; Table S3), in addition to ion and anion binding for GO-Molecular function (gene subsets related to bone and skin properties) and many cellular compartments. Enrichment analysis using the DAVID signature database collection noted that the most significantly enhanced ontology was "cell junction," a critical determinant of tissue morphology (Table S4). To evaluate the face genes enrichment in our selected 19,277 SNPs corresponding to 3,730 genes (Table S2), we gather all the genes related with face traits from recent data (Claes et al., 2018; Xiong et al., 2019; White et al., 2021), Facebase dataset (https://www.facebase.org/), and Genome-wide Association Study (GWAS) Central (study HGVST1841, http://www.gwascentral.org) and applied a hypergeometric test and a Monte Carlo simulation using 10,000 iterations (STAR Methods). In no iteration of random set of genes did we observe a number equal to or higher than the face genes represented in our 19,277 SNP selection ($p < 1e{-}4$). We observed a total of 1,794 face genes in our 19,277 SNP selection, constituting 26% of all the face genes present in the array (hypergeometric test p: 6.31e−172; Monte Carlo empirical $p < 1e{-}4$). When we added the reported face associated SNPs to our 19,277 SNPs, we observed that 11 of the 16 (68.7%) look-alike pairs clustered together (Figure S4), therefore adding two new couples.

The study of the functional nature of the SNPs loci shared by the ultra-look-alikes showed that 171 caused amino acid

changes, affecting 158 genes (Table S5). GOrilla analysis for GO-Molecular function found an enrichment in anion transport descriptors (Table S3). Using the GWAS catalog database (https://www.ebi.ac.uk/gwas/), we found that 113 SNPs corresponded to 130 GWAS associations and 84 traits (Table S6). These last traits included many related to facial determinants or physical features such as cleft palate/lip, eye color, hip circumference, body height, waist-hip ratio, balding measurement, and alopecia (Table S6) with an enrichment for lip and forehead morphology, body mass index, bone mineral density, and attached earlobe (Table S6). We observed an enrichment of traits that included the word morphologytagged to the terms nose, lip, mouth, facial, cranial vault, forehead, hair, and cheekbone (Fisher's exact test, odds ratio [OR] = 4.2, p = 0.04). Using the GWAS Central database (http://www.gwascentral.org), we found an enrichment (OR = 1.2782, p = 0.0007364) for SNPs associated with human facial variation (Adhikari et al., 2016). The analyses of the look-alike SNPs according to trait in GWAS Central showed an enrichment for the phenotype names "lip" (OR = 1.8321, p = 0.000327) and "forehead" (OR = 1.886, p = 0.010389). The identified look-alike SNPs were also enriched (OR = 2.201156, p = 0.04884) for genes included in the FaceBase dataset (https://www.facebase.org/). Finally, we studied the overlap between the herein discovered look-alike SNPs and expression quantitative trait loci (eQTLs). Using the Genotype-Tissue Expression (GTEx) Portal (https://www.gtexportal.org/home/), we observed that look-alike SNPs were more frequently associated with gene-expression changes than expected by random chance (Fisher's exact test, OR = 1.1, p = 0.0001). The enrichment was observed among different morphological structures and organs (Table S6). We also used the stratified linkage disequilibrium score regression (S-LDSC) (Finucane et al., 2015) to determine the enrichment of GWAS signals from the GWAS catalog for our SNPs. We observed that these SNPs were overrepresented for the pronasale-right chelion (enrichment score [ES] = 13.84, p = 0.018) and pronasale-left chelion (ES = 12.26, p = 0.04) face traits (Figure S4) (Xiong et al., 2019). The SNPs were also overrepresented for features that define 63 facial segments (Hoskens et al., 2021) considering the entire, mid, and outer face ($p < 0.05$) (Figure S4). These data indicate that the 19,277 characterized SNPs exert a major impact in the way the face of humans is defined.

The SNP microarray can also be used to determine copy-number variations (CNVs) (Feber et al., 2014). Unsupervised clustering heatmap with bootstrap clustered only one couple together of the 16 look-alikes according to CNVs (Figure 3A). Interestingly, three CNVs were shared by three look-alike pairs (Table S6), including a locus in chromosome 11 that targets genes involved in craniofacial dysmorphic features such as HYLS1 (Mee et al., 2005).

Other multiomics views of look-alike humans

Similar "identities" of look-alikes could also reside in other "omic" components such as the DNA methylome and the microbiome. According to DNA methylation patterns, only one of the sixteen (6.25%) look-alike pairs matched both individuals together, as shown in the unsupervised clustering heatmap (Figure 3B). This couple also clustered together according to SNP genotyping (Figure 2B). The comparison of DNA methylation patterns among the nine look-alike couples with the observed genetic overlap (Figure 2B) only clustered one additional pair (Figure S4). K-means algorithm represented using PCA and the t-SNE plot did not show significant clustering (Figure S5). Thus, overall, human look-alikes are diverse in their epigenome settings.

However, two avenues might provide a role for DNA methylation in facial morphology: epigenetic age and methylation QTL (meQTLs). The aging process changes facial morphology, and DNA methylation is used as a proxy for "biological age" that can or can not be directly related to the "chronological age." One example is the premature epigenetic aging observed in carriers of viral infections (Esteban-Cantos et al., 2021; Cao et al., 2022). We have calculated the intrapair absolute age differences in our 16 look-alike cohort according to chronological age (date of birth) or epigenetic age (DNA methylation clock) (Hannum et al., 2013). We found no differences in intrapair chronological age between the ultra-look-alike group and the non-ultra-look-alike group. In contrast, intrapair "epigenetic" age differences were smaller among ultra-look-alike pairs compared with the non-ultra-look-alike group (two-sided Mann-Whitney-Wilcoxon test, p = 0.0052) (Figure S6). DNA methylation is also associated with genetic variation (Villicaña and Bell, 2021) and could contribute to individual similarity acting as meQTLs. Using the methylation status of 1,379 CpG sites located within a window of +100 bp from the identified 19,277 SNPs, we observed that 3 of the 16 (18.7%) look-alike pairs clustered together (Figure S6). All three of these pairs were among the 9 ultra-look-alike couples (Figure 2B). Thus, DNA methylation, as a marker of biological age and meQTL, can also provide phenotypic commonality for ultra-look-alikes.

A similar scenario was found for the microbiome. From a qualitative standpoint (alpha diversity), according to the type of bacteria present in the studied oral sample (STAR Methods), only one look-alike pair clustered together (Figure 3C). This couple did not cluster together according to SNP genotyping (Figure 2B). From a quantitative standpoint, according to the amount of each bacteria strand present (STAR Methods), we found clustering of one look-alike pair (6.25%, 1 of 16) (Figure 3D). This couple also paired together by unsupervised SNP clustering (Figure 2B). The study of the nine couples with SNP similarity did not provide further pairing of look-alikes (Figure S6). K-means algorithm illustrated by PCA and t-SNE did not demonstrate clustering (Figure S7). Thus, look-alikes do not mostly share a microbiome. However, oral microbiome relates to obesity (Yang et al., 2019), and fat in the face could relate to similarities. We found that intrapair weight differences were smaller among ultra-look-alike pairs compared with non-ultra-look-alike pairs (two-sided Student's t test test, p = 0.035) (Figure S7). Thus, it is possible that the oral microbiome, through its relation to fat content, contributes to look-alike phenotypes.

Traits of look-alike humans beyond facial features

The likeness between the identified human pairs is not limited to the shared facial traits. All the recruited participants in the study completed a comprehensive biometric and lifestyle questionnaire (Methods S1), and the collected information is summarized in

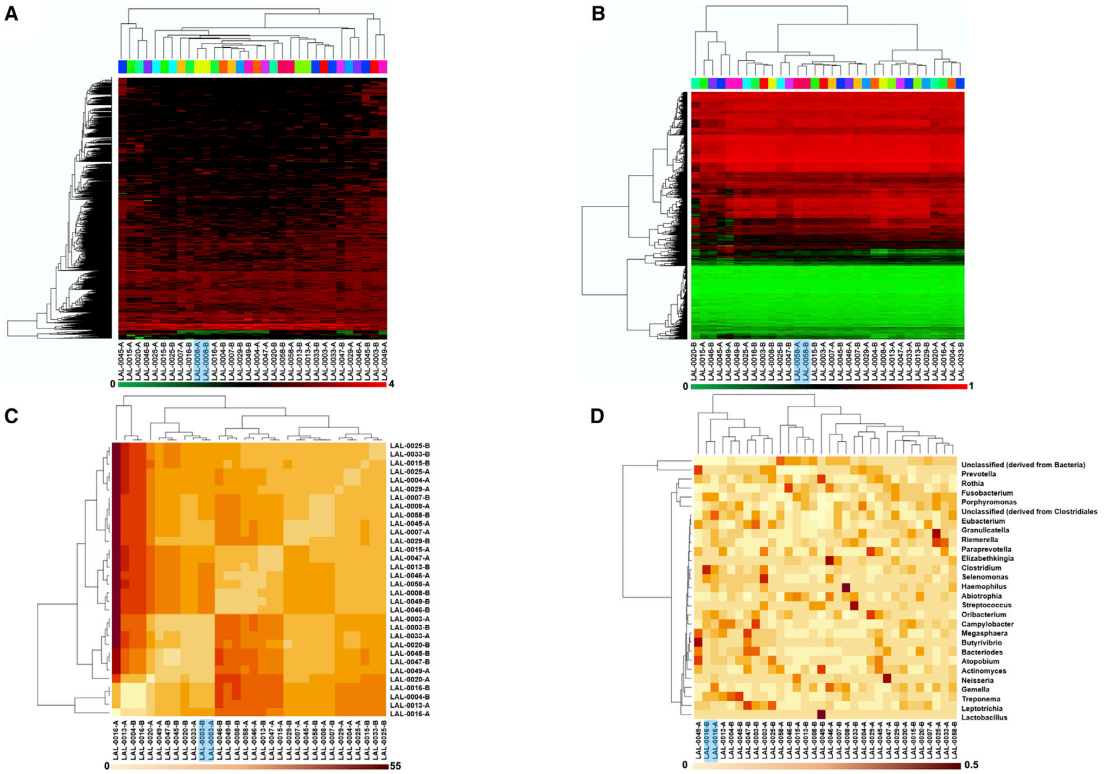


Figure 3. Copy-number variation, DNA methylation, and microbiome analysis of LALs

(A) Heatmap shows the hierarchical clustering of the samples based on the copy number (scale of 0–4) of all copy-number variation (CNV) regions, defined as regions in which at least one individual carried a different copy number. A random selection of one-fifth of such CNV regions is represented in this plot, but the clustering of samples had been obtained considering all CNV regions. The blue rectangle represents a LAL that clusters together.

(B) Heatmap shows unsupervised genome-wide DNA methylation hierarchical clustering with bootstrap of the 16 LALs, using the methylation β-values obtained from MethylationEPIC arrays. A random selection of 5000 CpGs is represented. Colors represent a continuous quantification of methylation beta values at each CpG site, where green highlights unmethylated CpGs (0), black, 50% methylated CpGs (0.5), and red, fully methylated CpGs (1). Clustered look-alikes are shown in a blue rectangle.

(C and D) Microbiome analysis of 16 LALs. Heatmaps show the distances from differences in pairwise bacterial counts of species found in the microbiome of each LAL (variation in alpha diversity scores) of counts from 0–55 (3C) and relative proportions of the taxonomic profiles at the genus level (3D) for each sample calculated on a scale of 0–0.5. Only the most represented genera are shown. Meta-genomic clustering of each look-alike sample was constructed using Euclidean distances and Ward.D2 hierarchical cluster method. Blue rectangle represents LALs whose microbiome is closely related.

Figure 4A. Overall, 68 parameters (Table S7) were included and converted to numerical or logical (0/1) variables (STAR Methods, (custom scripts GitHub: https://github.com/mesteller-bioinfolab/lookalike). The input curated questionnaire is shown in Table S7. We used a cosine similarity method (STAR Methods) to calculate likeness between the studied individuals according to the questionnaire answers. Studying the original 32 look-alike couples, we observed that the 16 look-alike pairs that matched together by all three facial recognition software showed shorter Euclidean distances within pairs ($p = 0.03475$) and higher cosine similarity scores ($p = 0.00321$) than those pairs that did not match by the facial algorithms (Figure 4B). According to their SNPs, the 16 look-alike pairs showed shorter Euclidean distances compared with those pairs that did not match by the three facial algorithms ($p = 0.00006$) (Figure 4B). Examples of independent questionnaire variables (such as height, weight, smoking habit, or level of education) further demonstrate that look-alike pairs are closer than non-look-alike pairs (Figure 4C). Thus, humans with a similar face might also share a more comprehensive physical, and probably behavioral, phenotype that relates to their shared genetic variants. Our study supports the concept of heritability estimation that individuals correlated at the phenotype level share a significant number of genotypic correlations (Visscher et al., 2008). Our results are germane to the ongoing efforts to predict biometric traits from genomic data (Lippert et al., 2017) and the diagnosis of genetic disorders using facial analysis technologies (Gripp et al., 2016; Hadj-Rabia et al., 2017; Hsieh et al., 2019; Gurovich et al., 2019).

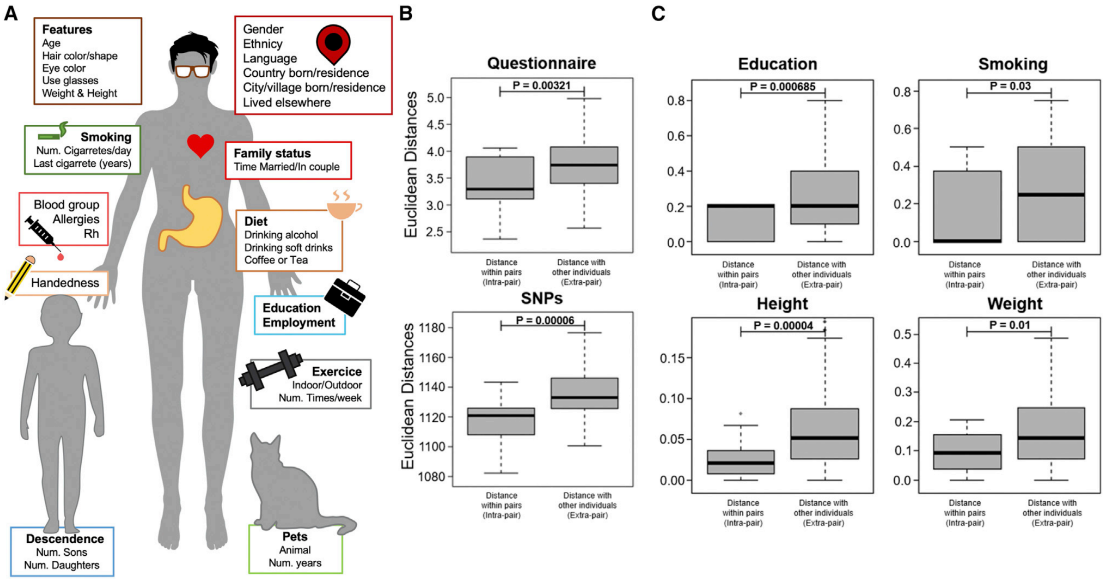


Figure 4. Biometric and lifestyle analysis of LALs using cosine similarity scores

(A) Representation of the biometric and lifestyle parameters considered to calculate cosine similarity scores.

(B) Euclidean distances between the individuals from a pair (intra-pair distance) compared with the distance between individuals from different pairs (extra-pair distance). Distances were calculated for questionnaire (top) and SNP data (below). Statistics by Student's t test.

(C) Distance boxplots for independent questionnaire variables generated by calculating, for all possible pairs of samples, their absolute differences for each variable. We then classified all pairs between pairs of look-alikes and pairs of non-look-alikes. Statistics by Wilcoxon rank sum tests.

DISCUSSION

Our study deciphers molecular components associated with facial construction by applying a multiomics approach in a unique cohort of look-alike humans that are genetically unrelated. Saliva DNA was subjected to genome-wide analyses of common genetic variation, DNA methylation, and microbiome analysis. We also performed a biometric and lifestyle analysis for all look-alike pairs. We found that 16 of the 32 look-alike pairs clustered in all three facial recognition software. Genetic analysis revealed that 9 of these 16 look-alike pairs (Figure 2B) clustered, identifying 19,277 common SNPs. Furthermore, analyses of these shared variants in GWAS and GTEx databases revealed enrichment for phenotypes related to body and face structures and an association with gene-expression changes. Together, this suggests that shared genetic variation in humans that look alike likely contribute to the common phenotype.

Historically, research into face morphology was heavily centered on craniofacial anomalies (Richmond et al., 2018). However, there is a recent growing interest into normal-range face variation, attributable to the necessity for facial recognition software for everyday life (smartphones, CCTV cameras, etc.). Easy access to low-cost, high-resolution pictures and advances in genotyping technology has ignited an age-old question: what makes humans look as they do? Association studies revealed low-frequency genetic variants with relatively small penetrance in facial features, suggesting a far more complex genetic role. Non-genetic factors can affect the expression of genes that form the face. Many epigenetic or imprinting disorders present craniofacial anomalies, such as patients with Prader-Willi or Angelman syndrome (Girardot et al., 2013), and microbial disruption is associated with developmental defects (Robertson et al., 2019). Despite evidence for epigenetic variation in human populations (Heyn et al., 2013) and development (Garg et al., 2018), only one look-alike pair clustered by DNA methylation. This pair also clustered together by SNPs, suggesting that the shared epigenetic profile is likely due to their underlining shared genetics (Lienert et al., 2011), as it was also supported by analyzing CpGs in the vicinity of the SNPs. In addition, ultra-look-alike pairs showed similar epigeneticages. Similarly, only one look-alike pair clustered by microbiome analysis, but ultra-look-alike pairs displayed similar weights, and microbiome composition could relate to obesity (Yang et al., 2019). These findings support a modest role for these biological components to determine facial shape; however, more evidence is required to discard a greater impact.

Finally, 68 biometric and lifestyle attributes from the look-alike pairs were studied. Physical traits such as weight and height as well as behavioral traits such as smoking and education were correlated in look-alike pairs, suggesting that shared genetic variation not only relates to shared physical appearance but may also influence common habits and behavior.

Overall, we provided a unique insight into the molecular characteristics that potentially influence the construction of the

human face. We suggest that these same determinants correlate with both physical and behavioral attributes that constitute human beings. These findings provide a molecular basis for future applications in various fields such as biomedicine, evolution, and forensics. Through collaborative efforts, the ultimate challenge would be to predict the human face structure based on the individual's multiomics landscape.

Limitations of the study

Due to the difficulty to obtain look-alike data and biomaterial, the sample size is small, restricting our ability to perform large-scale statistical analyses. Thus, some partially negative results, such as those derived from the non-genetic data, could relate to an underpowered study. The used headshots were two-dimensional, black and white images, and valuable information regarding three-dimensional constructs, subtle skin tones, and unique facial features are lacking. In addition, the used SNP array does not allow for the analysis of other genetic components such as structural variations and shared rare events. Another limitation is that our samples were mostly from European origin. Thus, the study could not effectively address the impact of the used multiomics in other human populations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Recruitment of look-alikes
- METHOD DETAILS
  - Facial recognition algorithms
  - Facial similarity
  - Sample preparation
  - HumanOmni5-Quad BeadChip
  - Infinium MethylationEPIC BeadChip
  - 16S meta-genomics sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Population-level vs shared SNPs in look-alike pairs
  - Copy number variant (CNV) calling and functional annotation
  - CNV clustering and heatmap
  - Genome-wide SNP arrays from monozygotic twins
  - Cryptic relatedness
  - Ancestry assessment
  - Kinship assessment
  - Functional enrichment of shared SNPs using Gene Ontology
  - Face gene enrichment in the identified SNPs
  - GWAS analysis
  - GWAS functional enrichment of shared SNPs using S-LDSC
  - DNA methylation age estimation
  - Multiomics clustering analyses
  - Questionnaires processing and similarity analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111257.

ACKNOWLEDGMENTS

We thank François Brunelle for providing the look-alike images. We thank CERCA Programme/Generalitat de Catalunya and the Josep Carreras Foundation for institutional support. This work was funded by the governments of Catalonia (2017SGR1080) and Spain (RTI2018-094049-B-I00, SAF2014-55000, and TIN2017-90124-P) and the Cellex Foundation.

AUTHOR CONTRIBUTIONS

M.E. conceived and designed the study; R.S.J., M.R., C.A.G.-P., M.C.d.M., D.P., S.M., V.D., P.C., M.F.-B., I.O., C.L.-F., A.N., C.F.-T., D.A., F.M.S., X.B., A.V., and M.E. analyzed multiomics and questionnaire data; R.J. and M.E. wrote the manuscript with contributions and approval from all authors.

DECLARATION OF INTERESTS

M.E. is a consultant of Ferrer International and Quimatryx. S.M. is an employee of Ferrer International. C.F.-T. is chief technical officer of Herta Security.

Received: July 16, 2021
Revised: June 5, 2022
Accepted: August 1, 2022
Published: August 23, 2022

REFERENCES

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R.B., Pérez, G.M., et al. (2016). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat. Commun. *7*, 11616. https://doi.org/10.1038/ncomms11616.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.

Beck, T., Shorter, T., and Brookes, A.J. (2020). GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic Acids Res. *48*, D933–D940. https://doi.org/10.1093/nar/gkz895.

Biswas, S., Bowyer, K.W., and Flynn, P.J. (2011). A study of face recognition of identical twins by humans. In IEEE International Workshop on Information Forensics and Security, Iguacu Falls, Brazil, pp. 1–6. https://doi.org/10.1109/WIFS.2011.6123126.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

Cao, X., Li, W., Wang, T., Ran, D., Davalos, V., Planas-Serra, L., Pujol, A., Esteller, M., Wang, X., and Yu, H. (2022). Accelerated biological aging in COVID-19 patients. Nat. Commun. *13*, 2135. https://doi.org/10.1038/s41467-022-29801-8.

Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14. https://doi.org/10.1016/j.cell.2017.05.011.

Claes, P., Roosenboom, J., White, J.D., Swigut, T., Sero, D., Li, J., Lee, M.K., Zaidi, A., Mattern, B.C., Liebowitz, C., et al. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat. Genet. *50*, 414–423. https://doi.org/10.1038/s41588-018-0057-4.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. *3*, e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. *10*, 48.

Esteban-Cantos, A., Rodríguez-Centeno, J., Barruz, P., Alejos, B., Saiz-Medrano, G., Nevado, J., Martin, A., Gayá, F., De Miguel, R., Bernardino, J.I., et al. (2021). Epigenetic age acceleration changes 2 years after antiretroviral therapy initiation in adults with HIV: a substudy of the NEAT001/ANRS143 randomised trial. Lancet. HIV *8*, e197–e205. https://doi.org/10.1016/S2352-3018(21)00006-0.

Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G.A., Thirlwell, C., Morris, T.J., Flanagan, A.M., Teschendorff, A.E., Kelly, J.D., et al. (2014). Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. *15*, R30. https://doi.org/10.1186/gb-2014-15-2-r30.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.

Fortin, J.-P., Triche, T.J., Jr., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics *33*, 558–560. https://doi.org/10.1093/bioinformatics/btw691.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA *102*, 10604–10609. https://doi.org/10.1073/pnas.0500398102.

Garrett-Bakelman, F.E., Darshi, M., Green, S.J., Gur, R.C., Lin, L., Macias, B.R., McKenna, M.J., Meydan, C., Mishra, T., Nasrini, J., et al. (2019). The NASA Twins Study: a multidimensional analysis of a year-long human spaceflight. Science *364*, eaau8650. https://doi.org/10.1126/science.aau8650.

Garg, P., Joshi, R.S., Watson, C., and Sharp, A.J. (2018). A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. *14*, e1007707. https://doi.org/10.1371/journal.pgen.1007707.

Girardot, M., Feil, R., and Lères, D. (2013). Epigenetic deregulation of genomic imprinting in humans: causal mechanisms and clinical implications. Epigenomics *5*, 715–728. https://doi.org/10.2217/epi.13.66.

Gripp, K.W., Baker, L., Telegrafi, A., and Monaghan, K.G. (2016). The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. Am. J. Med. Genet. *170*, 1754–1762. https://doi.org/10.1002/ajmg.a.37672.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nat. Med. *25*, 60–64. https://doi.org/10.1038/s41591-018-0279-0.

Hadj-Rabia, S., Schneider, H., Navarro, E., Klein, O., Kirby, N., Huttner, K., Wolf, L., Orin, M., Wohlfart, S., Bodemer, C., et al. (2017). Automatic recognition of the XLHED phenotype from facial images. Am. J. Med. Genet. *173*, 2408–2414. https://doi.org/10.1002/ajmg.a.38343.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367. https://doi.org/10.1016/j.molcel.2012.10.016.

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., and Lumey, L.H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc. Natl. Acad. Sci. USA *105*, 17046–17049. https://doi.org/10.1073/pnas.0806560105.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., et al. (2013). DNA methylation contributes to natural human variation. Genome Res. *23*, 1363–1372. https://doi.org/10.1101/gr.154187.112.

Hoskens, H., Liu, D., Naqvi, S., Lee, M.K., Eller, R.J., Indencleef, K., White, J.D., Li, J., Larmuseau, M.H.D., Hens, G., et al. (2021). 3D facial phenotyping by biometric sibling matching used in contemporary genomic methodologies. PLoS Genet. *17*, e1009528. https://doi.org/10.1371/journal.pgen.

Hsieh, T.C., Mensah, M.A., Pantel, J.T., Aguilar, D., Bar, O., Bayat, A., Becerra-Solano, L., Bentzen, H.B., Biskup, S., Borisov, O., et al. (2019). PEDIA: prioritization of exome data by image analysis. Genet. Med. *21*, 2807–2814. https://doi.org/10.1038/s41436-019-0566-2.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57. https://doi.org/10.1038/nprot.2008.211.

Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H.T., Wong, A.H.C., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. *41*, 240–245. https://doi.org/10.1038/ng.286.

Keegan, K.P., Glass, E.M., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol. Biol. *1399*, 207–233. https://doi.org/10.1007/978-1-4939-3369-3_13.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. *41*, e1. https://doi.org/10.1093/nar/gks808.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413. https://doi.org/10.1038/nature13673.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097. https://doi.org/10.1038/ng.946.

Lippert, C., Sabatini, R., Maher, M.C., Kang, E.Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., et al. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. Proc. Natl. Acad. Sci. USA *114*, 10166–10171. https://doi.org/10.1073/pnas.1711125114.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618. https://doi.org/10.1038/ismej.2011.139.

Mee, L., Honkala, H., Kopra, O., Vesa, J., Finnilä, S., Visapää, I., Sang, T.K., Jackson, G.R., Salonen, R., Kestilä, M., and Peltonen, L. (2005). Hydrolethalus syndrome is caused by a missense mutation in a novel gene HYLS1. Hum. Mol. Genet. *14*, 1475–1488. https://doi.org/10.1093/hmg/ddi157.

Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA methylation microarray for 850, 000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics *8*, 389–399. https://doi.org/10.2217/epi.15.114.

Müllner, D. (2013). Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. *53*, 1–18. https://doi.org/10.18637/jss.v053.i09.

Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Proceedings of the British Machine Vision Conference (BMVA Press), pp. 1–12. https://doi.org/10.5244/C.29.41.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

Quian Quiroga, R. (2017). How do we recognize a face? Cell *169*, 975–977. https://doi.org/10.1016/j.cell.2017.05.012.

R Core Team (2019). In R: A Language and Environment for Statistical Computing Computer Program, version 3.6. 1.

Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol. *11*, e1001555. https://doi.org/10.1371/journal.pbio.1001555.

Richmond, S., Howe, L.J., Lewis, S., Stergiakouli, E., and Zhurov, A. (2018). Facial genetics: a brief overview. Front. Genet. *9*, 462. https://doi.org/10.3389/fgene.2018.00462.

Rideout, W.M., 3rd, Eggan, K., and Jaenisch, R. (2001). Nuclear cloning and epigenetic reprogramming of the genome. Science *293*, 1093–1098. https://doi.org/10.1126/science.1063206.

Robertson, R.C., Manges, A.R., Finlay, B.B., and Prendergast, A.J. (2019). The human microbiome and child growth - first 1000 Days and beyond. Trends Microbiol. *27*, 131–147. https://doi.org/10.1016/j.tim.2018.09.008.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542. https://doi.org/10.1093/bioinformatics/btl117.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674. https://doi.org/10.1126/science.1119983.

Vedaldi, A., and Lenc, K. (2015). MatConvNet in Proceedings of the 23rd ACM International Conference on Multimedia (MM '15) (ACM Press), pp. 689–692. https://doi.org/10.1145/2733373.2807412.

Villicaña, S., and Bell, J.T. (2021). Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. *22*, 127. https://doi.org/10.1186/s13059-021-02347-6.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266. https://doi.org/10.1038/nrg2322.

Weir, B., Cockerham, C., and Feldman, M. (2014). Complete characterization of disequilibrium at two loci. In Mathematical Evolutionary Theory (Princeton: Princeton University Press), pp. 86–110. https://doi.org/10.1515/9781400859832-007.

White, J.D., Indencleef, K., Naqvi, S., Eller, R.J., Hoskens, H., Roosenboom, J., Lee, M.K., Li, J., Mohammed, J., Richmond, S., et al. (2021). Insights into the genetic architecture of the human face. Nat. Genet. *53*, 45–53. https://doi.org/10.1038/s41588-020-00741-7.

Wolff, G.L., Kodell, R.L., Moore, S.R., and Cooney, C.A. (1998). Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. FASEB J. *12*, 949–957. https://doi.org/10.1096/fasebj.12.11.949.

Xing, C., Huang, J., Hsu, Y.H., DeStefano, A.L., Heard-Costa, N.L., Wolf, P.A., Seshadri, S., Kiel, D.P., Cupples, L.A., and Dupuis, J. (2016). Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases. Eur. J. Hum. Genet. *24*, 1029–1034. https://doi.org/10.1038/ejhg.2015.244.

Xiong, Z., Dankova, G., Howe, L.J., Lee, M.K., Hysi, P.G., de Jong, M.A., Zhu, G., Adhikari, K., Li, D., Li, Y., et al. (2019). Novel genetic loci affecting facial shape variation in humans. Elife *8*, e49898. https://doi.org/10.7554/eLife.49898.

Yang, Y., Cai, Q., Zheng, W., Steinwandel, M., Blot, W.J., Shu, X.O., and Long, J. (2019). Oral microbiome and obesity in a large study of low-income and African-American populations. J. Oral Microbiol. *11*, 1650597. https://doi.org/10.1080/20002297.2019.1650597.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

Zhang, C., and Zhang, Z. (2010). A survey of recent advances in face detection. Microsoft Research.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Oragene DNA tubes	DNA Genotex	OG-500
Pico Green fluorescence kit	Life technologies/thermos	P7589
EZ DNA Methylation Kit	Zymo Research	D5003
Deposited data		
HumanOmni5-Quad BeadChip	This paper	GEO: GSE142304
Infinium MethylationEPIC BeadChip	This paper	GEO: GSE142304
16S metagenomics sequencing	This paper	BioProject: PRJNA596439
Custom scripts	This paper	https://github.com/mesteller-bioinfolab/lookalike
Look-alike photographs	www.francoisbrunelle.com/webn/e-project.html	https://github.com/mesteller-bioinfolab/lookalike/blob/master/FB_LAL_images.zip
Experimental models: Organisms/strains		
Humans (Homo sapiens)	Look-alike individuals upon consent.	N/A
Software and algorithms		
R	R Core team., 2019	www.r-project.org/
MatConvNet	VLFeat	http://www.vlfeat.org/matconvnet
Microsoft Oxford Project face API	Microsoft Azure	https://azure.microsoft.com/en-us/services/cognitive-services/face/
Herta CNN algorithm	Herta Security	www.hertasecurity.com
GenomeStudio (v2.0.4)	Illumina	https://support.illumina.com/downloads/genomestudio-2-0.html
pvclust	Suzuki and Shimodaira, 2006	http://stat.sys.i.kyoto-u.ac.jp/prog/pvclust/
hclust	Müllner, 2013	https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html
Kinship-based INference for GWAS (KING v2.2.3)	Manichaikul et al., 2010	http://people.virginia.edu/~wc9c/KING/
Minfi (v1.32.0)	Aryee et al., 2014 Fortin et al., 2017	https://bioconductor.org/packages/release/bioc/html/minfi.html
clusterProfiler	Yu et al., 2012	https://guangchuangyu.github.io/2016/01/go-analysis-using-clusterprofiler/
Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8)	Huang et al., 2009	https://david.ncifcrf.gov/
GOrilla	Eden et al., 2007, 2009	http://cbl-gorilla.cs.technion.ac.il/
GTEx portal (v7)	https://gtexportal.org/	N/A
GWAS catalog	Buniello et al., 2019	https://www.ebi.ac.uk/gwas/
GWAS central	Beck et al., 2020	https://www.gwascentral.org/
MG-RAST	Keegan et al., 2016	https://www.mg-rast.org/
Greengenes rRNA database	McDonald et al., 2012	https://greengenes.secondgenome.com/
Other		
François Brunelle website	www.francoisbrunelle.com/webn/e-project.html	N/A
University of Notre Dame twins database 2009/2010	https://cvrl.nd.edu/projects/data/	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for reagents and resource may be directed to and will be fulfilled by the lead contact, Dr. Manel Esteller (mesteller@carrerasresearch.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- SNP and DNA methylation data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Microbiome data have been deposited on the BioProject repository and are publicly available as of date of publication. Photographs of the look-alike pairs that were matched together for all three different independent facial recognition softwares have been deposited at GitHub and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Recruitment of look-alikes

32 Look-alike pairs (n = 64 individuals) that were initially recruited and photographed by François Brunelle (http://www.francoisbrunelle.com/webn/e-project.html) were enrolled to this study. All 64 individuals [42 females (65.6%) and 22 males (34.4%) with a median age of 40 years (range from 21 to 78 years), Table S7] were required to complete an extensive biometric and life-style questionnaire (Methods S1: Data collection questionnaire, related to STAR Methods) as well as provide legally signed consent forms approved by our bioethics committee for usage of both their facial images and DNA samples for this study. The study protocol was approved by the Clinical Research Ethics Committee of the Bellvitge University Hospital with the reference number PR348/16. To compliment this study, we were also provided with access to 100 monozygotic twin photos from the University of Notre Dame twins database 2009/2010 (https://cvrl.nd.edu/projects/data/). License agreements for data access were reviewed and signed by legal representatives of all entities involved in this study. 50 monozygotic twin pairs (n = 100) photographs were subsequently downloaded and analysed with the facial recognition algorithms detailed below.

METHOD DETAILS

Facial recognition algorithms

Three facial recognition algorithms were used to objectively analyze look-alike pairs: MatConvNet CNN algorithm, provided by the University of Pompeu i Fabra, Barcelona (Vedaldi and Lenc 2015); Microsoft Oxford Project face API by Microsoft; and the custom deep convolutional neural network Custom-Net (www.hertasecurity.com). The quantitative assessment of pairwise similarity between face photographs was calculated as follows. For the MatConvNet algorithm, the face biometric template from each photo was extracted from each processed face by means of a deep convolutional neural network (CNN) built into MatConvNet software. The resulting templates are represented as integer sparse descriptors of 8,192 values, which effectively encode the identity features of a face image (Vedaldi and Lenc 2015). Final pairwise similarity scores were set on a scale of 0–1 where 1 represents identical faces.

The custom deep convolutional neural network Custom-Net was developed by a leader in facial recognition platforms (www.hertasecurity.com). Firstly, a generic face detector optimized for unconstrained video surveillance scenarios was used to obtain the locations of all faces in each image (Zhang and Zhang, 2010). The threshold was adjusted to find all targeted faces in each photo, and a subsequent manual exploration was conducted to ensure that no false positives were included. Each face was cropped with a 25% extra margin from the original bounding box, converted to grayscale and resized to 250 × 250 pixels. Next, a face biometric template was extracted from each processed face by means of a deep convolutional neural network of 32 layers. The resulting templates were represented as integer sparse descriptors of 4,096 values, which effectively encode the identity features of a face image. Finally, the similarity score between a pair of images was computed as a negative mean square deviation between their template values. The final scores were mapped to a range 0–1, where 1 indicated identical faces, according to landmarks taken from the histogram of imposter pairs extracted from the well-known database (http://vis-www.cs.umass.edu/lfw/).

In the case of the custom deep convolutional neural network, the models have tens of millions of learned parameters and have been trained with more than 10 million facial images from over a hundred thousand subjects from different human populations, in a variety of unconstrained situations: differences of pose, expression, age and accessories within a subject. Moreover, the training process of a face recognition algorithm typically involves "data augmentation" operations, in which input images are randomly modified, e.g. by

artificially synthesizing glasses, adding facial occlusions, mirroring faces, etc. in order to add intraclass variability to the images and confer robustness to the resulting model. As a consequence, modern face verification algorithms have recently achieved near-perfect accuracy, as high as 99.97% on NIST's Facial Recognition Vendor Test (https://pages.nist.gov/frvt/html/frvt11.html#overview), for passport photo or mugshot scenarios, to the point that banks worldwide have widely adopted such systems for user verification. Particularly, these algorithms have become extremely reliable on controllable, almost ideal scenarios such as those captured by the photographer: 1:1 verification between large resolution images with good illumination, non-lateral poses (less than 60°) and without heavy occlusions; despite circumstancial similaritiy in interclass appearance like that given by glasses, facial expression or hairstyle. Thus, the impacts of these attributes, such as pose, hairstyle etc can be considered minimum, because the incorporated models have been exposed to these variations, in addition to additional features aspects such as colour styles, image degradations etc. The VGG dataset (https://www.robots.ox.ac.uk/~vgg/data/vgg_face/) shows examples of facial data used to train Matconvnet (Parkhi et al., 2015) and CustomNet (http://vis-www.cs.umass.edu/lfw/).

The Microsoft Oxford Project face API by Microsoft operates on a number of attributes that affect facial features such as age, gender, pose, smile, and facial hair along with 27 other landmarks for each face. These landmarks are left pupil, right pupil, nose tip, left mouth, right mouth, outer left eyebrow, inner left eyebrow, outer left eye, top left eye, bottom left eye, inner left eye, inner right eye, outer right eyebrow, inner right eye, top right eye, bottom right eye, outer right eye, left nose root, right nose root, top left nose alar, top right nose alar, left outer tip of nose alar, right outer tip of nose alar, top upper lip, bottom upper lip, top under lip and bottom under lip (https://azure.microsoft.com/en-us/services/cognitive-services/face/). The final similarity scores were also set on a scale of 0–1.

Facial similarity

Pair-wise facial similarity matrices were provided as an output for all three facial recognition software. Similarity scores were assigned as numerical values ranging between 0 – 1 where 1 represents identical images and 0, two opposed images. To obtain objective look-alike pairs, we performed unsupervised hierarchical clustering with bootstrap using the pvclust (Suzuki and Shimodaira 2006) in R statistical environment (v3.6.1) (https://www.R-project.org/).

Sample preparation

Genomic DNA from look-alike pairs in this study were isolated from saliva and self-collected into Oragene 500 DNA tubes and extracted according to the manufacturers instructions (DNA genotek). >10% of the extracted DNA corresponded to microbial DNA. DNA was quantified using Pico Green fluorescence kit/Qubit® 2.0 Fluorometer (life technologies). Bisulfite modification of genomic DNA was carried out with the EZ DNA Methylation Kit (Zymo Research) following the manufacturer's protocol.

HumanOmni5-Quad BeadChip

Comprehensive cross-examination of genome-wide single nucleotide variation of 4.3 million SNVs across all Look-alike pairs was performed using HumanOmni5-Quad BeadChip (Illumina). 400 ng of genomic DNA was applied to HumanOmni5-Quad BeadChip and scanned using HiScan SQ system (Illumina). The signal raw intensities for each array were assessed and analyzed with GenomeStudio Software (v2.0.4) (Illumina) using default normalization to generate X and Y intensity values for A and B alleles (generic labels for two alternative SNP alleles), respectively. Genotype calling were performed by using GenomeStudio GenCall method and only genotypes with high GenCall scores (GC) were selected (according to Illumina standards). The positions corresponding to Illumina internal controls were also removed from the analysis. In order to remove the positions shared between look-alike pairs by chance, a bootstrap look-alike control analysis was performed. Briefly, we generated 100 datasets of 16 random pairs extracted from the initial 32 pairs (64 individuals) used in the study and the complete SNP set from the Omni5 array (4M SNPs). The only requirement was that none of the generated random pairs in the 100 datasets included a candidate look-alike pair from the initial 32 couples. We applied to each of these new 100 "non-look-alike" datasets the same SNP selection protocol used in the look-alike datasets, i.e. removing monomorphic and non-autosomal positions and selecting the shared inter-look-alike genotypes for each of the 16 pairs. This iterative process produced 100 independent SNP datasets that represented shared genotypes between non-look-alike pairs. Each of the SNP lists obtained contained an average of 5000 SNPs. The plot of the cumulative distribution of these shared SNPs after 100 iterations shows that the number of observed SNPs tends to plateau, indicating that we are reaching a maximum number of SNPs shared by the non-look-alike pairs is being reached. Next, we pooled all 100 SNP datasets into one table removing all redundant variants. This table of unique SNPs was considered as the SNP positions shared between pairs independent of their look-alike status (by chance) and were subsequently removed from our analysis of the look-alike pairs. Then the XY and monoallelic positions for the 16 original pairs were removed. Finally, the SNPs with identical genotypes in each of the 16 pairs and located in genes were selected for further analysis. CNV calling was performed by using PennCNV plugin in GenomeStudio with default parameters.

Infinium MethylationEPIC BeadChip

Genome-wide DNA methylation interrogation of >850,000 CpG sites was performed using the Infinium MethylationEPIC BeadChip (Illumina) according to manufacturer's recommended protocol, as previously described (Moran et al., 2016). Briefly, 600 ng of DNA was used to hybridize to the EPIC BeadChip and scanned using HiScan SQ system (Illumina). Raw signal intensity data were initially QC'd and pre-processed from resulting idat files in R statistical environment (v3.6.1) using minfi Bioconductor package (v1.32.0).

A number of quality control steps were applied to minimize errors and remove erratic probe signals. Firstly, interrogation of sex chromosomes was performed to identify potential labeling errors. Next, the removal of problematic probes was carried out, such as failed probes (detection p value > 0.01), cross-reacting probes and probes that overlapped single nucleotide variants within +/− 1bp of CpG sites followed by background correction and dye-based normalization using ssNoob algorithm (single-sample normal-exponential out-of-band). Lastly, we removed all sex chromosomes. Final DNA methylation scores for each CpG were represented as a β-values ranging between standard 0 and 1 where 1 represents fully methylated CpGs and 0, fully unmethylated. All downstream analyses were performed under R statistical environment (v3.6.1).

16S meta-genomics sequencing

We identified and compared bacterial populations from diverse microbiomes from all look-alike pairs using 16S metagenomics sequencing (Illumina) (Klindworth et al., 2013). Salival DNA was extracted and bacterial libraries prepared following the Illumina 16S Library preparation protocol. The variable V3 and V4 regions of 16S rRNA was amplified in order to obtain a single amplicon of approximately 460 bp that underwent paired-end sequencing using MiSeqDx (Illumina). Resulting fastq files were analysed using MG-RAST. The counts corresponding to taxonomic abundance profiles for each sample were retrieved by using MG-RAST tools. Particularly, we retrieved the bacterial counts from sequences aligned to Genus taxonomic categories in the Greengenes rRNA database with the following cutoffs: an alignment length of 15 bp, a percent identity of 60% and an e-value equal or lower to 1×10^{-5}. The relative proportions for each genus and sample were calculated and only the most represented genus were used.

QUANTIFICATION AND STATISTICAL ANALYSIS

Population-level vs shared SNPs in look-alike pairs

In order to define the number of SNPs shared between non look-alike pairs by chance we generated 55 random combinations of the 9 ultra look-alike pairs avoiding in each dataset the presence of a look-alike pair. We selected the SNP positions with the same genotype for each of the 9 non look-alike pairs in any of the 55 control datasets, obtaining the percent of randomly shared variants in a data set of 9 non look-alikes. Finally, we calculated the statistical significance of the comparison between SNPs shared in look-alike and non look-alike pairs by a Pearson's chi-squared test (p value $<2.2\ 10^{-16}$). However, since different pairs of look-alikes were from multiple different ethnicities, but individuals in the same look-alike pair shared the same ethnicity, we also performed the enrichment analysis to determine if the number of shared SNPs was more than expected by chance accounting to ethnicity. Thus, we tested pairs of European ancestry individuals with other Europeans and repeated the same for each of the different ethnicities. To this end, we downloaded the most recent set of Omni genotypes from 1000 Genomes available in the phase 3 release directory (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/). The downloaded 1000 Genomes phase 3 vcf file was transformed to Genomic Data Structure (GDS) format using the function seqVCF2GDS from SeqArray R package (version 1.36.0). Look-alike PLINK PED files were also transformed to GDS format using the fucntion snpgdsPED2GDS from SNPRelate R package (version 1.30.1). The 1000 Genomes genotyping data was merged with the "ultra" look-alikes genotyping data and the remaining dataset held 67,312 common SNPs. Finally, for each ethnicity we generated 55 random combinations of non look-alike pairs to test if the number of shared SNPs in our "ultra" look-alike population was more than expected by chance. Considering the European ancestry of the majority of "ultra" look-alike (6 out of 9) and non-"ultra" look-alike (7 out of 7) pairs in our study, we used the 7 non-"ultra" look-alike pairs with European ancestry to create 55 random combinations of 6 random non look-alike pairs to compute the number of shared SNPs with the same genotype as a proxy for the European population. For East Asia, Central-South Asia and Hispanic populations, we generated 55 random combinations of 1 random non look-alike 1000 Genomes pair to compute the number of shared SNPs in each of the aforementioned populations. Finally, the number of SNPs shared by "ultra" look-alike pairs in each population was tested for statistical significance enrichment against the background number of shared SNPs in each non look-alike population by means of the Pearson's chi-squared test.

Copy number variant (CNV) calling and functional annotation

The impact of CNVs on genes was calculated in two different ways. First, we looked at whole-gene CNVs, and then partially-overlapping CNVs. Copy number of all genes in the genome was calculated by first establishing CNV breakpoints. Breakpoints were assigned to the outermost SNP positions of regions with the same copy number. The breakpoints were calculated separately for each sample. Using these coordinates, the copy number of whole protein-coding and RNA genes was calculated for all individuals. Gene coordinates were obtained from Ensembl v75 (build GRCh37). We took the genes that had a shared copy number in all pairs of look-alikes (both individuals within the pair had the same number of copies), and we selected those genes for which at least one pair of look-alikes had a different number of copies than the rest of the pairs. For example, to look for partially-overlapping CNVs, we selected all positions in the genome in which the copy number matched within all pairs, but for which at least 2 pairs of lookalikes had a different copy number to the rest of the pairs. We then looked for overlaps with partial overlaps with coding or non-coding genes. As an example, region chr11:125778219-125780253, which overlaps with a lncRNA that has a regulatory relationship with the HYLS1 gene, there are three pairs of look-alikes that carry three copies of this lncRNA, while the remaining pairs have two copies of it. All custom R scripts for CNV analysis are deposited in GitHub repository: https://github.com/mesteller-bioinfolab/lookalike.

CNV clustering and heatmap

Clustering of CNVs was done after filtering out all positions with the same copy number in all samples and merging all contiguous positions with the same copy number. Positions from the X and Y chromosomes that showed the same copy number in all males and the same copy number in all females were also filtered out. The clustering of the samples was calculated using pvclust (Suzuki and Shimodaira 2006). Variants represented in the heatmap are a random selection of one fifth of the total number of variants.

Genome-wide SNP arrays from monozygotic twins

We obtained single nucleotide polymorphism (SNP) data for 38 monozygotic twins from two publicly available studies. Both were downloaded from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession No. GSE33598 and GSE9608. The signal raw intensities for each array were assessed and analyzed with GenomeStudio Software (v2.0.4) (Illumina) using default normalization to generate X and Y intensity values for A and B alleles (generic labels for two alternative SNP alleles), respectively. All downstream analyses were performed in the R statistical environment (v3.6.1) (https://www.R-project.org/).

Cryptic relatedness

Robust relatedness inference and genetic correlation estimates between monozygotic twins, look-alike pairs and random non look-alikes were calculated using the software KING (Kinship-based INference for GWAS) (version 2.2.3). Student's t-test was applied to calculate statistical significance between populations.

Ancestry assessment

Genotyping was performed using GenomeStudio v2.0.5; PACKPED Plink files were created using the software PLINK Input Report Plug-in v2.1.4 (https://emea.support.illumina.com/downloads/genomestudio-2-0-plugins.html). To analyze the look-alike pairs in the context of world-wide genetic diversity, their genomic data was merged using with 1,980 West-Eurasian, Asian and Native American individuals genotyped in the Affimetrix HO array (Lazaridis et al., 2014); the remaining dataset held 175,469 common SNPs. Principal Component Analysis (PCA) was generated with the HO individuals. Look-Alike individuals were then projected onto the first two components (PC1 and PC2) using options 'lsqproject: YES' and 'shrinkmode: YES' of smartpca built-in module of EIGENSOFT (v. 7.2.1) (Patterson et al., 2006; Weir et al., 2014) (https://www.hsph.harvard.edu/alkes-price/software/).

Kinship assessment

Kinship coefficients between look-alike pairs was first estimated with PLINK. PLINK uses a method-of-moments approach where the total proportion of shared SNPs IBD is calculated based on the estimated allele frequency of all SNPs in a dataset assumed to be homogeneous (Purcell et al., 2007). PLINK-indep-pairwise option was used with parameters 50 5 1.5. to generate a pruned subset of genotypes in low linkage disequilibrium of 282,122 SNPs in comparisons with 1000G dataset and 103,256 in comparisons with HO dataset; pairwise relatedness between individuals of each pair was calculated with the –genome–min-0.05 command to detect pairs with levels of IBD sharing compatible with up to a 3rd degree relationship (Manichaikul et al., 2010). Potential relatedness between pairs was subsequently explored by estimating long (>10 cM) IBD blocks that might be indicative of co-ancestry among individuals occurring in the last few hundreds or years (Ralph and Coop, 2013).

Functional enrichment of shared SNPs using Gene Ontology

Enrichment analysis was done with the enrichGO function from the clusterProfiler R package (Yu et al., 2012), using the org.Hs.eg.db genome annotation. The tested 3,730 genes annotated to the 19,277 SNPs with a matching genotype in all pairs of look-alikes. The background list of genes were all genes annotated to SNPs detected in HumanOmni5-Quad BeadChip analysis. Parameters minGSSize and maxGSSize from the enrichGO function were set to 1 and 22000, respectively, in order to capture all gene ontologies. Additional enrichment analyses were done using DAVID v6.8 and GOrilla.

Enrichment of eQTLs in the look-alike SNPs set was calculated using data from the GTEx portal, release v7 (GTEx_Analysis_v7.metasoft.txt.gz). eQTLs with a fixed effect model p-values < 0.05 were selected for the analysis. A Fisher's test was performed to calculate if the overlap between look-alike SNPs and eQTLs was bigger than expected by chance. The same enrichment analysis was done with each tissue independently, considering the eQTLs with a tissue-specific p-value <0.05. Gene ontology analysis was performed using GOrilla.

Face gene enrichment in the identified SNPs

In order to statistically evaluate the face genes enrichment in our selected 19,277 SNPs corresponding to 3,730 genes shared by all "ultra" look-alike pairs, we gather all the genes related with face traits (face genes) from recent comprehensive genomic screenings related to facial shape (Claes et al., 2018; Xiong et al., 2019; White et al., 2021), the Facebase dataset (https://www.facebase.org/) and GWAS central (study HGVST1841, http://www.gwascentral.org) and applied two different approaches. In the first approach, we applied a hypergeometric test, as it is implemented in the R "phyper" function, from the package "stats". In the second, we also performed a Monte Carlo simulation using 10,000 iterations. In each iteration, we selected a random set of 3,730 genes (the same number of genes in our 19,277 SNPs) from the total genes represented in the array (23,774 genes) and we counted the number of face genes found in this random selection. All the analyses were performed in R statistical programming language v.4.0.3.

GWAS analysis

The overlap between matching sets of SNPs called from look-alike pairs and GWAS SNPs was performed using data from two GWAS databases: GWAS Catalog and GWAS Central. In GWAS Catalog v1.0.2, all GWAS SNPs were retrieved and lifted over from GRCh38 to GRCh37 using the R package liftOver. To calculate trait enrichment, we performed Fisher's exact tests, computing matching genotypes from look-alike pairs against all SNPs detected in the HumanOmni5-Quad BeadChip. For GWAS Central analysis, studies related to facial morphology (HGVST1044, HGVST1625, HGVST1841, HGVST1892, HGVST1933, HGVST2265, HGVST2325, HGVST2359, HGVST2363 and HGVST2597) were selected. Fisher's exact tests were performed to calculate significant overlaps in the different studies and correction for multiple testing was done with Benjamini and Hochberg's adjustment method ($\alpha = 0.05$). All custom R scripts for SNP functional analysis are deposited in GitHub repository: https://github.com/mesteller-bioinfolab/lookalike.

GWAS functional enrichment of shared SNPs using S-LDSC

In order to determine the enrichment of GWAS signals for specific annotations we used the stratified LD score regression (S-LDSC) tool (github.com/bulik/ldsc). S-LDSC is a method to estimate heritability enrichment for selected functional annotations. To this end, we followed the partitioned heritability analysis tutorial (github.com/bulik/ldsc/wiki/Patitioned-Heritability) using the last and recommended version of the baseline-LD model (version 2.2) with 97 annotations. To asses the heritability enrichment of our 19,277 SNPs, we included a "look-alike" custom functional annotation, defined by the set of 19277 SNPs, on top of the baseline-LD model v2.2. Since S-LDSC is typically applied to large annotations, we included a 500-bp window around the set of 19,277 SNPs to define our custom "look-alike" functional annotation category, following the annotation format of the baseline-LD model v2.2. Considering the European ancestry of the majority of samples in our study, we performed the S-LDSC analysis using European LD scores and allele frequencies from the 1000 Genomes Phase 3 project. Full summary statistics available for "facial morphology" trait in European ancestry individuals were downloaded from GWAS Catalog, corresponding to two studies (Xiong et al., 2019; Hoskens et al., 2021). Finally, partition heritability analysis was performed with default parameters and facial traits with ES >1 and enrichment p value < 0.05 were considered.

DNA methylation age estimation

Epigenetic age estimation was computed using the Hannum method using the function methyAge from the ENmix R package (version 1.32.0).

Multiomics clustering analyses

To genetically, epigenetically and metagenomically categorize inherent similarities between all look-alike pairs, shared SNV, CNV, DNA methylation and microbiota profiles, robust correlations and unsupervised hierarchical clustering with bootstrapping were performed with R function packages pvclust (Suzuki and Shimodaira 2006). Euclidean distance scores and ward.d2 minimum variance method were applied to attain hierarchical clustering represented as heatmaps using R statistical environment (v3.6.1). K-means clustering was also performed and represented using the first two dimensions of a Principal Component Analysis (PCA). To perform k-means clustering, 16 "centers" (clusters) were indicated. The SNP set was also visualized using t-SNE representation, selecting 2 dimensions and adjusting "perplexity" parameter to 6 and "max_iter" to 5,000. All the analysis were performed in R statistical programming language v.4.0.3 using the packages "SNPRelate", "gdsfmt","stats", "Rtsne", "ggfortify" and "ggplot2".

Questionnaires processing and similarity analysis

Data obtained through questionnaires was transformed into a table, which was processed and transformed into numerical format with a custom script (deposited in GitHub; https://github.com/mesteller-bioinfolab/lookalike). In this script, all logical variables were transformed to 0 (False/No) and 1 (True/Yes). When the variables could be ordered (e.g. Never - Sometimes - Often), they were assigned numbers (0–1 - 2 in the example) that were afterwards normalized to 1. For non-sortable variables, the categories were split into logical columns (e.g. Employment category was split into three logical variables - Executive, Salaried and Own business). Finally, empty boxes were filled with the mode for each variable. Cosine similarity was calculated using the numerical matrix between all individuals. The look-alike intra and extra-pair distance analysis were defined and calculated as follows. Intra-pairs were defined as look-alike pairs that clustered in all three facial recognition software (n = 16). The extra-pairs were defined as all other combination pairs of non look-alikes in the initial 16 pairs. For 32 individuals, pairs of same individuals and their look-alike pair counterpart were removed, leaving 30 possible combinations per 16 pair (n = 480). The euclidean distances between each individual and all other samples were calculated using the dist function from the R package pvclust (Suzuki and Shimodaira 2006). Distances were calculated on SNP, CNV, methylome, quantitative and qualitative microbiome and questionnaire data. Intra-pair distances were compared to extra-pair distances using Student's T test. Distance boxplots for independent variables were generated by calculating, for all possible pairs of samples, their absolute differences for each variable. We then classified all pairs between pairs of look-alikes and pairs of non-look-alikes. Finally, we calculated if the differences were significant with Wilcoxon rank sum tests.

Co-authored publications

The following is a list of publications to which I contributed as a co-author during my doctoral program.

1. Martinez-Verbo, L. *et al.* PVR (CD155) epigenetic status mediates immunotherapy response in multiple myeloma. *Leukemia* (2024) doi:10.1038/s41375-024-02419-z
2. Malla, S. *et al.* The scaffolding function of LSD1 controls DNA methylation in mouse ESCs. *Nat Commun* **15**, 7758 (2024).
3. Noguera-Castells, A. *et al.* DNA methylation profiling of myelodysplastic syndromes and clinical response to azacitidine: A multicentre retrospective study. *Br J Haematol* **204**, 1838–1843 (2024).
4. Noguera-Castells, A. *et al.* Epigenetic Fingerprint of the SARS-CoV-2 Infection in the Lung of Lethal COVID-19. *Chest* **165**, 820–824 (2024).
5. Gallardo-Gómez, M. *et al.* Serum DNA methylome of the colorectal cancer serrated pathway enables non-invasive detection. *Mol Oncol* (2023) doi:10.1002/1878-0261.13573.
6. Noguera-Castells, A., García-Prieto, C. A., Álvarez-Errico, D. & Esteller, M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* **18**, 2185742 (2023).
7. Pham, V. N. *et al.* Formaldehyde regulates S-adenosylmethionine biosynthesis and one-carbon metabolism. *Science* **382**, eabp9201 (2023).
8. Salz, L. *et al.* Culture expansion of CAR T cells results in aberrant DNA methylation that is associated with adverse clinical outcome. *Leukemia* **37**, 1868–1878 (2023).
9. Ortiz-Barahona, V. *et al.* Epigenetic inactivation of the 5-methylcytosine RNA methyltransferase NSUN7 is associated with clinical outcome and therapeutic vulnerability in liver cancer. *Mol Cancer* **22**, 83 (2023).
10. Bueno-Costa, A. *et al.* Remodeling of the m6A RNA landscape in the conversion of acute lymphoblastic leukemia cells to macrophages. *Leukemia* **36**, 2121–2124 (2022).
11. Tottone, L. *et al.* A Tumor Suppressor Enhancer of PTEN in T-cell Development and Leukemia. *Blood Cancer Discovery* **2**, 92–109 (2021).