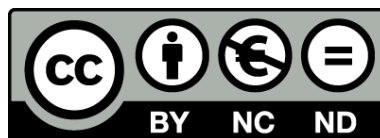




UNIVERSITAT^{DE}
BARCELONA

Somatic processed pseudogenes and micropeptides in cancer: insights from large-scale genomic studies

Ana Dueso Barroso



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

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

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

UNIVERSITAT DE BARCELONA
FACULTAT DE MEDICINA
DEPARTAMENT D'ANATOMIA PATOLÒGICA, FARMACOLOGIA I MICROBIOLOGIA
PROGRAMA DE DOCTORAT EN BIOMEDICINA. LINEA DE RECERCA
BIOINFORMÀTICA

Somatic processed pseudogenes and micropeptides in cancer: insights from large-scale genomic studies

Memòria presentada per Ana Dueso-Barroso per optar al grau de doctora per la
Universitat de Barcelona.

Programa de doctorat en Biomedicina, departament d'anatomia patològica,
farmacologia i microbiologia.

DIRECTOR DE TESI



David Torrents Arenales

DOCTORANDA



Ana Dueso-Barroso

TUTOR



Elías Campo Güerri



UNIVERSITAT DE
BARCELONA



Barcelona
Supercomputing
Center

Centro Nacional de Supercomputación

Acknowledgments

Aquest és l'apartat de la tesi que més ganes he tingut sempre d'escriure. I també el que he volgut escriure últim. Han sigut gairebé 8 anys, així que és probable que sigui llarg. Tant de bo no deixar-me a ningú.

Tancar una etapa com aquesta no ha sigut fàcil emocionalment. Tampoc redactar tota la feina feta quan no sents que li donin prou valor, et carregues de frustració i bloqueig, i es fa molt més llarg del que pensaves en començar. Però sempre he tingut a les millors persones que podria haver trobat acompanyant-me científicament, i personalment. Res seria el que és, si no fos per totes elles.

Per començar, agrair al Dr. Elias Campo Güerri per ser el tutor d'aquesta tesi i a la Dra. Núria López-Bigas i el Dr. Miguel Vázquez per haver fet el seguiment del meu treball durant cada curs i comentar amb rigorositat els resultats. També als membres del tribunal, la Dra. Maria del Mar Albà Soler, el Dr. Abel González-Pérez i al Dr. Jose Antonio Seoane Fernández, per acceptar la proposta i donar-me una resposta tan ràpida. Va ser un alleujament tenir la confirmació de tot el tribunal a la primera. També a la Dra. Lara Nonell, i al Dr. Eduard Porta per ser part del tribunal com a suplents. A totes les persones amb les quals he col·laborat durant aquests anys, i en especial a la Marion Martínez, per ser també un punt de referència en començar etapa al VHIO.

A totes les dones científiques invisibilitzades que han obert el camí perquè avui puguem ser-hi aquí.

M'agradaria continuar agraint al David Torrents, el director d'aquesta tesi, haver-me donat l'oportunitat de formar part del seu grup i animar-me a provar que tal això de la bioinformàtica, sense por a llançar-me a la piscina. Per cada lliçó i per acompanyar-me, a la seva manera, mentre em formava com a científica.

Cada dia de feina durant aquests anys no hauria sigut el mateix sense totes les persones del grup de Computational Genomics amb les que m'he anat creuant. A la Lidia, Juan, Flo, Héctor, Álex, Paula, Silvia, Txema i Álvaro. A la Mercè, Elias i Marta, perquè van ser els primers estudiants de doctorat amb els que compartia temps i espai. L'energia amb què vivien aquells anys de doctorat, quan jo acabava d'aterrar al grup, va animar-me a fer una tesi doctoral. Al Jordi, per ser un referent en perseverança, i per saber treure també sempre un tema de discussió mentre dinàvem. A la Michelle i la Laia, per deixar-me acompanyar-les durant els seus treballs de grau i màster. A Ceci, por todo el conocimiento compartido, siempre era un gusto escucharte hablar de ciencia. To Leila, for such an amazing energy and light, for every life lesson and conversation, for knowing me better than I, and trusting, even the short time we shared at BSC. A la Montse, per ser un pilar pel grup. Per cada estona de cafè i conversa, per cada vegada que m'ha ajudat amb tot allò més tècnic, per ensenyar-me a treballar amb un supercomputador i explicar-me tantes vegades com fes falta, quantes cpus necessitava. Gràcies, de veritat per seure al meu costat i dedicar tantes hores com necessités. A la Romina, per tot el coneixement que ha compartit i pel temps que ha dedicat sense cap problema a ajudar-me tot i el volum de feina que sempre tenia, per la seva perseverança i ganes d'aprendre, però també per preocupar-se sempre per com estava. Per enviar un missatge, un e-mail o passar-se pel nostre lloc a preguntar. I per fer i portar-nos els millors pastissos, d'això no me n'oblido. A la Lorena, per ajudar-me amb tot allò relacionat amb l'estadística i les matemàtiques. Però també per cada estona compartida al sol, fent un cafè i xerrant de ciència, i de la vida en general. Pel seu positivisme inacabable. A la Luisa, amb qui he compartit temps dins i fora de la feina, festivals, passejos per la ciutat, galetes salades amb arequipe, i sessions d'acrioga. Gracias por estar y compartir todos estos años, conocimiento, experiencia, y tiempo. A Dani, Daniel, que suerte haberte tenido al lado. Por cada

abrazo, por ser una figura esencial durante estos años y seguir ahí ahora. Aún no tengo claro porque somos amigos, pero me hace feliz saber que aunque estes lejos, siempre puedo escribirte y compartir penas y alegrías. A Migue, que alegría encontrarte cada día en el BSC, que alivio saber que estabas ahí para lo que hiciera falta. Gracias por estos últimos años, por ser compañero pero sobretodo amigo, por cada llamada, por ser el punto de unión del grupo de amigos que hemos acabado construyendo, por tener un titulo en liar a la gente para hacer planes.

Fora de computational genomics, ha sigut sempre una alegria i un plaer trobar-me a altres persones pels passadissos del BSC, també tenir equips i persones disposades a ajudar i fer-nos la feina més fàcil. Al José, l'encarregat de seguretat a capella, per cada bon dia, per entretenir-me a mig matí mentre feia la seva ronda i fer-nos riure. A l'Eva Navarrete i a Maria Emilia, por ayudarme con cada gestión. A l'equip de suport, per donar-me un cop de mà instal·lant programes i fent funcionar els algoritmes. I a l'equip de helpdesk, per ajudar-me SEMPRE amb qualsevol cosa que li passés al meu ordinador. Recordo els mesos en els quals cada dia tenia problemes, i sempre hi éreu allà per donar un cop de mà. En especial, al Yassine, per ensenyar-me a encendre l'ordinador quan pensava que em faltaven cables, per venir a solucionar el que calgués al moment, però sobretot per haver sigut un gran amic. Per cada conversa i riures BSC-Sants estació, per les hores de berenar i per adoptar-me a l'hora de dinar. Gràcies, per ser el millor helpdesk, per portar-nos menjar marroquí i per creure't el millor psico, ho fas prou bé. Gràcies també per continuar sent el meu helpdesk no-oficial. Al millor grup d'organització de campaments, sopars, sortides, festes, i el que calgui, Sergio, Sergi, David, Joan, Ángel, Blanca, Elisa, Ignasi, Pepe, Wini, Roc (i Migue, clar), que guai i divertit compartir amb vosaltres tantes estones i continuar-ho fent (tot i ja sigui una exBSC). A la Wini, també per escoltar-me quan he necessitat desfogar-me d'allò que sentia poc just. A Nataly, por ser una referente, por cada consejo, por valorarme y mojarse por

mi, pese a las consecuencias que eso tuvo. Gracias de verdad, todo siempre se pone en su lugar con el tiempo. I en general, a tothom del BSC amb qui he compartit alguna conversa.

A totes i tots els companys de feina del VHIO, perquè tot i que aquesta tesi encara no ha acabat, la vida ha continuat amb una nova etapa. Gràcies per acollir-me i fer més fàcil començar a un lloc nou. A Fran Martínez, per la confiança, l'oportunitat de continuar el meu camí a la ciència tot i no haver acabat encara aquest doctorat, i la paciència tot i que aquest final hagi sigut més llarg del que pensava.

Agrair també a la Dra. Sabine Oertelt-Prigione, a Linda Modderkolk i a totes les companyes amb les quals vaig compartir una setmana de summer school a Nijmegen. Gràcies per la inspiració, l'energia, i totes les lliçons apreses sobre un camp tan important en la ciència i la salut com és el sexe i el gènere.

La sort d'aquests anys, ha estat poder compartir temps també amb gent estimada fora de la feina i la rutina. I encara més, quan aquestes persones també m'entenien i acompanyaven en el procés. Han sigut molts anys, cadascú ha compartit amb mi una part del seu temps, d'una manera o una altra, quedant-se, o marxant. Però ho valoro i agraeixo.

De l'escola, a la Maria, per ser-hi quan ho he necessitat i també quan no. A les amigues de la universitat, que després de ja 11(?) anys, continuen fent camí amb mi. Clara, gràcies per escoltar i comentar sempre les nostres respectives tesis, compartir angoixes, hores de teletreball, per cada sortida, viatge, i passeig. Vetux, gràcies per ser la mami (ara ja de veritat), compartir maduresa, riures, dinars i vermut. Prunus, gràcies per ser-hi, per reunir-nos, i escoltar-nos sempre a totes, amb paciència i sense judicis. Lenux, per les festes i els balls juntes, però també per ser un espai segur on compartir, per compartir moments bons, i també els que no ho han sigut tant, per la confiança. Chikis, gràcies per

cada trucada, tot i la distància sempre t'he sentit a prop, per cada consell sincer, per ballar cada cançó i sobretot, per cada abraçada. Probablement, avui no seria escrivint això, si no fos per la Judit Pinteño, qui em va animar a escriure i fer pràctiques al grup de Computational Genomics del BSC. A la Carla, per seure el primer dia de màster al meu costat i donar-me conversa. Gràcies per quedar-te, per sempre estar al dia, per escoltar-me amb tant respecte i carinyo.

A les Young IT Girls, quin viatge juntes. Ja fa uns anys que caminem en equip. He après tant de totes vosaltres. Gràcies per la confiança, i per compartir motivació. En especial, mencionar a l'Alejandra, per haver sigut una referent, per fer-nos sempre riure, per la seva sinceritat i per haver-me enredat però també per haver confiat sempre. Gracias por compartir dentro, y fuera de YITG. A la Elisa, perquè a més de YITG, també la tenia cada dia al BSC. Per seure amb mi a la terrassa de la 4^o planta aquell dia, i començar a tenir-nos més presents. Per cada hora de dinar, berenar, i el que fes falta. A María Fernández, por cada concierto juntas y por ser la mejor secretaria que una presi puede tener, a la Olivia, per la seva motivació i ganes, i per acollir-me l'etapa que comença (o ja ha començat) postdoctorat, a l'Ona, per pensar en mi i proposar-me activitats tan xules i enriquidores, a l'Helena, la Maria Gil, l'Esther, la Marina, i l'Anna Canal.

A aquelles persones que es creuen amb tu, i comencen a fer-se un lloc o agafar una posició rellevant. A l'Igsus, per escoltar els meus contes de genètica i ensenyar-me sobre música, per les truites de tardor, i per compartir l'inici d'any a Gàmbia, les meves últimes vacances abans de començar aquest tancament. Per donar-me el Nobel sense ser encara ni doctora. A mis dos tatuadores preferidos. Fer(sito), gracias por ayudarme a encontrar las palabras, por dejarlas marcadas en mi piel, y deja el recuerdo d todo este final de etapa. Por ver en mi más allá de lo que yo veo y confiar. Que ilusión tenerte. Pedri, gracias por compartir conmigo muchas de tus primeras veces en Barcelona, Costa Brava y Pirineos, pero sobretodo por enseñarme sin quererlo, a vivir el presente de forma tan intensa,

pese a ser la persona que más vive pensando en el futuro. A tots els nens i nenes d'oncologia infantil de Vall Hebron, ja van 7 anys acompanyant-vos i aprenent de totes i tots vosaltres. Gràcies per posar-hi llum, per ser la motivació per la qual treballar en això, per deixar-me seure al vostre costat i donar-vos la mà. Al Ricard, el meu terapeuta i psicòleg, per tot i ser la seva feina, escoltar-me amb paciència, mostrar-me el camí i ajudar-me a conèixer-me. Gràcies per cada sessió, per sostenir-me fins aquí i repetir-me tantes vegades com ha fet falta cada lliçó. A la Ela per ensenyar-me a ballar dancehall, per fer d'animadora i motivar-nos a cada classe, i a la Xènia per ensenyar-me a fer ceràmica i gestionar la frustració quan una peça no surt com esperes o es trenca.

A qui ha sigut casa des de que va començar aquesta etapa, i per sort, des de que tinc 12 anys i em recordava quins deures teníem per l'endemà. Èlia, gracias por estar, por ser mi amiga desde hace tantos años, por crecer conmigo no solo porque los años pasen, por tantas cosas compartidas, escucharme, entenderme y ponérmelo fácil. Por hacerme un hueco a tu lado, en la mesa, en el sofá, o en la cama cuando me despertaba con ansiedad.

A la Alba, por darme siempre el empujón que necesito. Que suerte tenerte desde siempre. Que suerte poder contar contigo aunque no estés cerca. Gracias por escucharte cada uno de mis audios aunque duren 15 minutos, y por enviarme otros de vuelta y así andar por la calle contigo, aunque no estés físicamente. Gracias por ayudarme en cada decisión, por conocerme tan bien y darme siempre el mejor consejo, por entenderme, por la paciencia y la sinceridad. Por cada verano, y desde hace años, invierno, primavera o lo que haga falta. Por hacerme de espejo y ir de mi mano.

Al Roc, per ser-hi, acompanyar-me, cuinar-me els millors esmorzars, dinars i sopars, i cuidar-me, en general, durant aquest tancament. Gràcies per confiar en mi, per sentir-te sempre orgullós, per escoltar-me i compartir. Per les ganés i

el flow conjunt. Per animar-me a fer les següents passes, i per donar-me la mà i caminar amb mi. Gràcies per cada abraçada i ser un espai segur.

A la meva família, per ser casa des de que vaig néixer i acompanyar-me en cada etapa vital i professional. Als meus avis per sempre estar orgullosos de mi. A les meves germanes. Miriam, gracias por la paciencia, por ser referente en perseverancia y esfuerzo, por ser la hormiga que llega donde se propone, por poner calma siempre que hace falta. Judit, gracias por tu valentia, por ser de las tres, la que no tiene miedo al cambio y la aventura, por el carácter. A las dos por darme siempre la mano, ir a mi lado y confiar. A vosotros, papa y mama, por darme la libertad de escoger el camino que quisiera, por educarme todo lo bien que habéis sabido, por hacer todo lo posible para que no nos haya faltado nada y por abrir vuestros brazos siempre que lo necesito.

I a mi. Al meu cap, al meu cos i al meu cor. Per haver escollit això, arribat fins aquí i voler seguir. Per l'esforç, la capacitat i la motivació.

Fin.\n

*Al Z., la L., la F., la M., i el S., de la planta
d'oncologia pediàtrica de l'Hospital de
Vall Hebron.*

Abstract

Cancer, a complex disease, arises from accumulated somatic genomic and epigenomic changes within tumor cells, typically acquired during an individual's lifetime. These alterations confer growth advantages, transforming normal cells into cancerous ones. Differences among tumors originated in the same tissue, have been demonstrated and characterized in diverse studies using large cohorts of patients, such as The International Cancer Genome Consortium (ICGC) or The Cancer Genome Atlas (TCGA). Furthermore, it is known that each tumor can be formed by many cell populations, each accumulating different somatic genetic mutations. This knowledge has put into question the traditional classification of tumors, and how they are treated. Advancements in genome technologies, such as next-generation sequencing, have played an important role in generating vast amounts of tumor datasets, allowing sophisticated and ambitious bioinformatic analyses. These technologies have been essential to comprehend tumor formation and progression, and their potential translation into the clinics.

Using large-scale and public initiatives of cancer data, and through the combination of genomic and transcriptomic analysis, we have been involved in diverse cancer-related studies, primarily focused on the identification and interpretation somatic genomic events. Therefore, the general goal of the work described in this thesis is to expand the understanding of the genomic basis behind tumors, through the analysis of somatic events, like somatic processed pseudogenes and other previously unexplored genomic elements, i.e. micropeptides.

First, in collaboration with Dr. Elias Campo from IDIBAPS, we participated in a longitudinal study of Chronic Lymphocytic Leukemia. In particular, we were focused on the analysis of somatic structural variants, to define and quantify their cell frequency and incorporate them in the study of the subclonal architecture of

CLL patients. Using diverse variant calling pipelines and experimental validations, we first identified SVs and observed an increase in them during tumor progression, particularly evident once the patients transformed into a more aggressive form known Richter's syndrome. We then designed a strategy to calculate SV variant allele frequencies. This involved exploring coverage variability and read alignment within these mutated genomic regions. Based on this analysis, we could observe stable or decreased SV frequencies at diagnosis, contrasting with increase at Richter transformation.

Another part of the thesis has been conducted in the context of the Pancancer Analysis of Whole Genomes initiative, where we studied the landscape of processed pseudogenes in 2585 cancer genomes and assessed their potential functional impact. PPs represent mRNA copies randomly integrated into the genome through retrotransposition. Prior to our study, these events were described as somatic in only a few tumor types. We established a protocol based on automatic rules applied to somatic structural variants and manual inspection of the genomes, to detect such somatic event. We found evidence for 433 candidates somatic PPs across 251 tumor genomes, uncovering new cancer types not examined before. Additionally, as a first approximation to study their functional impact and using RNA-seq data exploration, we identified evidence of expression of 17 PPs across 6 tumor types. The reconstruction of the potential PP-host gene fusion transcripts allowed us to predict that these insertions generally generate premature stop codons within the coding region of the host.

Finally, we focused on the identification of novel micropeptides, a recently discovered class of genetic elements. Micropeptides are small open reading frames of less than 300 nucleotides that can code for stable and functional small proteins. Among other observed functions, it has been shown that these small peptides can suppress cancer growth and have important roles in cancer. We used publicly available genomic and transcriptomic data to identify new micropeptides,

focusing on non-annotated DNA regions. First, in collaboration with Dra. Maria Abad from VHIO, we defined a catalog of more than 1.000.000 candidate micropeptide sequences in non-annotated regions. To do so, we performed de novo transcriptome assembly of 6 RNA-seq samples from pancreatic adenocarcinoma human tissues, merged the predicted transcripts and in-silico translated their sequences. Results were filtered to remove sequences overlapping with known coding sequences and depending on their expression values. The dataset was then used for analyzing pancreatic tumor samples with mass spectrometry analysis. Secondly, complementing this collaboration, we lead a different study focusing on the identification of new small ORFs within non-annotated regions of the human genome. Based on evolutionary conservation features at DNA and protein level, we identified a set of 8.289 candidate smORFs within intergenic regions of the human genome. We then also analyzed their potential transcription on 135 normal samples from the GTEX project, including 28 tissues. From this data, we could find expression evidence for 260 candidate smORFs in at least one normal sample. Lastly, with the aim of exploring the role of micropeptides in cancer we analyzed recurrence of somatic SNVs from the ICGC. However, to date, we have not identified any cancer driver mutations within these smORFs. We hope that extending this comparison to other collections of somatic variants related to cancer can identify candidate cancer smORFs

Collectively, the presented thesis offers a comprehensive description of somatic genomic events in cancer focusing on structural variation and processed pseudogenes, as well as the evaluation of novel gene elements, providing a foundation for future investigations.

Abbreviations

A	Adenine
aa	amino acid
AML	Acute Myeloid leukemia
BAM	Binary alignment map
BKP	Breakpoint
Blastn	nucleotide Basic local alignment search tool
Blat	Basic local alignment tool
BOCA-UK	Bone cancer - United Kingdom project
BRCA	Breast cancer
BRCA-FR	Breast cancer - France project
BSC	Barcelona Supercomputing Center
BTCA-SG	Biliary tract cancer - Singapore project
C	Cytosine
CCF	Cancer cell fraction
cDNA	complementary DNA
CDS	Coding sequence
ChIP-seq	ChIP-sequencing
chr	Chromosome
CLL	Chronic Lymphocytic Leukemia
CLLE-ES	Chronic Lymphocytic Leukemia - Spain project
CML	Chronic myelogenous leukemia
CNA	Copy number alteration
CNIO	Centre Nacional de Investigaciones Oncológicas
CNV	Copy number variant
COAD	Colorectal adenocarcinoma
COSMIC	Catalog of Somatic Mutations in Cancer
cw	cluster window
dbSNP	Single Nucleotide Polymorphism Database
DKFZ	German Cancer Research Center
DLBCL	Diffuse large B-cell lymphoma
dn/ds	substitution ratio

DNA	Deoxyribonucleic Acid
dORFs	downstream ORFs
ENCODE	The Encyclopedia of DNA Elements
ESAD	Esophageal adenocarcinoma
ESAD-UK	Esophageal adenocarcinoma - United Kingdom project
e-value	expected value (blast analysis)
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
G	Guanine
GACA	Gastric cancer
GACA-CN	Gastric cancer - China project
GBM	Glioblastoma multiforme
gnomAD	Genome Aggregation Consortium
GO	Gene Ontology
GRIPs	Gene retrocopy insertion polymorphisms
GTEx	Genotype-Tissue expression
GTF	General Feature Format
GWAS	Genome-wide association studies
HapMap	Haplotype Mapping
HGP	Human Genome Project
HNSC	Head and neck squamous cell carcinoma
HPC	High-performance computers
ICGC-ARGO	Acceleration Research in Genomic Oncology
IGHV	Immunoglobulin heavy variable
IGV	Integrative genome viewer
Indels	Insertions and deletions
IQR	Interquartile range
ITH	Intratumor heterogeneity
JSON	JavaScript Object Notation
Kb	Kilobase
KS	Kolmogorov-Smirnov
L1	LINE-1
LINC	Liver cancer
LINE	Long interspersed elements
LN	Lymph node

IncORFs	Long non-coding ORFs
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
LUSC-KR	Lung squamous cell carcinoma - South Korea project
LUSC-US	Lung squamous cell carcinoma - United States project
MAPQ	mapping quality
Mb	Mega base
MS	Mass spectrometry
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
NSCLC	non-small cell lung carcinoma
ORF	Open-reading frame
OV	Ovarian carcinoma
OV-AU	Ovarian carcinoma - Australian project
PACA	Pancreatic adenocarcinoma
PACA-AU	Pancreatic adenocarcinoma - Australian project
PACA-CA	Pancreatic adenocarcinoma - Canada project
PAML	Phylogenetic Analysis by Maximum Likelihood
PB	Peripheral blood
PCAWG	Pan-Cancer analysis of Whole Genomes
PE	Paired-end
PP	Processed pseudogene
QQ-plot	Quantile-quantile plot
Q3	3rd quartile
RBH	Reciprocal Best Hit
REST	REpresentational State Transfer
Ribo-seq	Ribosome sequencing
RNA	Ribonucleic Acid
RNA-seq	RNA-sequencing
RPFs	Ribosome-protected RNA fragments
RT	Richter transformation

SAM	Sequence alignment map
SBS	Sequencing by Synthesis
simw	simulation window
SKCM	Skin cutaneous melanoma
smORF	small open reading frame
SMuFin	Somatic Mutation Finder
SNV	Single Nucleotide Variant
STAD	Stomach adenocarcinoma
SV	Structural variant
SvABA	Structural variation analysis by assembly
sw	smoothing window
T	Thymine
tblastn	translated nucleotide blast
TCGA	The Cancer Genome Atlas
TE	Transposable element
TMB	Tumor mutational burden
TPM	Transcripts Per Million
TraFiC	Transposome Finder in Cancer
tsv	Tab-separated values
UCSC	University of California Santa Cruz
UNICORNs	Unannotated intergenic constrained regions
uORFs	Upstream ORFs
UTCA-FR	Uterine Cancer - France project
UTR	Untranslated region
VAF	Variant Allele Frequency
VCF	Variant Caller Format
VEP	Variant Effect Predictor
VHIO	Vall Hebron Institute of Oncology
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
ws	windows size

Table of contents

Acknowledgments	iii
Abstract	xiii
Abbreviations	xvii
1. Strategy and thesis trajectory	1
2. Introduction	5
2.1 The information storage system of humans: the genome.....	7
2.1.1 History of genetics: from Darwin to the Human Genome Project.....	9
2.1.1.1 Women’s contribution to genetics.....	15
2.1.2 The post-genomic era	17
2.2 DNA and RNA studies in the post-genomic era.....	19
2.2.1 Reading nucleotides: Next-Generation Sequencing technology	20
2.2.2 Assembly process: reconstructing the sequence	24
2.2.3 Integrating sequencing data in biomedical sciences. Progression towards precision medicine.	26
2.3 Cancer: a collection of complex diseases.....	30
2.3.1 The hallmarks of cancer: decoding the complexity	33
2.3.2 Somatic variation in the human genome.....	37
2.3.2.1 Types of somatic variants.....	39
2.3.2.2 Variant calling analysis to describe the somatic variation landscape of tumors.....	42
2.3.2.3 Public databases and catalogs of genomic variants.....	45
2.3.3 Driver and passenger mutations in cancer	46
2.3.3.1 Identification of cancer driver genes through bioinformatic approaches.....	47
2.3.4 Intratumor heterogeneity and clonal dynamics	49
2.3.4.1 High-throughput sequencing analysis to decipher cell populations	52
2.3.5 Large-scale initiatives promoting cancer research.....	54
2.3.6 Challenges in cancer research.....	56

2.4 Processed pseudogenes: a by-product of L1 retrotransposition	57
2.4.1 Somatic retrotransposition events in cancer	64
2.4.2 Using NGS data to identify somatic retrotransposition events	67
2.5 Translated small open reading frames: micropeptides	71
2.5.1 Classification of small ORFs	75
2.5.2 Identification of micropeptides.....	77
2.5.2.1 Computational annotation through in-silico evolutionary approaches.....	78
2.5.2.2 Ribosome profiling to monitor translation	83
2.5.2.3 Mass spectrometry to directly detect peptides	85
2.5.3 Published databases to study micropeptides	88
3. Motivation and objectives.....	93
4. Methods	99
4.1. Analysis of somatic structural variants in CLL and their incorporation into subclonality studies.....	101
4.1.1 Chronic lymphocytic leukemia longitudinal study cohort.....	102
4.1.1.1 Disease course of one pilot CLL case.....	103
4.1.2 Whole genome sequencing and alignment	104
4.1.3 Somatic structural variants identification.....	105
4.1.3.1 Variant caller programs.....	105
4.1.3.2 Variant validation through manual inspection of aligned sequencing reads.....	106
4.1.3.3 Filtering, merging and consensus variant calling results	108
4.1.3.4 Rescue of somatic structural variants from longitudinal samples	109
4.1.4 Inferring structural variant allele frequencies to analyze intratumor heterogeneity.....	110
4.1.4.1 Analysis of aligned tumor reads in an in-silico sample	111
4.1.4.2 Calculating the variant allele frequency for in-silico structural variants to define a strategy	112
4.1.4.3 Applying the designed strategy to CLL longitudinal samples.....	114

4.1.4.4 Deducing cancer cell fraction of structural variants and clonal dynamics for one pilot CLL case	114
4.2. Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact.....	117
4.2.1 Genomic and transcriptomic cancer data	118
4.2.2 Somatic processed pseudogenes identification.....	119
4.2.2.1 Genomic data analysis.....	120
4.2.2.1.1 Candidate PP selection through VCF files	120
4.2.2.1.2 Manual validation: inspection of tumor sequencing reads	122
4.2.2.2 Generation of an automatic protocol	125
4.2.2.2.1 Pilot exploration of one candidate PP	125
4.2.2.2.2 Protocol development for the complete analysis	126
4.2.2.2.3 Final PP searching strategy.....	129
4.2.3 Expression evaluation of acquired PPs.....	131
4.3. Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data	133
4.3.1 Transcriptomic data from pancreatic adenocarcinoma.....	134
4.3.2 De novo transcriptome assembly	134
4.3.3 Transcriptome combination of multiple samples analyzed	136
4.3.3.1 StringTie transcript merge mode	136
4.3.3.2 In-house strategy to obtain a consensus set.....	137
4.3.3.2.1 Merging step through transcript clustering	137
4.3.3.2.2 Definition of a consensus sequence and selection of representative transcripts.....	138
4.3.4 <i>In-silico</i> 3-frames translation of de novo consensus transcripts.....	139
4.3.5 Local alignment search to remove overlap with annotated CDS.....	141
4.3.6 Candidate micropeptides selection based on expression for MS analysis	142
4.3.7 Strategy and final parameters to build candidate micropeptides datasets	143
4.3.8 Collection of known and conserved intergenic human regions.....	145

4.3.9 <i>In-silico</i> translation of intergenic constrained regions.....	146
4.3.10 Searching for orthologs on <i>Mus Musculus</i> using Reciprocal Best Hit approach	147
4.3.11 Inference of purifying selection based on dn/ds ratio	150
4.3.11.1 Expected dn/ds ratio on known protein coding genes	151
4.3.11.2 Selection of candidate functional micropeptides	152
4.3.12 Expression analysis of candidate functional micropeptides in normal tissues.....	152
4.3.13 Exploring somatic cancer SNVs within candidate micropeptides to assess their role in tumorigenesis.....	153
4.3.13.1 Applying OncodriveCLUSTL to published small ORFs.....	155
4.3.13.2 Evaluation of recurrent variants within novel candidate micropeptides	157
5. Results	159
5.1. Analysis of somatic structural variants in CLL and their incorporation into subclonality studies.....	161
5.1.1 Identification pipeline for somatic structural variants.....	162
5.1.1.1 Evaluation of the structural variant identification pipeline	162
5.1.1.1.1 Fine-tuning specific parameters used by DELLY2.....	163
5.1.1.1.2 Comparative analysis of somatic structural variant callers	164
5.1.2 Exploring intratumor heterogeneity from structural variant allele frequencies.....	166
5.1.2.1 Sequencing coverage variation in normal and tumor genomes ...	166
5.1.2.2 Identification of variant supporting reads in an in-silico sequenced sample	170
5.1.2.3 Variant allele frequency estimation of artificial structural variants	174
5.1.3 Applying the define methodology to longitudinal CLL samples: case 63	177
5.1.3.1 Somatic structural variant landscape.....	177
5.1.3.2 Frequency and evolution of structural variants during tumor progression.....	181

5.2. Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact.....	185
5.2.1 Analysis of a lung squamous cell carcinoma genome	186
5.2.1.1 Identification of somatic structural variants supporting PPs formation	186
5.2.1.2 Reconstruction of CNH4 pilot processed pseudogene	187
5.2.2 Automatic search of PPs across all LUSC tumors based on diverse criteria combinations	192
5.2.2.1 Dataset 1: evidence of one insertion point.....	192
5.2.2.2 Dataset 2: evidence of one insertion point and splicing events ...	193
5.2.2.3 Dataset 3: evidence of two insertion sites.....	194
5.2.2.4 Dataset 4: evidence of both insertion sites and splicing events within the source genes	194
5.2.3 Identification of somatic processed pseudogenes in all PCAWG tumor genomes.....	197
5.2.3.1 Manual validation of candidate PPs previously identified across PCAWG cohort.....	199
5.2.4 Evaluation of potential PP-host gene fusion transcripts.....	204
5.3. Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data	207
5.3.1 Predicting non-reference-based novel transcripts for pancreatic adenocarcinoma samples.....	208
5.3.1.1 Calibrating StringTie	209
5.3.2 Assessment of transcript clustering based on different criteria	211
5.3.3 Small open-reading frames datasets: insights from two different criteria	213
5.3.3.1 Dataset version 1: a more conservative set of small ORFs	216
5.3.3.1.1 De novo transcript prediction allowing multi-mapped reads....	216
5.3.3.1.2 Consensus set of predicted transcripts using StringTie algorithm	216
5.3.3.1.3 In-silico translation of coding DNA.....	218

5.3.3.1.4 Filtering micropeptides based on their overlap with known CDS	219
5.3.3.2 Dataset version 2: inclusion of non-canonical start codons and expression-based filtering for small ORFs	220
5.3.3.2.1 De novo transcript prediction based on unique mapped reads	221
5.3.3.2.2 Consensus set of predicted transcripts applying an in-house merging strategy	221
5.3.3.2.3 In-silico translation of coding DNA considering non-canonical start codons.....	224
5.3.3.2.4 Filtering micropeptides based on their expression and overlap with known CDS	225
5.3.4 UNICORNs: highly evolutionary constraint intergenic regions	229
5.3.5 In-silico translated small ORFs located in intergenic regions.....	231
5.3.6 Candidate ortholog sequences of translated intergenic small ORFs ...	232
5.3.7 Calculated dn/ds ratio on known protein-coding genes.....	235
5.3.8 Catalog of candidate intergenic micropeptides from highly conserved regions.....	237
5.3.9 Preliminary evidence of expressed candidate functional micropeptides	239
5.3.10 Detection of significant clusters of somatic cancer mutations in published smORFs	242
5.3.11 Low number of somatic SNVs acquired in intergenic novel candidate micropeptides	246
6. Discussion.....	249
7. General overview	279
8. Conclusions.....	285
9. References.....	291
10. Annex.....	311
10.1 Supplementary figures	313
10.2 Supplementary tables	318
10.3 Publications	329

1. Strategy and thesis trajectory

Before starting, I would like to describe the trajectory of the presented thesis to expose the bases of our decisions and the strategy we followed. The central concept of my thesis, shared between all the projects I have been working on, is studying somatic variations in cancer using large-scale datasets. These studies are the basis to better understand tumor formation and progression and later apply genomic oncology at the clinics. Under this wider objective, and for the sake of clarity, we have divided this thesis in three chapters, each covering a specific goal and strategy. As described below, the first one deals with our contribution in the field of somatic SVs identification and characterization in tumor, in particular within intratumor heterogeneity; the second covers our study of somatic processed pseudogenes in cancer and their potential functional impact; and the third one, divided in two different studies, describes our current and final project where we aim at finding and characterizing new micropeptides in the context of cancer.

Our group was involved in diverse studies regarding the Pan-Cancer analysis of Whole Genomes (PCAWG) project since the beginning of this international collaboration. For this reason, we had the opportunity to work with all the data from PCAWG before its publication and explore somatic variation in more than 20 tumor primary sites. Motivated by recent publications about retrotransposition and processed pseudogenes (PP) in cancer, we decided to explore these last somatic events using the PCAWG data. Starting from somatic structural variants previously identified by the consortium, I designed a strategy to identify somatic processed pseudogenes in diverse cancer types. Then I used aligned tumor and normal genome sequencing reads to manually validate the candidates we obtained. Finally, I explored if they could have any functional impact. To do so, I analyzed RNA-seq data from the same patients to look for fusion genes. This work

ended up in one of the main PCAWG articles, published in Nature Genetics in 2020 (1).

Before this work was published, we started a collaboration together with Dr. Elias Campo and Dr. Ferran Nadeu from the IDIBAPS. The project was focused on Chronic Lymphocytic Leukemia (CLL) and the study of intratumor heterogeneity (ITH) and progression of disease. Using longitudinal samples collected at different time points from around 20 patients, the main goal was to understand Richter's transformation, an aggressive alteration of CLL with dismal prognosis. Considering the experience I gained with the previous project, my tasks within this collaboration were centered on the identification of structural variants and their classification, as to VAF. Firstly, I analyzed the results obtained from diverse variant callers to filter and merge them. I performed this validation by combining automatic and manual inspection. Therefore, we could end with an accurate set of structural variants for each sample used for later published analysis (2,3). Secondly, led by us and together with Romina Royo, we aimed to include structural variation into the characterization of subclones and the description of CLL intratumor heterogeneity. At that time, ITH was mainly studied using single nucleotide variants and all pipelines used for identifying subclonality systematically avoided structural variants. Using data from the CLL project, I explored how to calculate the frequency of structural variants by identifying sequencing reads covering each variant. Counting supporting reads was a challenge because of the coverage variation among each sequenced sample. Our strategy was based on including somatic structural variants within subclones previously identified using single nucleotide variants. But while doing this work, a computational method for inferring SVs cancer cell fraction was published. In fact, it was also one of the articles from PCAWG project (4). We expected to explore ITH including structural variants in diverse cancer types and using longitudinal

samples. This clearly limited our publication options to the point that we decided to prioritize another project.

In parallel, in 2020, we collaborated with Dra. Maria Abad and Marion Martínez from the Vall d'Hebron Institute of Oncology (VHIO). The project was aimed at the identification of micropeptides in pancreatic adenocarcinoma using mass spectrometry (MS) analysis. In this project were also involved Dr. Hector Peinado and Dr. Javier Muñoz from the CNIO. In this collaboration, my task was to create a dataset of novel and non-annotated candidate micropeptides to be used for the interpretation of the Mass Spec results. I used RNA-seq data from the same cancer type to be tissue-specific, since transcription and translation are. By doing *de novo* transcriptome assembly of 6 patients, I obtained a set of transcripts including both known and novel ones. I performed an *in-silico* translation of the transcripts to obtain candidate micropeptides and I removed those overlapping with known protein-coding genes. The set of candidates has been used for MS analysis at CNIO. Nowadays, experimental validation of interesting results is done at VHIO to end with a publication.

The knowledge we acquired regarding micropeptides, and the fact that they have not been studied at the genome-wide level in cancer, opened the possibility of searching and identifying these unexplored genetic elements in other cancer types. We could foresee options of publication working on this strategy, as the annotation of micropeptides in general is still highly imprecise and incomplete. For this reason, we decided to focus on the identification of micropeptides and to characterize their potential role in cancer. I mainly focused on micropeptides during my second half of the thesis. Since nearly all published micropeptides are identified within annotated genes, we decided to start the project through the identification of novel candidate micropeptides by exploring highly conserved intergenic regions across the entire human genome to later investigate their potential role in cancer based on somatic mutations acquired within them.

Although this was an ambitious project, we considered we had the tools to start with it. In this line, we looked for micropeptides within conserved regions previously identified in The Zoonomia Project (5). Based on the search of ortholog sequences, we defined a set of candidate human micropeptides. To add more supportive information on these candidate micropeptides, I explored expression levels of our set of conserved and translated regions by analyzing RNA-seq data from GTEX project (6). Finally, intending to explore the role of these novel candidate micropeptides in cancer disease and tumorigenesis, I tested and evaluated OncodriveCLUSTL (7), a driver discovery algorithm and explored all somatic SNVs from The International Cancer Genome Consortium. However, and despite the need for more analysis, we could not identify significant signals due to the low number of somatic mutations present within these candidate micropeptides.

During my PhD trajectory I have also mentored other students. I provided guidance to Michelle Tomaselli for her final degree project. She tested the published algorithm SVClone for calculating structural variants cancer cell fraction. After this, I mentored her final master project focused on the analysis of single nucleotide variants in Endometrial Cancer using topological whole exome sequencing samples. This work was led by Dra. Rosaura Esteve-Puig and Dr. Xavier Matias-Guiu from IDIBELL. I worked on the last figures and analysis together with Romina Royo once Michelle finished her master project, as well as collaborated on writing part of the manuscript. Finally, I served as the tutor of Laia Ollé during her final degree project. She worked on the characterization of micropeptides candidates in pancreatic adenocarcinoma, in particular, through their annotation using ribosome profiling data from the public database sorfs.org. I reviewed their written projects and oral presentations before presenting it at the university.

2. Introduction

2.1 The information storage system of humans: the genome

Deoxyribonucleic Acid (DNA)(8) is the chemical name of the molecule carrying genetic instructions in all living organisms. DNA is their central information storage system, and it consists of a right-handed double helix formed by two strands that wind around one another. The DNA helix is anti-parallel, meaning the 5' end of one strand is paired with the 3' end of its complementary, and vice versa. Both strands are made of a sugar known as deoxyribose and phosphate backbone, which have bases sticking out from it. The strands are held together by hydrogen bonds between the bases; adenine (A) with thymine (T), and cytosine (C) with guanine (G). These nitrogen bases are also exposed and available for hydrogen bonding for other molecules that play vital roles in replication and expression. Structurally it is organized into chromosomes and functionally, into genes. Genes are the basic unit of heredity (9), containing the information for building one or more molecules that help the body work. Genes usually code for proteins, each of them with a particular characteristic or function. They are located on a chromosome and consist of nucleotides arranged in a linear manner. It is estimated that humans have about 20.000 genes (10). All the genetic material of an organism, including genes and other elements controlling the activity of genes is its genome, which in humans, plants and animal cells is housed in the nucleus of nearly all of its cells.

Genetic information together with environmental factors characterize the observable traits of each individual organism, also known as phenotype. However, those characteristics do not come from DNA itself, but from the result of a specific flow of information named The Central Dogma of Molecular Biology. This

describes the transfer of information stored in genes as DNA, which is transcribed into ribonucleic acid (RNA), and translated into proteins (Fig 1). The concept was developed by Crick in 1958, and it states that information cannot be transferred from protein to either protein or nucleic acid (11). The transcription(12) of a subset of genes into RNA molecules also describes a cell's identity and its biological activity. All these RNA molecules are collectively defined as the transcriptome, which can differ between cell types and time. The transcriptome is also essential for understanding development and disease. Although most of the observable trait's information comes from genes and individual proteins, it is now known that untranslated RNAs can actually be involved in the phenotype.

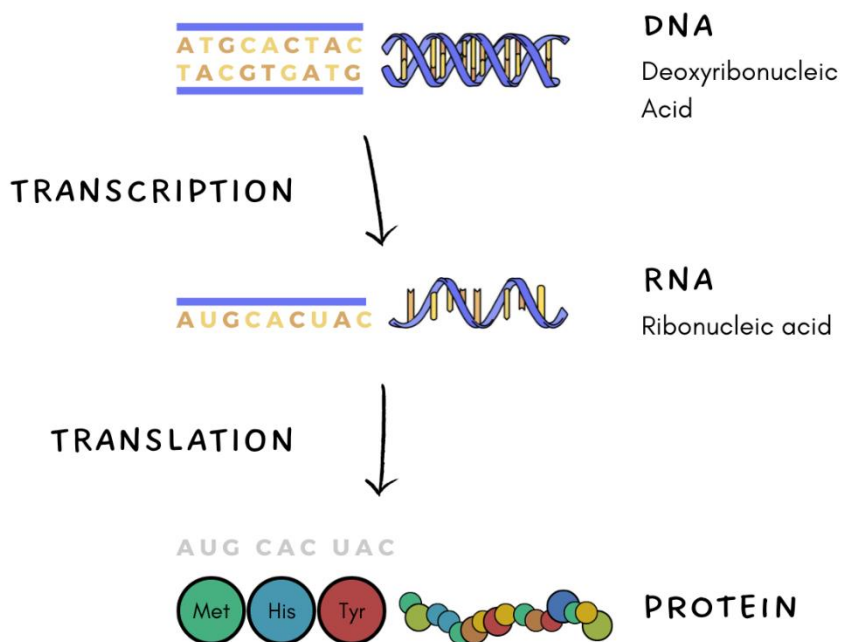


Figure 1. Central dogma of molecular biology.

2.1.1 History of genetics: from Darwin to the Human Genome Project

Based on the definition of the National Institute of General Medical Sciences (10), genetics is the scientific study of genes and inheritance in living organisms, and in particular of how certain qualities and traits are passed from parents to offspring as a result of changes in DNA sequence.

In the 19th century, scientists started questioning why children resemble one parent more than others or why some species have similarities between them more closely than others. Scientists could observe similarities between the offspring of animals and plants, but they could not understand why this happened. These observations were the starting point of genetics.

In particular, it was in 1858, when Charles Darwin received a manuscript from Alfred Russel Wallace exposing an evolution theory based on natural selection. This theory coincided with the ideas about the evolution of species Darwin was working on. One year later, Darwin presented and published together Wallace's work, "The origin of species," which describes how new species arose via evolution and how natural selection uses natural variation to evolve new forms (13).

A few years later, in 1865, Gregor Mendel presented his research on inheritance in pea plants in the scientific journal *Verhandlungen des naturforschenden Vereines*. Mendel tracked several phenotypes in peas across different generations, developing homozygous lines, observing the offspring of each kind of parent and analyzing the data statistically. This was the first empirical evidence that traits were passed down measurably from parent to offspring and the only approach utilized to understand genetic inheritance. Gregor Mendel could describe the unit of heredity as a particle that does not change. Together with Darwin's work, his study suggested that all species might be related between

them, and because of inheriting different traits, they might drift apart through natural selection. At the same time, Haeckel predicted that the hereditary material was located in the nucleus (13).

It was in 1869 when Swiss physiological chemist Friedrich Miescher tried to isolate and characterize the protein components of leukocytes. During his experiments, he came across a substance from the cell nuclei. It has chemical properties unlike any protein, including a much higher phosphorus content and resistance to proteolysis. With this, Miescher identified in 1871 what he called “nuclein” and demonstrated the material in the nucleus was what we now know as nucleic acid (14). However, during this century, research was usually performed in isolation, and genetics advanced slowly.

In 1900, other scientists performing similar experiments to Mendel’s work arrived at the same conclusions and cited his work in their publications. Subsequent to the rediscovery, linkage, lethal genes, and maternal inheritance were described.

By the early 20th century, powerful light microscopes allowed scientists to see into a cell’s nucleus. The observation of chromosomes combined with The Chromosomal Theory of Inheritance (Walter Sutton and Theodor Boveri, 1904), which defines the chromosome as the location of genes, linked them with trait inheritance. They could also observe that chromosomes occur in matched pairs in humans, one from the mother and one from the father. Chromosomal abnormalities such as duplications, deletions, translocations, or inversions were reported for the first time.

Although Miescher determined the material in the nucleus was nucleic acid in the early 1870s, the community did not widely appreciate his discovery. In 1910, Albrecht Kossel was awarded for his discovery of the five nucleotide bases: adenine, cytosine, guanine, thymine, and uracil (15). From the 1920s through the

1950s, other scientists continued to investigate the chemical nature of the molecule, and diverse experiments concluded that DNA was indeed the genetic material within the nucleus. The Russian biochemist Phoebus Levene proposed in 1919 that nucleic acids were composed of a series of nucleotides, which were formed of one of four nitrogen-containing bases, a sugar molecule, and a phosphate group. This was Levene's "polynucleotide" model. Therefore, he was the first to discover the order of the three major components of a single nucleotide: phosphate-sugar-base (14). In 1943, Oswald Avery together with Colin MacLeod and Maclyn McCarty, proved that DNA carries genetic information. Although at that time no one knew how it worked, they could demonstrate that hereditary units, genes (16), are composed of DNA, not protein or RNA. Some years later, Erwin Chargaff reached two major conclusions (Chargaff 1950). First, the composition of DNA varies among species, and the same nucleotides do not repeat in the same order. Second, almost all DNA maintains certain properties even within different organisms or tissue types. The amount of adenine is usually similar to the amount of thymine, and the amount of guanine approximates cytosine (14). This second conclusion explains that A is bound to T, and C is bound to G in the DNA structure.

New advances in genetics were applied in medicine, leading to the beginning of modern human genetics in 1949. Moreover, the same year, the first textbook of human genetics was published, and the American Journal of Human Genetics was founded (17).

At King's College in London, by the early 1950s, chemists Rosalind Elsie Franklin and Maurice Wilkins worked with X-ray diffraction to study DNA. They beamed X-rays through the molecule and obtained a shadow picture of the DNA structure by how the X-rays bounced off the components. In January 1953, Wilkins showed the resulting picture, known as "Exposure 51", to James Dewey Watson without Franklin's knowledge. Chargaff's second conclusion together

with this X-ray crystallography work, were crucial to Watson and Crick's proposal regarding DNA structure. In April 1953, Watson and Crick published their famous paper in *Nature*, proposing that the DNA molecule was composed of two chains of nucleotides paired to form a double helix. They also explained how the DNA molecule could replicate itself with high accuracy. For their work, Watson, Crick, and Wilkins were awarded the Nobel Prize in 1962. However, regardless of her contribution, Rosalind Franklin was not named a prize winner (16).

After the discovery of the double-helix, the breaking of the genetic code was the second most important advance in molecular biology. In 1955, Severo Ochoa isolated RNA polymerase, the enzyme that transcribes molecules of DNA into RNA. Ochoa could then make the first synthetic RNA molecules which were essential for deciphering the genetic code. Interpreting the genetic language was the work of Marshall Nirenberg and his team at the National Institutes of Health. In late 1960, Nirenberg and Heinrich Matthaei observed that introducing RNA into a cell-free system resulted in synthesizing proteins, whereas adding DNA did not. After this achievement, they added *E. Coli* extract to 20 test tubes containing a mixture of all 20 amino acids. Each amino acid was radioactively tagged in one test tube. Then, they added synthetic RNA made of uracil to each test tube, finding unusual activity in the tube containing phenylalanine. The UUU triplet was the first codon deciphered. In 1964 Nirenberg and Philip Leder discovered how to determine the sequence of the nucleotides in each codon. Two years later, Nirenberg had deciphered the 64 RNA codons for all 20 amino acids. Together with Khorana and Robert Holley, Nirenberg won the Nobel Prize in 1968 "for their interpretation of the genetic code and its function in protein synthesis"(18–20). Also, in this period, Margarita Salas was working as a postdoctoral in Ochoa's lab. Salas not only found that replication, transcription and translation read DNA in only one direction but also helped showing that UAA triplet represents a stop codon (21) and therefore the end point of translation.

Simultaneously, during 1959 and 1960 new methods for analyzing chromosomes such as cytogenetics and new biochemical assays using cultured cells revealed genetic causes behind many human diseases including cancer. Moreover, the fundamentals of mammalian sex determination were defined. Individuals without a Y chromosome were shown to be female, whereas those with a Y chromosome were male. Culture cells became widely used to study monogenic human diseases (22).

Also, in the mid-20th century, the Darwinian theory of evolution was confirmed. Scientists demonstrated experimentally that mutations could be induced. Therefore, understanding the role of variation together with environmental constraints allowed them to solidify the concept that natural selection was a major factor in evolution (13). Modern Evolutionary Synthesis linked Charles Darwin's theory of evolution with Gregor Mendel's studies regarding genetic inheritance and variation. The term was the result of combining Dobzhansky and Fisher's work. In 1968, Kimura proposed the neutral theory of molecular evolution, which contends that at the molecular level, evolutionary changes and polymorphisms are caused by random genetic drift (23).

In the 1970s, Arber (24) discovered restriction enzymes, molecules that recognize and cut specific short sequences of DNA (25). At the same time, Smith isolated and characterized the first Type II restriction endonuclease (*HindII*) and determined the sequence of its cleavage site (26). Independent studies led to the discovery of reverse transcriptase in retroviruses by Baltimore and Temin, revolutionizing molecular biology (27). Moreover, in 1972, Berg assembled the first DNA molecules combining genes from different organisms (28). This technology, known as recombinant DNA, involves cutting DNA sequences using restriction enzymes and fusing the strands with DNA ligases. The development of recombinant DNA technology opened the way to genetic engineering, allowing researchers to give new abilities or eliminate traits to organisms.

In the late 20th century, the availability of reading nucleotides from the genome becomes the next break in genetics. DNA sequencing techniques were first described in 1977 by Sanger and Gilbert after Salas and her colleague Luis Blanco isolated the DNA polymerase enzyme from the bacterial virus phi29. This enzyme is involved in DNA replication copying each strand into identical DNA molecules. The method published by Sanger (1977) is based on random incorporation of chain-terminating inhibitors by DNA polymerase during in vitro DNA replication, and it has been widely used for 40 years (29). Three years after publishing Sanger sequencing, the first genome was sequenced by Sanger Group. In particular, it was the bacteriophage Φ X174 of *E. coli*. The enzyme isolated and patented by Salas and Blanco is also widely used in forensics, studies of ancient DNA and oncology. Three years after publishing Sanger sequencing, Wally Gilbert, Paul Berg and Fred Sanger shared the Nobel Prize for Chemistry, for pioneering DNA sequencing methods (15). Combining linkage analysis, fine mapping within large pedigrees and Sanger sequencing, diverse human genes linked to rare, monogenic and syndromic diseases were discovered (17).

Thanks to the discovery of another polymerase enzyme (Taq) that can withstand high temperatures without denaturing, the PCR (polymerase chain reaction) technique was reported by Mullins (1983). Due to all these findings, in 1986 (30), Hood, Smith and Hunkapiller launched the first automated DNA sequencer. Researchers worldwide came together in consortiums and collaborative groups, and the US Government together with the National Institutes of Health (NIH) established the Human Genome Project (HGP) (1990). The aim of the project was to sequence and map all the genes of our species, *Homo Sapiens*. The first drafts of the human genome sequence, both from the public HGP (31) and private Celera Genomics (32) were published in February 2001. However, it was in 2003 when the Human Genome Project (33) was

completed covering 99 percent of the euchromatic portion of the genome and is accurate to 99.99 percent (34).

2.1.1.1 Women's contribution to genetics

It is well known that women also contributed to genetics research during 19th and 20th centuries. However, institutionalized sexism has prevented them from the recognition they deserve. Most of them are still not mentioned in reviews, articles or books summarizing the history of genetics. Despite significant progress, UNESCO reported on 2021 that only around 33% of the world's researchers are women (35).

Higher education was opened to women in the last three decades of the 19th century, allowing the entry of women into the scientific workforce. However, access to studentships, grants, fellowships, and established careers in universities was absent for them. As the field of genetics was not yet institutionalized, it was one of the earliest emerging disciplines to benefit from their contribution, specifically in Mendelian genetics and heredity. Despite this, the system restricted women to certain roles, even they gained a master's or doctoral degree. The job titles they held were "assistant", "technician", "stockkeeper" or unpaid working wife (36,37).

Considering the importance of their work, to highlight and bring women that contributed to genetics to the fore, and although this section is not essential for the understanding of the presented thesis, a summary of scientific women who made important discoveries is provided here in alphabetical order (38–40).

Barbara McClintock (1902 – 1992): Her studies in maize cytogenetics showed how traits were suppressed or expressed between generations. She also discovered transposable elements, DNA sequences that can change position within a genome. For this work, she received The Nobel Prize in Physiology or

Medicine in 1983, becoming the only woman who has received an unshared Nobel Prize in this field.

Charlotte Auerbach (1899 – 1994): In collaboration with A. J. Clark, and J. M. Robson, demonstrated that mustard gas could induce mutations in *Drosophila melanogaster*. She was most known for discovering mutagenesis.

Edith Rebecca Saunders (1865 – 1945): She was the first collaborator of the geneticist William Bateson, playing an active role in re-discovering Mendel's laws and the study of trait inheritance in plants. Together with Bateson defined terms like alleles, heterozygote, and homozygote.

Elizabeth Blackburn (1948 –): She is most known for her work on telomeres and the co-discovery of telomerase. She was awarded the Nobel Prize in Physiology or Medicine (2009) for this discovery.

Liane Russell nee Brauch (1923 – 2019): Producing many strains of mutated mice she could demonstrate that in mammals, the Y-chromosome determined the animal's male gender.

Margaret Oakley Dayhoff (1925 – 1983): Known as the founder of bioinformatics and one of the first scientists to combine mathematics, computation, and biochemistry. She created the one-letter code for amino acids and originated point accepted mutations. These are replacements of single amino acids in the primary structure of a protein.

Margaret Wu: She developed a statistical tool known as Watterson estimator that approximates the level of genetic diversity in a population contributing to population genetics.

Marie Maynard Daly (1921 – 2003): Her research was focused on the creation of proteins, as well as histones, proteins known to help package DNA into

chromosomes. Daly's research contributed to research into the structure of DNA. She was the first Black woman to earn a doctorate in biochemistry in the U.S.

Martha Cowles Chase (1927 – 2003): Together with Alfred Hershey, they published a paper showing DNA was the biochemical material that transmitted genetic information and therefore DNA was the genetic material of life. The Hershey-Chase experiment helped inspire Watson and Crick to solve the 3-D structure of DNA. Hershey was awarded the Nobel Prize for the discovery, but Chase was not included.

Mary Frances Lyon (1925 – 2014): Working with mice she could demonstrate X-chromosome inactivation, a process by which one X-chromosome is not activated in some female mammals including humans.

Nettie Maria Stevens (1861 – 1912): Using as an experimental model the yellow mealworm, she discovered that the combination of X and Y chromosomes determined the sex of an individual. Her work expanded the fields of modern genetics.

Ruth Sager (1918 – 1997): She investigated how cancer cells grow, multiply and reduce their ability to maintain their chromosome structures. She theorized that a set of genes might be key to halting the growth of cancer and identified over 100 of them. These genes are now named tumor suppressor genes.

2.1.2 The post-genomic era

The publication of The Human Genome Project in 2003 transformed biology and accelerated advancements in the genetic field. HGP was the starting point of the post-genomic era.

Deciphering almost the entire sequence of the human DNA allowed scientists to examine all genes, genetic variants and diseases, initiating the comprehensive discovery and cataloguing of many parts of the human genome.

Questions with implications for biology and medicine became approachable, and experiments that were inconceivable years ago the publication started to be routine. Establishing well-founded correlations between sequence variation and phenotypes enable understanding the architecture of common complex diseases such as diabetes, asthma or cancer as well as rare diseases or behavioral traits. All this knowledge evolves in the personalization of therapies, early detection of disease, ability to follow progression and treatment responses and stratification of disease and patients.

The Human Genome Project inspired subsequent large-scale initiatives and big science projects integrating cross-disciplinary efforts towards human genomics and health. During the post-genomic era, advances in biomedical sciences and expansive research have lent great contributions to better understanding of the human condition and the causes and solutions for several genetic diseases. In addition to GenBank, other online repositories such as the University of California Santa Cruz (UCSC) (41) and Ensembl (42) were created to host genome data in 2002. Projects including Haplotype Mapping (HapMap) (43), 1000 Genomes (44), The Genome 10K (45), or The Cancer Genome Atlas (TCGA) (46) illustrated these great efforts in genomics and the progress of knowledge (47–49).

The community predicted that individual genome sequences will play a larger role in medical practice, and this has happened. In 2011 the first patient saved by DNA sequencing was reported, as his one in 1 billion genetic mutation of *XIAP* gene resulted to be treatable with cord transplant (47,48,50).

The Human Genome Project not only opened avenues in biology and medicine but also in technology and computation. By 2000, the internet was reachable, bandwidth adequate to move genome data and processing power accessible (48). Although these developments were rapidly incorporated into biology,

advancements in bioinformatic tools to store, process, analyze and visualize sequencing data were essential. Therefore, bioinformatic experts and computational biologists emerged. Research groups focused on genomics and working with NGS start combining multidisciplinary experts and usually require sufficient computational infrastructures for data storage and analysis.

2.2 DNA and RNA studies in the post-genomic era

Sanger sequencing, developed in the late 1970s by Frederick Sanger, allowed scientists to read the DNA sequence of genes accurately and was instrumental in several groundbreaking discoveries. However, Sanger sequencing had limitations, primarily in its cost and throughput, making it impractical for large-scale genomic studies. Following the interest generated by the HGP and, because of all the research opportunities sequencing DNA could provide, the development of new technologies rapidly evolved. Sanger sequencing needed to turn into a more automatic, rapid and affordable technology. Companies realized this field could be a successful business, so the market competition gave birth to an overabundance of technologies with progressively higher sequencing throughput at lower costs. Collectively, they were named as Next-Generation Sequencing (NGS) (47). NGS revolutionized the field of genetics by enabling high-throughput, cost-effective sequencing and making large-scale projects feasible. NGS has then become the cornerstone of modern genetic research, allowing scientists to explore complex genomic landscapes with unprecedented depth and speed, and affecting bioinformatic analysis.

2.2.1 Reading nucleotides: Next-Generation Sequencing technology

Pyrosequencing was the first generation of NGS and measured the enzymatic luminometric signal generated by pyrophosphate release during DNA polymerization. 454 Life Sciences commercialized this technique and introduced in 2003 the first DNA sequencer. In a single 4-hour run, the system could produce around 400-500 bp-long reads with 99% accuracy and up to 25 million bp. The same year, a new approach known as Sequencing by Synthesis (SBS) was developed by Solexa and three years later, they launched their first commercial sequencer named Genome Analyzer. This sequencer had a higher throughput in a single run but reads were shorter. Nevertheless, they sequence both DNA strands of each fragment providing paired-end reads separated by a known distance and that enables more accurate read alignment. Solexa company was acquired by Illumina in 2007, and its SBS commercialized approach supports massively parallel sequencing and detect single bases as they are incorporated into growing DNA strands. Illumina HiSeq X Ten sequencer machine allowed in 2014 large-scale whole genome sequencing (WGS) for \$1000 per genome and has the capability to sequence tens of thousands of genomes per year. In 2015, the company was responsible for generating 90% of the world's sequencing data and 70% of the market for DNA sequencers. Nowadays it is still the best sequencing company of the market providing diverse platforms for different applications, all of them with high output (1,2-6000 Gb), high accuracy, low cost per base and diversity of library preparation configurations (47,51).

The classical protocol (Fig 2) for all NGS technologies initiates randomly breaking DNA and creating fragment or mate-pair templates for single or paired-end sequencing respectively. The protocol is followed by size selection and adapters ligation to the end of the fragments. After that, DNA amplification is

generally done. Fragments are PCR amplified only from one end (single-end read), or both (paired-end reads) (Fig 3). These first steps are known as library preparation. Templates are immobilized to a solid surface to allow thousands to billions of reactions to be performed simultaneously. During the sequencing step, nucleotides containing a fluorescent tag and a terminator that blocks incorporation of the next base, bind through natural complementarity to the DNA template. The use of this reversible terminator nucleotide permits one nucleotide to be incorporated at a time and is one of the adaptations over Sanger sequencing. The fluorescent signal indicates which nucleotide has been added and the terminator is then separated to allow the next base to bind to the template. This step is repeated for the length of the fragment end is being sequenced, typically resulting in read lengths between 100 – 400 bp. If paired-end is performed, reads are washed away after reading the forward DNA strand, and the process repeats for the reverse strand resulting in two sequenced ends per template (51–53).

The NGS protocol can be either applied to sequence the entire genome (whole genome sequencing), only the known coding regions (3% of the whole genome), that means the exome (whole exome sequencing, WES), or specific regions (target sequencing). While WGS requires more time and higher costs, WES, which is cheaper, works under the assumption that alterations in proteins usually have a deleterious impact on genome regulation (54). Moreover, next-generation sequencing allows reading not only DNA molecules but also RNA (RNA sequencing, RNA-seq). Sequencers from PacBio or Nanopore can detect nucleotide base modifications in RNA by monitoring reverse transcription in real time, whereas the Illumina approach, among others, sequences the complementary DNA (cDNA) obtained after converting RNA. NGS can also be used to sequence DNA regions where proteins such as transcription factors or chromatin-associated proteins are bound to regulate gene expression (ChIP-seq)

or RNA regions covered by ribosomes (Ribo-seq) to study actively translated mRNAs. Although some steps may change between these techniques to, for example capture a specific molecule, in all cases generated NGS reads are usually outputted in FASTQ files combining the nucleotide base sequence and associated base quality scores.

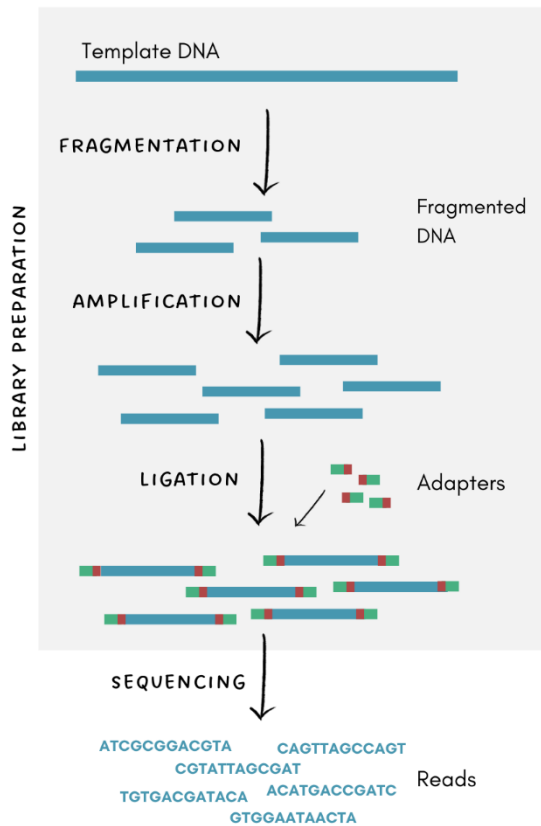


Figure 2. Classical protocol applied in all next-generation sequencing technologies.

Although sequencing methodologies have evolved rapidly to cover many applications and to be reasonable for diverse studies, there are still some challenges to be overcome. The quality and properties of the sample clearly influence the obtained results. However, the two most important challenges of NGS are the read length and the error rates.

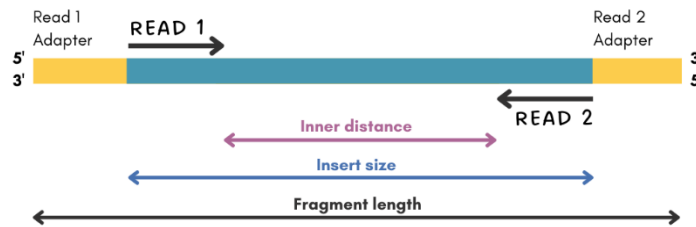


Figure 3. Scheme for paired-end sequencing. Both ends of a DNA fragment are sequenced, and distance between in nucleotides them is known.

After sequencing, obtained reads range between 75 and 900 bp being the most used average length of 100 and 400 bp (55). While short read-lengths comprise of shorter overlapping ends complicating the determination of the preceding and following reads, longer reads simplify this assembly step and require less rounds of the overall process. The shorter the reads are, the more sequences may be similar in nucleotides resulting in ambiguity regarding their precise position and in the inability to resolve repetitive regions. Additionally, short reads are prone to miss larger variants such as insertions or deletions. Paired-end sequencing can help solve these issues by sequencing the same fragment from both ends and providing more positioning information than single-end sequencing. In the other hand, library preparation and the sequencing process itself are associated with sequence errors. The most common type is substitution, where a nucleotide is replaced by another making the identification of variants more difficult. Errors can appear to be platform specific, and distinct sets of nucleotides such as GGT or GGC for Illumina technique, can be associated

with poor sequencing performance. Base misincorporations or rearrangements can also occur during the massive and simultaneous PCR amplification step. For this reason and in order to increase the fidelity of the template sequence, PCR-free library preparation can be applied for short-read sequencing. Lastly, the error rate can also increase when the maximum read length of the platform is approached. Specially if NGS is used for clinical diagnostics or treatment decision making, variants should be validated by visual inspection of the aligned sequencing data or additional Sanger sequencing of the candidate regions (53,56,57).

2.2.2 Assembly process: reconstructing the sequence

Nucleotide reads generated by sequencers are usually far shorter than the size of the genomes investigated when applying NGS protocols. By overlapping these reads, the complete sequence can be deduced. This process is defined as assembly, and it was developed to resolve limitations of current technologies that are not able to sequence the whole genome on a single read. Depending on the sample and type of raw data, this process has diverse flavors including genome, transcriptome or metagenome assembly (47). Nevertheless, it usually starts by filtering low quality reads and correcting errors from library preparation or sequencing and continues computing a set of overlaps to find out the best arrangement. The assembly process produces files that enable visualization and interrogation of the sequence and are human-readable. The resulting files are known as sequence alignment map (SAM) file or its binary version (BAM) and have a smaller size than FASTQ files.

Independently of the sample analyzed but according to the availability of the reference sequence or to the goal of the study, the assembly has two main approaches. Reference- based sequence assembly is used when the reference sequence from the same organism, or closely related species has been previously

obtained. In this case, the reference sequence serves as a guide for the reconstruction of reads. On the contrary, de novo sequence assembly does not involve using a reference sequence. Is a more complicated process and requires more computational resources (53,55). Short- single-read sequencing approaches, for example, make it impossible to assemble human genome sequences de novo because of its length and complexity. Therefore, sequencing reads are usually compared against a reference (50).

Reference-based assemblies are usually performed for analyzing the human genome or transcriptome to reduce computational resources and avoid reconstruction issues. Despite, changes in the reference sequence may require revalidation of the assembly, explaining for example, why the community is mainly still using the GRCh37 human version instead of the latest one (GRCh38) (56). Moreover, it should be considered that 70% of the human reference sequence corresponds to a single individual and it does not represent global human genomic variation. Dependence on a single assembly creates reference biases, reducing the accuracy of genetic analyses. Even so this problem has not been resolved yet, The Human Pangenome Reference Consortium is working to create a more complete human reference genome representing global genomic diversity (58). Another major limitation is the complication when mapping short reads within repetitive or poorly characterized regions. As mentioned before, paired-end reads can partially solve this issue assuming one of the reads of the pair maps in a unique region (Fig 4).

Once NGS data is generated, the challenge remains on comprehensively analyze and interpret the sequences, as well as in the large and powerful computing environments needed to process the data. Following read alignment to a reference genome or de novo assembly, data usually undergoes different quality control steps with bioinformatic programs. Quality control includes inspecting depth coverage of the sample, defined as the average number of reads

that align to, or “cover,” known reference bases (Sequencing Coverage for NGS Experiments, n.d.), base call quality scores, mapping quality, duplication rate and strand bias. After this, to convert sequences into meaningful biological results, diverse tools can be used depending on the goal of the project. As an example, in order to identify variants related to a genetic disease, the sequenced and the reference genome will be compared using variant calling algorithms capable of detecting nucleotide sequence differences. Later, these changes will be annotated and interpreted to understand their impact on the cell (McCombie et al., 2019). More information regarding analysis protocols applied in cancer genomics and transcriptomics is explained in the following section.



Figure 4. Comparison of single-end and paired-end reads in NGS alignment. PE sequencing provides additional information on fragment length and read orientation.

2.2.3 Integrating sequencing data in biomedical sciences.

Progression towards precision medicine.

NGS has been a changer in genetics. Before NGS, reading DNA was slow and expensive, limiting our understanding of the genome. Using next-generation sequencing, we are able to read DNA faster and at a lower cost. Therefore, we can study genes in more detail and discover new genetic elements, as well as finding changes that cause disease. NGS has opened up a whole new world of genetic knowledge and possibilities.

In 2021, a total of 3.278 unique animals have had their nuclear genome sequenced, assembled and publicly available in the GenBank database (60,61).

The availability of complete genome reference sequences, together with faster, cheaper and accurate NGS technologies to produce large amounts of sequencing data, and bioinformatic tools to analyze them, have opened new horizons within genetics research and lead to planning higher level projects. Next-generation sequencing is used to study genomes of humans, animals, plants, microbes and viruses. The number of applications is nearly limitless including searches for new genes and their functions, discovery of diversities among individuals and disease-related genes or locating common and rare variants that influence the risk of developing complex diseases, as well as variants acquired during lifetime in specific cell populations that can drive tumorigenesis. Specific techniques such as RNA-Seq provides direct cell- and tissue-specific gene expression features, quantification of transcripts, detection of splice variants and novel transcript isoforms and chimeric gene fusions. The evaluation of the transcriptome profiles is also valuable for understanding diseases (53,55,57,62). Integrating both DNA and RNA analysis provides further evidence of altered function of mutated genes, allowing for accurate definition of the basis of the disease.

Several human diseases are associated with genetic variants that can be inherited from carrier parents (germline variants) or acquired during lifetime (somatic variants). Despite all genetic diseases that can be studied using NGS, the approach used for each can differ. Inherited rare diseases, which are those affecting a low number of individuals, are usually the result of single-gene mutations directly affecting a protein sequence. Therefore, target sequencing or whole-exome sequencing are sufficient to identify the precise exonic mutations causing the disease. Since whole-exome sequencing accounts for approximately 2% of the entire human genome, many disease-causing variants can be discovered using this NGS approach. However, for interrogating non-coding regions of the genome or transcriptome, as well as for studying large variants that can even involve different chromosomes, whole-genome sequencing is needed

(53,55). WGS is the most comprehensive NGS approach, so is commonly applied to study genetics of complex diseases, which are caused by a combination of variants distributed in coding and non-coding regions, and environmental factors (63). This last approach is used more often in research than in the medical field. For example, for diagnosis and stratification of cancer patients, the NGS assay applied is typically targeted sequencing panels, which interrogates dozens or hundreds of interesting genes that are known to be related with the disease.

Not only should the NGS assay differ depending on the disease, but also other features, such as the coverage at which the sample is sequenced. To detect most germline hetero- or homozygous variants, 30x coverage, meaning around 30 reads aligned across each sequenced nucleotide, is enough. However, to identify rare somatic cancer variants, present in only a cell population and in low frequency, higher coverage is needed (64).

DNA and RNA sequencing, as well as other omics data, are now mainstream and contribute not only to biology but also to medicine for diagnosis, prognosis, follow up and treatment decision. In both fields, research and patient care, multiple traditional molecular assays may have to be performed for studying multiple mutations and a large amount of tissue is needed. Using NGS, hundreds and thousands of genes, target regions, or whole genome, can be interrogated in one single test from small biopsy samples. NGS experiments link experimental design with data analysis, and they can be combined with other classical methods to have a greater insight into biological disease. Biomedical research projects have changed the way they are designed. Therefore, instead of focusing on just one or few variants, genes or proteins and using the function-to-genetic approach, they aim to explore many regions at once to provide a wider genome representation of variants (or genes) to later associate with disease. Moreover, due to its capacity to massively sequence regions or genomes faster and cheaper, NGS is an

important tool for precision medicine and offers new opportunities that can be applied in patient care (56,57,65).

Precision medicine has revolutionized how we improve health and treat disease. Although nowadays, the “one-size-fits-all” classical approach is still how most medical treatments are designed, it is known that it is not effective in many cases. Treatments can be very successful for a group of patients, while for others not. On the contrary, precision medicine considers individual differences in people’s genes, microbiomes, environments, family history and lifestyle. (Fig 5) This information allows clinicians to make diagnostic and apply therapeutic strategies precisely for each individual patient (66). In order to apply precision medicine, biomedical research should be done to understand genetics and biology behind a disease.



Figure 5. One-size-fits-all and precision medicine approaches. Contrary to the classical approach, in precision medicine differences between individuals are considered to make a diagnostic and apply therapeutic strategies.

2.3 Cancer: a collection of complex diseases

Through cell division, human cells grow and multiply in the body to form new cells to replace those growing old or damaged. However, this orderly process can break down and abnormal or damaging cells grow and multiply uncontrolled. These cells may form tumors, lumps of tissue, which can be cancerous or not (benign) (67). Clinical differences between benign and malignant tumors were described by Gabriele Fallopius, who identified cancer cells to have irregular shape, multi-lobulation, adhesion to neighboring tissues and more blood vessels surrounding the lesion (68).

In the late 1800s, three fundamental theories described the cause of cancer, proposing that was a product of chronic irritation, hypothesizing that was the result of displaced embryonal tissue or suggesting cancer was caused by infectious or pathogenic agents. Bernardino Ramazzini observed that nuns suffered from high rates of breast cancer, which was attributed to their celibate life. Harting and Hesse documented in 1879 that miners in the Black Forest regions in Germany died due to lung cancer. Other non-occupational agents as tobacco were associated with, in this case, nasal cancers, as well as viral infections due to sexual promiscuity were also correlated with risk of cervical cancer. Variations in the type of cancer found in different areas of the world were also observed, and people who migrated to other countries developed types of cancer common in their adopted countries, rather than their homelands (69).

Alfred Armand Louis Marie Velpeau, after examining malignant and benign tumors under the microscope, wrote that cancer cells were merely a secondary product rather than the essential element in the disease, and that there must exist another intimate element which science would need to define the nature of cancer. He was anticipating the genetic bases of cancer. Following the view of Velpeau's, Theodor Boveri first proposed a role for somatic mutations, those

acquired during lifetime, in cancer development. Boveri suggested that loss of key cellular attributes due to these mutations were important driver events in the formation and progression of cancer, and that inheritance of germline variants could play a role in disease susceptibility. It had taken 50 years of work for Boveri to validate Velpeau's intuition, and another half century for the emergence of molecular biology and molecular genetics to confirm Boveri's theory on the nature of cancer (68,70). Viewing tumor as a cellular rather than an organ problem led to confirm that cancer is a genetic disease involving dynamic changes that lead to malfunctioning of the cellular properties.

Cancer is then defined as a collection of complex diseases. Is characterized by uncontrolled cellular growth and division of abnormal cells due to the accumulation of genetic and epigenetic changes and their subsequent natural Darwinian selection when conferring advantages for the tumor cell. It can be caused by genetic predisposition due to inherited mutations, also named germline, somatic variants acquired during lifetime due to environmental factors such as UV light or tobacco, genome instability, infections, chronic irritation, aging or the combination of various of the mentioned factors (67,71). Genetic changes usually tend to affect oncogenes, tumor suppressor genes and DNA repair genes. Whereas alterations in oncogenes cause the activation of them allowing the cell to grow and survive when they should not, variants in tumor suppressor genes cause the loss of their function and lead to malignancy. Inherited mutations inactivating one allele of a tumor suppressor increases the probability of developing a tumor. Efforts regarding the identification of potential therapeutic target genes have been mainly focused on oncogenes (70,72). As an example, the most common loss of proapoptotic regulator through genomic mutation involves the p53 tumor suppressor gene. The functional inactivation of this gene is seen in more than 50% of human cancers. Epigenetic modifications can also influence gene expression and contribute to cancer development. Moreover, interactions

within the tumor microenvironment which consists of surrounding normal cells, play an essential role in progression, survival of cancer cells, promote angiogenesis and modulate immune responses (73).

There are more than 100 types of cancer, usually named depending on the organ or tissue where tumor arises, and also described by the type of cells that formed them. As an example, carcinomas, the most common type of cancer are formed by epithelial cells, those that cover inside and outside surfaces of the body, whereas sarcomas are developed in bone and soft tissues such as muscle or fat, leukemias begin in the bone marrow and lymphomas in lymphocytes (T and B cells) (67). Each cancer type has its own characteristics, progression, and treatment responses, and even within the same type of cancer there can be significant diversity between patients and at molecular level. These differences, that have been well described in large-scale studies, are defined as intertumor heterogeneity (Fig 6) (64). Understanding this heterogeneity allows for precise treatment approaches tailored to each patient's specific tumor characteristics, increasing treatment efficacy and reducing adverse effects.

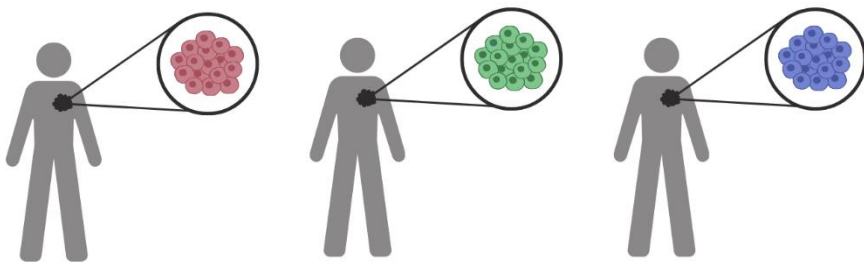


Figure 6. Intertumor heterogeneity, showing genetic and molecular differences across patients and tumors.

2.3.1 The hallmarks of cancer: decoding the complexity

Cancer genomes are altered at multiple sites due to tumorigenesis, a multistep process by which normal cells undergo a series of genetic and epigenetic changes leading to the formation of a tumor. Mutations or alterations disrupt the normal regulatory mechanisms of cells, enabling a set of distinctive traits or characteristics that are commonly observed in most malignant tumors, collectively named as the hallmarks of cancer.

Six essential alterations in cell physiology that dictate malignant growth were first described in 2000 by Hanahan and colleagues (73). These shared hallmarks of cancer (Fig 7) that are acquired during tumor development include:

- 1) Self-sufficiency in growth signals: cancer cells acquire molecular strategies to achieve autonomy stimulating their own growth signals continuously and reducing their dependence from normal tissue microenvironment. Many oncogenes in the cancer catalog act by mimicking normal growth signaling, leading to unregulated cell division and therefore, tumor formation.
- 2) Insensitivity to antigrowth signals: soluble growth and immobilized inhibitors present in the extracellular matrix or surfaces nearby cells operate as antiproliferative signals to maintain cellular quiescence and tissue homeostasis in normal tissues. However, cancer cells can evade these mechanisms, associated with the cell cycle clock, and bypass natural controls on cell division.
- 3) Evading apoptosis: the apoptotic program is present in all cell types throughout the body, including a series of steps that cause programmed death of old and dysfunctional or unnecessary cells. Cancer cells become resistant to apoptosis allowing them to survive and accumulate even under unfavorable conditions.

- 4) Limitless replicative potential: independent of the cell-to-cell signaling pathways that limit multiplication, cells have a finite replicative potential. Therefore, once they progressed through a certain number of duplications they stop growing. This process is called senescence. Cancer cells can maintain their telomeres, protective caps on the ends of chromosomes, and consequently replicate unlimited and prevent cellular senescence.
- 5) Sustained angiogenesis: via an “angiogenic switch”, tumors induce the formation of new blood vessels to ensure a dedicated blood, and consequently, oxygen and nutrients, supply.
- 6) Tissue invasion and metastasis: during tumor development, primary tumor masses spawn pioneer cells that move out, invading nearby tissues and spreading to distant sites in the body. Tumors can succeed in these sites and found new colonies, forming secondary tumors named metastases.

Later in 2011 and 2022, and due to the continuous study of tumor biology, four novel attributes (2 in 2011, and 2 in 2022) of cancer cells were proposed and added to the list of core hallmarks (Fig 7) (74,75). These new emerging hallmarks are:

- 1) Reprogramming energy metabolism: uncontrolled cell proliferation also involves adjustments on the metabolism, shifting the energy production to fuel cell growth and division. Cancer cells reprogram their glucose metabolism through what is called the Warburg effect, limiting their energy metabolism mostly to glycolysis facilitating the biosynthesis of macromolecules and organelles required for assembling new cells.
- 2) Evading immune destruction: although both the innate and adaptative cellular arms of the immune system are able to contribute to immune surveillance and tumor eradication, solid tumors managed to avoid detection and evade destruction by the immune system.

- 3) Unlocking phenotypic plasticity: during organization of cells into tissues, the end result of cellular differentiation is antiproliferative, being a barrier to continuing growing. Cancer cells unlock the restricted capability for phenotypic plasticity to escape from the terminal differentiation.
- 4) Senescent cells: senescence not only shuts down the cell division cycle, but also evokes changes in cell morphology and metabolism, involving the release of proteins such as chemokines, cytokines, and proteases. Therefore, senescent cancer cells contribute to proliferative signaling, avoiding apoptosis, inducing angiogenesis, stimulating invasion and metastasis, and suppressing tumor immunity. A transitory state of senescence is well documented under therapy resistance, representing a form of inactivity of proliferating cancer cells, but more operative in other tumor stages of development, progression, and metastasis.

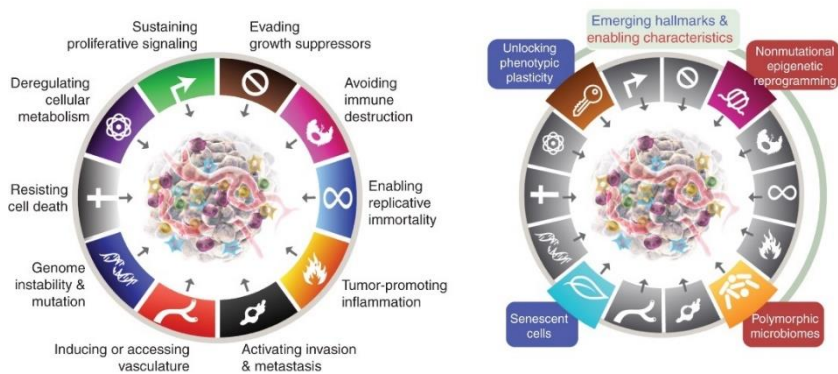


Figure 7. Hallmarks of cancer and enabling capabilities of cancer. Figure from Hannah et al. 2022 (75).

The acquisition of these hallmarks of cancer is made possible by four enabling characteristics. The first two enabling characteristics were described in 2000 and included genome instability and tumor-promoting inflammation. Cancer cells often increase the rates of mutation acquired in their genomes, including changes and the loss of function in a growing number of genes involved in sensing and repairing DNA damage, assuring correct chromosomal segregation in mitosis, and in general affecting to genomic “caretaker” systems. These genetic events can occur early in some tumor progression pathways and late in others. On the other hand, chronic inflammation can supply bioactive molecules to the tumor microenvironment, including growth factors, or enzymes that facilitate angiogenesis, invasion, and metastasis (73). The last two well described enabling characteristics consider epigenetic reprogramming, involving epigenetically regulation of gene expression that facilitates the acquisition of hallmark capabilities, and the presence of polymorphic microbiomes. Increasing evidence has shown that variability in the microbiomes between individuals can have an impact on cancer phenotypes. Some bacterial species stimulate proliferative signaling and modulate growth suppression by modifying tumor suppressor activity (75).

Understanding these hallmarks provides insights into the complex nature of cancer and guides research efforts in the development of targeted therapies and diagnostic approaches. In fact, targeted therapeutics can be categorized according to the effects on one or more hallmarks. Most of these therapies have been delivered directly to molecular targets involved in enabling particular capabilities. However, not because of inhibiting one key pathway the tumor may completely shut off a hallmark capability, and cells eventually adapt to the selective pressure resulting in relapse (74).

2.3.2 Somatic variation in the human genome

Everyone is born with a collection of genetic variants that define our genotype. This determines many aspects of our biology and our life, and together with environmental factors predispose us to different kinds of disease as well as prevent for others. These genetic variants are known as germline (Fig 8), and they are normally studied in the context of rare and complex diseases.

Somatic variants (Fig 8) are genetic alterations that occur after conception and therefore are acquired in somatic cells but not in germ cells (76). Unlike germline variants, which are inherited from parents and present in every cell of an individual, somatic variants are only present in certain cells or tissues and are not passed on to offspring. These variants arise as spontaneous stochastic events during lifetime, because of specific factors, including exposure to carcinogens (UV light, pollution, chemical agents), chronic inflammation, lifestyle choices such as smoking, diet and physical activity, DNA replication errors or the impact of external stressors. Moreover, the risk of acquiring somatic mutations increases with diverse inherited variants. These acquired variants can lead to dysregulation of multiple essential cellular processes, including cell cycle control, DNA repair or apoptosis, and therefore are one of the main causes of cancer in combination with other genetic and epigenetic changes (72,77,78).

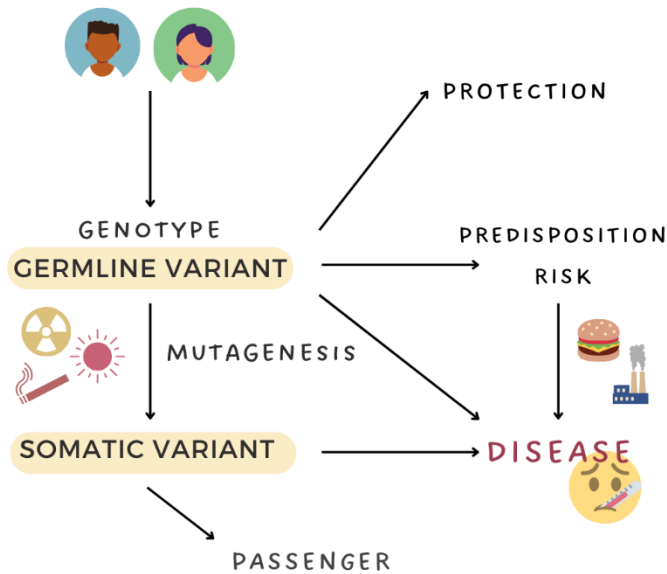


Figure 8. Variation in the human genome. Germline variants are those inherited from the parents, while somatic variants are acquired during lifetime. Whereas some germline variants can be protective, others can directly cause rare diseases, or increase the risk of developing a complex disease in combination with environmental factors. The majority of somatic variants acquired due to mutagenesis do not result into a disease (passenger variants), whereas others are the cause of cancer.

In cancer, characteristic patterns of somatic mutations found in the genomes of the tumor are referred to as mutational signatures (54,79). These patterns result from particular mutational processes that can be caused by factors such as exposure to mutagens, defective DNA repair mechanisms or other cellular processes. As an example, signature 1 represents a clock-like mutational process (aging) and it is widely observed across all types of cancers. Each mutational signature is characterized by a distinct combination of mutation types and the specific nucleotide context in which these mutations occur. Their study provides insights into the biological processes driving cancer development and progression and helps to identify the causative processes and environmental exposures contributing to tumor formation.

Understanding the impact on cellular function of somatic variants itself, and of mutational signatures, enables cancer classification, patient prognosis, identification of therapeutic targets, prediction of response to certain treatments and consequently, advancing precision medicine approaches.

2.3.2.1 Types of somatic variants

Genomic alterations are classified according to the type of DNA change in single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations (CNAs) and structural variants (SVs) (54). A brief description of them is provided below.

- Single nucleotide variants (Fig 9) occur at a single nucleotide in the DNA sequence and involve its substitution with another at a specific position in the genome. SNVs are the most common types of genetic mutations, the smallest and the most easily detectable. Depending on their location in the genome, they can lead to alterations in the genetic code, impacting the function of genes and proteins, or may be benign with no detectable effect. Generally, can be classified as missense variants if the altered codon is translated into a different amino acid, stop gain or loss when they produce a new stop codon within the sequence or delete it, or synonymous in case the substitution does not implies a change in the translated amino acid.

- Small insertions and deletions (Fig 9) result from the insertion or deletion of one or more nucleotides in the DNA sequence, usually up to 50 bp. They can have significant consequences when appear genes, causing frameshift mutations, altering the reading frames, disrupting splice sites and introducing premature stop codons resulting into to the production of truncated or non-functional proteins.

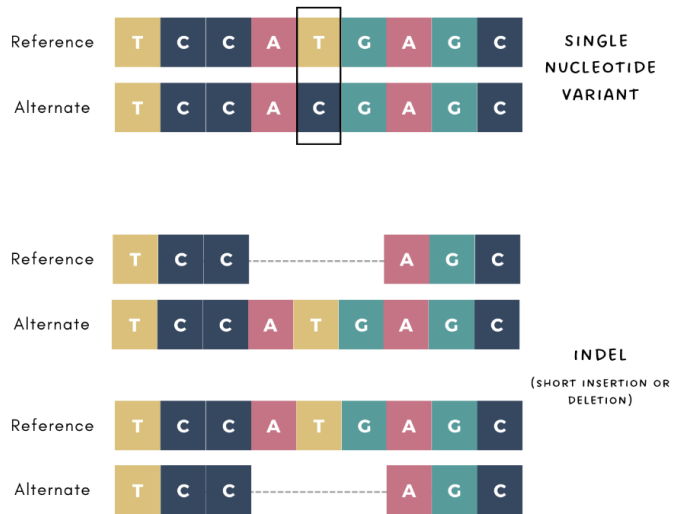


Figure 9 Types of somatic variants. Up, representation of a single nucleotide variant. Down, visualization of a short insertion (below) and deletion (above). Small insertions and deletions are named as Indels.

- Copy number alterations refer to change (duplications or deletions) in the number of copies of a particular segment of DNA within the genome. CNVs (Copy Number Variants) can range in size from a few hundred base pairs to large segments of DNA containing multiple genes, as well as entire chromosomes. They can have significant effects on gene dosage and expression levels. Thus, they can alter gene function.
- Structural variants (Fig 10) are the most complex type of genetic alterations and encompass DNA breaks and sequence reassembling elsewhere in the genome. Unlike SNVs and indels, structural variants can encompass much larger regions. Structural variants include large deletions and insertions, that can involve new DNA from exogenous sources like viruses, duplications, inversions and translocations where more than one chromosome is involved. Different types of translocations can be also defined. Whereas there is no loss of genetic material in balanced translocation events, unbalanced translocations cause the loss of DNA. Moreover, two-way exchanges between non-homologous chromosomes are known as reciprocal translocations, whether a one-way transfer of a segment into a non-homologous chromosome is defined as nonreciprocal translocation. Generally, structural variants can have significant consequences on gene regulation, gene fusion and genomic stability, which is a hallmark of cancer.

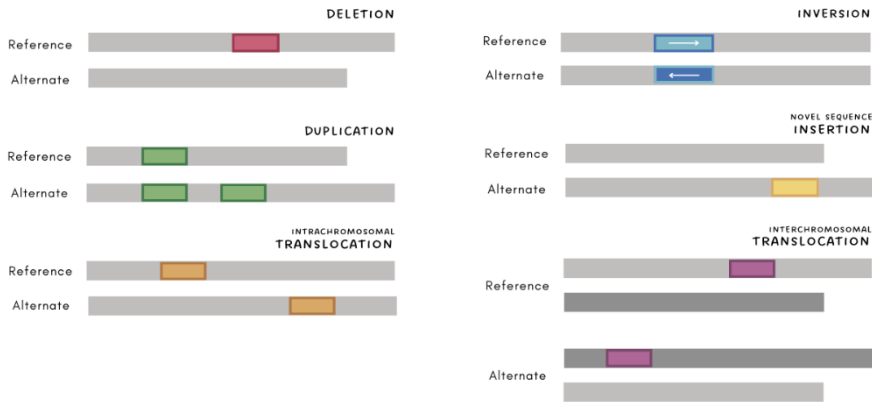


Figure 10. Types of somatic structural variation.

2.3.2.2 Variant calling analysis to describe the somatic variation landscape of tumors

Three different research approaches are required to assess the genetic profile of a disease depending on its type. To study the relationship between complex diseases and inherited germline variants, genome-wide association studies (GWAS) are used. These variants are used to be enriched in a subgroup of individuals that increase their risk of developing a phenotype or complex disease such as asthma, cardiovascular diseases or type 2 diabetes. A large number of patients showing the selected phenotype as well as a subset of control patients not presenting the phenotype are needed to compare their genetic variation landscape, mainly within their exons. On the other hand, rare diseases, health conditions with a very low prevalence in the population, can be studied using a more cost-effective whole-exome sequencing or even target sequencing to only identify pathogenic variants in coding regions of the genome, which are expected to occur at high penetrance. The number of patients evaluated is lower and mutations are not recurrently identified among all the cohort nor the entire genome but can occur in specific genes. Finally, the analysis of somatic variants,

which is mainly related to cancer, is done through the comparison of an individual genome, evaluating those genomic positions where an alternate allele is supported by cancer cells and not present in normal and healthy cells. Thus, in order to identify recurrent variants across patients, a large number of sequenced normal and their matched tumor samples from diverse patients is needed.

In the era of NGS, the general strategy to study cancer genomes and somatic mutations starts, as mentioned, by comparing normal and matched tumor DNA previously extracted and sequenced. Both sequences are aligned separately, and the BAM files obtained are the inputs for the variant calling, which is the main step for DNA alteration discovery.

The massive amount of data generated by NGS required the development of algorithms based on statistical methods and computationally efficient. Variant caller algorithms are bioinformatic tools used to detect and identify genetic variants from high-throughput sequencing data. During the past years, many tools have been created for both germline and somatic variant detection, with the main difference being the usual need of a normal and matched tumor sequence to identify acquired genetic variants in an individual. These methods can also be grouped based on the variant type they are able to detect: SNVs, indels, CNVs and SVs. Whereas some tools are dedicated to one single class, others can detect different kinds of variants. However, the difficulty in finding each type is different, being SNVs the easiest and SVs the most complex ones. Mismatches between aligned reads and the reference genome allow to detect point mutations and short insertions and deletions. Structural variants are called based on split reads, part of a read maps to a different region or it appears as unmapped, and paired-end read discrepancies regarding orientation, mapping chromosome and/or insert size.

Many factors can complicate variant calling steps, starting with technical determinants such as sequencing errors and alignment artifacts and followed by low frequency variants due to tumor heterogeneity or low purity of the sample. Each variant caller applies a specific criterion to call variants with a determined confidence, based on read depth, that is the number of unique reads in a reference nucleotide, base quality or variant frequency. Variant callers have their own strengths and limitations preventing the detection of false positive and false negative events (53). For this reason, researchers typically combine the results of diverse algorithms, applying a multi-variant calling approach to increase sensitivity and reduce the rate of false negatives (54,79) and occasionally reviewing the results through manual inspection. This approach has been implemented in many institutional pipelines and main large-scale cancer genomics projects.

Variant caller algorithms play a fundamental role in genomic research and personalized medicine, enabling the identification of genetic variants associated with disease. Nevertheless, once variants are detected, their potential to activate oncogenesis, association with drug response or the disease evolution and outcome must be evaluated through annotation and functional analysis. Variant annotation is the process of assigning information to DNA variants and assessing their possible pathogenicity. This step is a crucial point and is a challenging bridge between machine and human-readable format (54). Although this interpretation can be used by researchers and clinicians to tune precision therapies, and despite numerous efforts to provide guidelines and best practices, its application to the clinics is still complex and problematic. Diverse algorithms including Variant Effect Predictor (VEP) (80), ANNOVAR (81) or PAVE (<https://github.com/hartwigmedical/hmftools/tree/master/pave>), have been also created to annotate genomic variants using information from diverse public databases and estimating the consequence and impact of each variant.

2.3.2.3 Public databases and catalogs of genomic variants

As the field of cancer genomics progresses, the intricate process of sequencing, alignment, and variant calling has illuminated the need for robust information technology infrastructures and sophisticated computational tools (78). These components are vital for transforming raw data into meaningful insights within the context of characterizing tumors in large-scale cohorts. Moreover, the imperative to biologically understand the results and disseminate findings has led to the development of public databases. These repositories facilitate the search and sharing of results and drive the advancement of cancer genomic research, underscoring their pivotal role in making such research not only possible but also deeply impactful.

A wide range of databases collecting genomic variation have been developed, including all kinds of variant annotations, and cancer specific catalogs. Comprehensive databases of human genetic variation such as dbSNP (Single Nucleotide Polymorphism Database) (82) or gnomAD (Genome Aggregation Consortium) (83) can be used to annotate known germline variants and their population frequencies. This information also allows to filter false positive events identified as somatic. For functional annotation analysis, resources include information found in literature and curated annotations. As an example, ClinVar (84) is a widely used and freely available archive from the National Center for Biotechnology Information (NCBI) that provides information for interpretation and clinical significance genetic variants. Diverse tools for predicting the potential impact of a variant have also been created, including PolyPhen-2 (85), for amino acid substitutions or SIFT (86).

Focused on somatic mutations and cancer, one of the most well-known databases is The Catalog of Somatic Mutations in Cancer (COSMIC), which includes almost 6 millions of coding mutations across 1.4 million cancer samples.

It also includes non-coding mutations, copy-number alterations, gene-fusions and mutational signatures (87). Moreover, a catalog of driver genes (the Cancer Gene Census) is also available to search or download. Although it is not a dedicated annotation database, The Cancer Genome Atlas (Weinstein et al., 2013a) also provides comprehensive genomic data for various cancer types, allowing to analyze and annotate genetic variation. Other databases and web portals such as IntoGen (89) are available to evaluate cancer driver genes previously identified in large-scale cohorts.

2.3.3 Driver and passenger mutations in cancer

The number of genetic mutations present in the DNA of a tumor sample is quantified in cancer genomics and is known as tumor mutational burden (TMB). It includes the total count of somatic mutations; those acquired during lifetime and therefore not present in all the cells of the body. TMB is typically expressed as the number of mutations per mega base (Mb) of DNA and it is often obtained using NGS techniques that allow the analysis of the tumor genomic profile. High mutational load, representing a large accumulation of mutations, is usually associated with environmental DNA damage and in clinical practice it can be related with a better prognosis and longer survival (64).

The accumulation of specific combinations of genetic alterations or the presence of mutations in a defined set of target cells results in higher propensity for malignant progression. Therefore, not all somatic mutations promote cancer development (77). Two categories of somatic mutations were defined.

Driver mutations directly or indirectly play significant roles in oncogenesis. They occur in genes (called cancer driver genes) that regulate key cellular processes, primarily enabling the previously mentioned hallmark capabilities (64,71). Driver mutations likely occur at different stages of tumor evolution. Diverse studies have revealed that normal cells frequently harbor one or more

cancer driver mutations, and that the landscape of drivers and their expansion greatly varies between tissues (<5% in colon cells carry drivers compared to >50% in endometrium) (90). Most genes act as drivers in one or two tumor types, and only around ten genes can drive more than 20 malignancies through mutations. Moreover, mutations can drive tumorigenesis only under specific selective constraints. On the other hand, passenger mutations are genetic alterations resulting from genomic instability but do not have an impact on tumor growth. The majority of somatic mutations found in cancer likely represent passenger variants and only a minority are drivers. Therefore, passenger variants can provide valuable information about the evolution of the tumor, can aid in understanding the complex genomic landscape of cancer, and can be used in research to identify specific mutational patterns. Because of this, they may also have implications in precision medicine and target therapies.

2.3.3.1 Identification of cancer driver genes through bioinformatic approaches

The identification of cancer driver genes is crucial in cancer genomics to advance our understanding of the biology behind tumor formation and progression, and to guide precision treatment and diagnosis approaches and develop effective therapies that target specific genes driving the tumor, to ultimately improve the patient's quality of life. The search for gene abnormalities that can lead to cancer development is one of the pillars of cancer research since the discovery of a point mutation in *HRAS* gene that causes the activation and transforming capacities in human bladder carcinoma (91). The improvement of DNA sequencing technologies and the advance in the annotation of the human genome enables us to reveal the landscape of somatic mutations in tumors. While only a few tens of cancer driver genes were characterized through biochemical

and molecular assays in the span of two or three decades, hundreds of cancer genes have been identified using cancer genomics in less than two decades.

Since tumorigenesis follows a Darwinian evolution, spontaneous somatic mutations acquired in diverse cells are positively selected when conferring selective advantages for them. As a result, the patterns of mutations in specific genes, those driving tumorigenesis, deviate from their expectation under neutral mutagenesis. Following this assumption one common strategy to identify cancer driver genes involves the analysis of somatic mutations across large-scale cancer genomic datasets together with statistical methods to seek for genes mutated at abnormal high frequencies across the cohort.

Driver discovery methods focus on one or more features of the mutational pattern of genes. Bioinformatic tools can be used to detect unexpected clustering of mutations in specific protein regions, to determine a bias towards the accumulation of variants with high functional impact or deviation in the frequency of trinucleotide changes (Fig 11). The obtained results allow to prioritize those genes that are more likely to have a role in cancer to explore them deeply (71,92). Mutational features may also reveal different tumorigenic mechanisms of the same driver gene across tumor types.

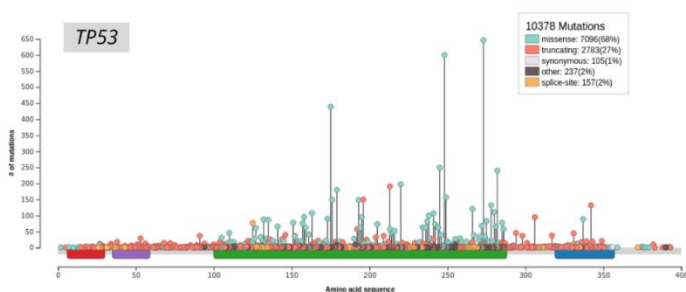


Figure 11. Observed mutations in different tumor types and across TP53 cancer driver gene. Clustered and recurrent mutations have been identified within the gene by multiple algorithms. Image from www.intogen.org.

In 2020, the compendium of driver genes obtained through the analysis of cancer exosomes comprised between 500 and 600 mutational drivers. Although genes mutated at frequencies higher than 10% have already been discovered, it is predicted that the number of identified drivers will increase. New drivers could be derived from genes mutated at lower frequencies or in populations that have been biased against in tumor genome sequencing projects, as well as of conditions not profiled and new clinical samples including metastatic or relapse tumors. Integrative approaches incorporating multi-omics data, integrating large-scale cancer genomic datasets and functional genomics information will allow to identify novel cancer driver genes.

2.3.4 Intratumor heterogeneity and clonal dynamics

Genomic differences among cancer patients diagnosed with the same tumor type have been demonstrated and characterized and are known as intertumor heterogeneity (Fig 12) (93).

Furthermore, it is well known that tumors are formed by many cell populations (94–97), and each of these can accumulate different somatic genetic variants including passenger and driver mutations. This phenomenon is called intratumor heterogeneity and refers to the presence of genetic, phenotypic, morphological and functional diversity within the cells of a single tumor mass.

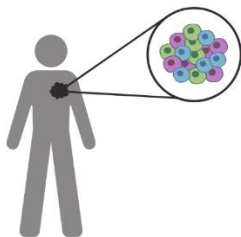


Figure 12. Intratumor heterogeneity. Tumors formed by diverse cell populations including genetic and molecular differences.

Clonal dynamics, also known as tumor evolution (Fig 13), can be depicted as a succession of clonal expansion rounds, where every round is driven by the acquisition of additional mutational events. Mutations are acquired stochastically because of proliferation and increased genomic instability. Then, as a Darwinian evolution process, these mutations are selected, and cell populations named clones are adapted resulting in ITH (90,98). The study of subclonality can reveal a tumor's life history and the temporal order of the acquired somatic events. In the early phase of cancer evolution, founder mutations are acquired. This common ancestor or trunk of the evolutionary tree branches into subclones due to genetic instability and alterations in the tumor microenvironment, accumulating new mutations and leading the heterogeneity within the tumor tissue. Usually, mutations in driver genes are identified as founder mutations and consequently are present in all cells being clonal (64,97). Even though the Darwinian process can explain the history of tumors to some extent, the full spectrum of cancer evolutionary trajectories is not sufficiently encompassed. Non-Darwinian mechanisms have been also described and considered a form of evolution through one-hit catastrophic events that bring multiple genetic alterations at the same time. These macroevolutionary events can drive tumor initiation and progression (99).

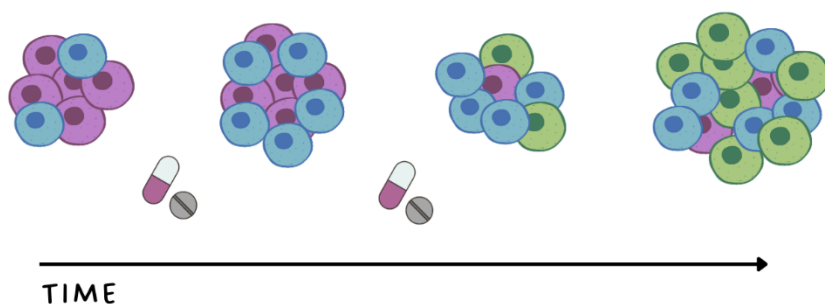


Figure 13. Tumor evolution during time. Cell populations at the beginning (left) grow and/or disappear during time and depending on the received treatment. Relapse is shown at the end (green cells).

Intratumor heterogeneity has significant challenges in clinical management and is likely the major cause of therapeutic resistance and tumor relapse. Clonal diversity can provide a more diverse material on which selection can work, allowing the tumor to therapeutic adaptation instead of extinction. The co-existence of genetically distinct clones, that may interact between them also modulates progression and therapeutic responses (98). Although in theory, cancer therapy reduces genetic variation in cells, generally it only removes sensible clones, eliminating competition for growth and resulting in the expansion of subclones (98,100). The constantly changing environment of tumors underlies their ever-changing dynamics, where clones that were dominant reach a bottleneck and are depleted, whereas other minor subpopulations achieve a favorable position later and become dominant. ITH is also associated with the aggressiveness of the disease, as it has been demonstrated for example in prostate cancer (96). Based on several studies, we now expect that most tumors present a certain level of ITH.

An example of the strong relationship between subclonality and therapeutic resistance can be found in chronic myelogenous leukemia (CML). Patients diagnosed with CML show notable response to a treatment known as imatinib mesylate, but a fraction of these patients relapse. Analysis of the tumor genome of these patients showed the presence of resistant cell subpopulations, which are selected by pressure once the treatment is given and leads to the expansion of therapy insensitive cells causing relapse (98). Moreover, it has also been seen that clonal evolution is more frequent in tumors receiving chemoimmunotherapy than treatment-naïve tumors, where the clonal architecture can be in equilibrium (100).

Understanding the complexity of intratumor heterogeneity is critical for effective therapeutic strategies and precision medicine approaches that directly target the diverse subclones within a tumor to achieve better patient outcomes. Especially in the early stages of a tumor, identifying resistant clones could avoid tumor relapse and improve cure rate.

2.3.4.1 High-throughput sequencing analysis to decipher cell populations

Intratumor heterogeneity is then another level of complexity when studying cancer. Analyzing genome sequencing data of bulk tumor samples and based on the cancer cell fraction (CCF) of a set of somatic mutations, the subclonal structure of tumors can be identified. Mutations with similar CCF will probably represent the same cell population. Thus, clustering mutations based on their CCF yields the subclonal architecture of a tumor sample. While mutations present in all cells will be defined as clonal and are supposed to be from the initiating tumor cell, mutations with a CCF lower than 1 and therefore present only in a subset of cells will be named subclonal and acquired during tumor progression. Cancer cell fraction could be estimated by adjusting variant allele frequencies (VAF) for local copy number variation and sample purity (101).

The variant allele frequency of a set of somatic mutations can be directly estimated from NGS read counts. It is the result of dividing the number of reads supporting the variant allele by the total number of reads covering the genetic position or region. The value can be multiplied by 100 to get the VAF as a percentage. Since somatic mutations are mainly heterozygous, they are present only in one allele and consequently should be identified in half of the total number of reads covering their location. Therefore, if a somatic heterozygous mutation is clonal meaning it is present in all cells, its VAF will be around 0,5 or 50%. Lower

VAF values suggest the mutation is subclonal and a minor population of cells is carrying the genetic variant.

Studying ITH implies new methodological challenges, especially if subclonality is studied with standard sequencing depths. Low frequency variants, i.e., subclonal mutations, are difficult to detect with high confidence when using around 30x coverage samples. Moreover, variant allele frequencies are normally calculated only for single nucleotide variants, but small insertions and deletions and large structural variants could also be used. However, calculating the frequency for these large variants is not as easy and becomes a challenge because of read count is not straight forward with SVs.

Sequencing of a tumor sample only provides a static snapshot of its genetic landscape. The subclonality analysis of multiple tumor samples from the same cancer patient obtained from physically separate regions or different time points of the tumor development, allows not only a better and precise reconstruction of the cell populations but also their spatial distribution or their evolution during time (2,102,103). Comparing the VAF of a set of mutations representing a subclone and tracing them among longitudinal samples, researchers can evaluate how these cell populations change, expand or disappear from the tumor mass. The study of clonal dynamics together with clinical data can decipher, for example, whether a specific therapy results in relapse because of a specific resistant subclone.

Diverse bioinformatic tools (104,105) have been designed to cluster mutant allele frequencies and reconstruct tumor evolution using NGS data of one or more samples from a patient.

2.3.5 Large-scale initiatives promoting cancer research

As a response to the heterogeneous nature of cancer, advancements in technology and the collective effort to uncover the genetic landscape of cancer, large-scale studies have emerged during the last years.

The development of high-throughput sequencing technologies, such as NGS, becomes an increasing number of generated data ready to explore and analyze. Alongside sequencing technologies, computational tools and algorithms to analyze and interpret the data rapidly evolve, allowing cancer community to process efficiently vast amounts of data.

Moreover, due to the heterogeneity of cancer patients, and in order to later translate research into clinics applying precision medicine and focusing on the genetic makeup of each individual patient, large numbers of samples are needed. Therefore, the analysis of this data provides greater statistical power to detect rare genetic variation, significant associations and recurrent variants linked to cancer risk, prognosis and treatment response. Large-scale studies facilitate the identification of biomarkers and therapeutic targets specific to certain cancer subtypes that might not be evident in smaller studies.

Collaborative efforts among researchers, institutions, and countries became essential to tackle the complex nature of cancer genetics. To promote cancer research, diverse initiatives led by big consortia have organized international and national projects collecting and sharing omics data. In this framework, the most important and well-known initiatives have been the International Cancer Genome Consortium (106) and The Cancer Genome Atlas (107), both of which aim to coordinate cancer research projects including tens of cancer types and being collaborative.

Whereas the ICGC is a global initiative involving many countries (Fig 14), each leading the project and analysis of one cancer type, TCGA was launched by the

National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) and was based and coordinated within the United States. Spain could contribute in the ICGC providing Chronic Lymphocytic Leukemia samples and coordinating the project. Both initiatives aim to comprehensively characterize genomic variation in multiple cancer types including sequencing samples from around 20 and 60 different primary sites respectively and facilitating cross-disciplinary research collaborations. At this time, sequencing data can be downloaded and used for research after an approved application, and results including genomic variation can also be explored through their websites and data portals.

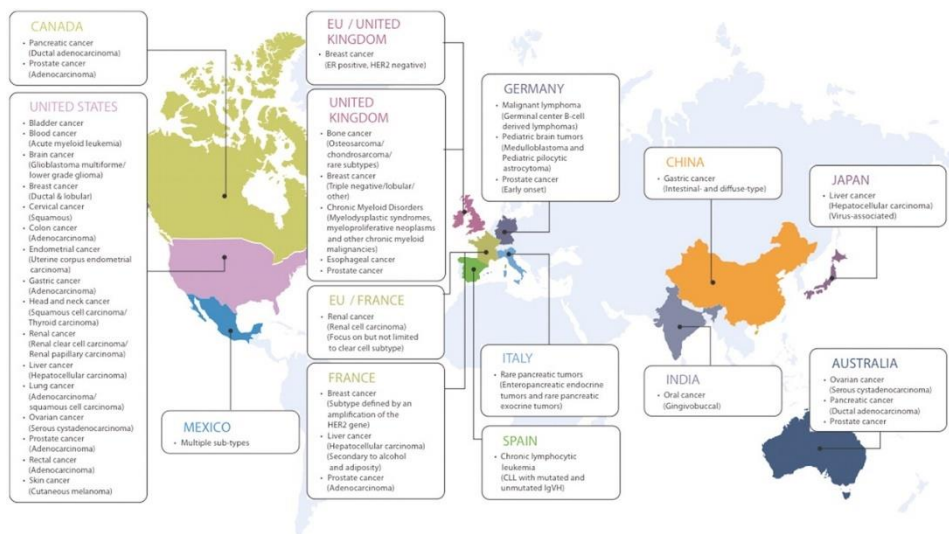


Figure 14. ICGC cancer projects and corresponding countries. Image from ZHANG P. et al, (2011).

As an evolution of these large-scale projects mainly generating sequencing data from thousands of tumors, the ICGC launched a new worldwide initiative named the Pan-Cancer Analysis of Whole Genomes (PCAWG). In this new phase, researchers intend to jointly analyze more than 2.600 normal-tumor whole genome pairs across 38 cancer types. Data was harmonized, annotated and homogeneously analyzed to later compare results among patients and cancer

types. Following the main goals of cancer genomics, PCAWG aims to deeper understand the molecular mechanisms behind tumor formation and evolution, including genetic patterns, driver mutations and key pathways that span various cancers. This study provides valuable insights on the identification of therapeutic targets, contributing to prevention, diagnosis and treatment through precision medicine and more effective cancer therapies. The PCAWG is the most comprehensive analysis of cancer whole genomes up to date, and required an infrastructure capable of performing large-scale analysis, enabling the storage of high amounts of data and their study using computational and data access tools.

Recently, to strengthen cancer research and its translation into the clinics, a new effort from the ICGC has been defined. ICGC-ARGO (Acceleration Research in Genomic Oncology) aims to coordinate the integration of homogenic genomic analysis and phenotypic data on 200.000 cancer patients. This dataset will be used to decipher key clinical and biological questions.

2.3.6 Challenges in cancer research

Many years have passed since researchers started applying NGS data to cancer research. However, in the field of omics data, the utilization for next-generation sequencing technology, the implementation of large-scale studies, and the translation of this research into the clinics, many challenges should still be faced.

Storage, analysis and interpretation of large datasets can be managed thanks to cloud-based solutions and local high-performance computers (HPC) clusters as well as new bioinformatic tools. However, at a global level, data sharing and standardization of the data is usually an obstacle in terms of methodological and legal aspects. Combining data from different projects and platforms for meta-analysis requires careful consideration of data harmonization, normalization and correction of batch effects to ensure valid comparisons. Moreover, the huge

number of tools, and the variety of for example variant caller algorithms that have discrepancies among the results, complicate the integration of the analysis. Other problematic procedures include the demanding characteristics of tumor samples, such as low purity, formalin-fixed paraffin-embedded archival material that could reduce sample quality, lack of matched-normal data. On the other hand, with the increasing availability of genomic data, keeping patient privacy and addressing ethical considerations related to data sharing and informed consent become a must. Although these ethical concerns are essential for research, they can also serve as hindrance since addressing them is neither immediate nor rapid, and their resolution can vary across countries.

Lastly, the results obtained should be easily translated into clinical applications. Not only are sophisticated algorithms needed for accurate interpretation but also easy-to-use in clinical environments.

While these challenges exist, the use of NGS and large-scale studies in cancer research holds great promise uncovering the complexity of the human genome and advancing precision medicine.

2.4 Processed pseudogenes: a by-product of L1 retrotransposition

The human genome is comprised of repetitive sequences, some of which are thought to originate from viruses. These sequences have the capability to transpose within the genome, generating multiple copies. Their study has been crucial in understanding the evolutionary history of human genes. Mobile elements have played a role in shaping the genome by promoting genomic diversity and providing insights into ancient genetic events. Their transposition within the genome could result in the formation of new functional genes or the inhibition of coding sequences. Investigating their impact on the genetic

landscape has shed light on the mechanisms underlying genetic innovation and adaptation, offering valuable clues about human origins and evolutionary development (108,109).

Mobile repetitive DNA, such as long interspersed elements (LINE) or Alu repeats, form a considerable proportion of the human genome. In particular, LINE-1 (L1) (Fig 15) composes about 17% of the entire human DNA content and 20% of the mouse genome. Although most repeat elements in the human genome are inactive because of truncations, point mutations and rearrangements, it is estimated that between 50 to 120 L1 are currently active being the most functional autonomous retrotransposons in mammalian genomes. When transcribed and translated, functional LINEs encode two proteins that coordinate reverse transcription of their RNA template and integrate them back into the genome (110–112). This process, involving the insertion of a DNA sequence mediated by an RNA, is known as retrotransposition or “copy and paste”. In humans, this is carried out through the mentioned proteins encoded by LINE elements (LINE-1 ORF2 and ORF1) which function as reverse-transcriptase and endonuclease (113). This machinery allows LINE elements to propagate in the genome as parasitic units, usually at a distant site from the original element, shaping the human genome over evolutionary time. It is estimated that about 79% of human genes contain at least a segment of an L1 element within its transcription unit (114).

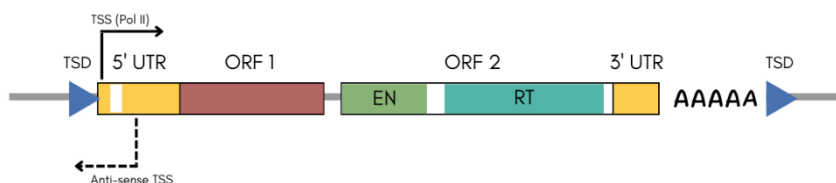


Figure 15. A human L1 element is 6Kb in length and encodes two ORF flanked by 5' and 3' UTRs.

Pseudogenes are complete or partial copies of genes, usually unable to code for functional polypeptides (115). Following the theory of neutral evolution by Kimura in 1968, over time, pseudogenes accumulate random mutations that can often cause disruption of the original reading frame. Therefore, these elements seem to be unconstrained by selection. It is known that mammalian genomes contain thousands of them (113) being the average density detected of 6.5 per mega base for the whole human genome. In fact, there is a strong correlation between the number of pseudogenes and the size of the chromosomes (116). Most frequent pseudogenes come from multigene families with large copy numbers. In general, housekeeping genes expressed in a wide range of tissue types are more likely to generate retrotransposed copies.

Depending on the mechanisms they have been formed, pseudogenes are classified as non-processed or processed. Those of the first category are the result of segmental duplication of genes and subsequent loss of function by mutations. A small fraction of duplicated genes will remain functional, being a source for the formation of new gene functions and expression profiles and considered one of the main drivers of evolution and a source for functional variability. The second category, processed pseudogenes (PP), are formed through the retrotransposition of mature mRNAs using L1 machinery (Fig 16). LINE-1, a still active retrotransposon in humans, is able not only to mobilize its own transcripts (*cis* preference), but also other repetitive elements such as Alu, SINE-VNTR-Alu, and nonrepetitive sequences including mRNA from other genes (*in trans*). Therefore, processed pseudogenes are a by-product of LINE-mediated retrotransposition. PPs are found to be complementary DNA copies of mRNA transcripts randomly integrated into the genome (110,112,117). Considering the mechanism processed pseudogenes are formed, most of these sequences share the following characteristics. Since PP are the result of reverse transcription of an mRNA, they completely lack intron sequences and upstream promoters found in

their functional paralogous gene, and many of them have a poly-A sequence after the 3' end. Usually, because they are mobilized and inserted into the genome using the LINE1 machinery, they are flanked by repeat elements of 7-17bp that were also present at the source region (110,113,115).

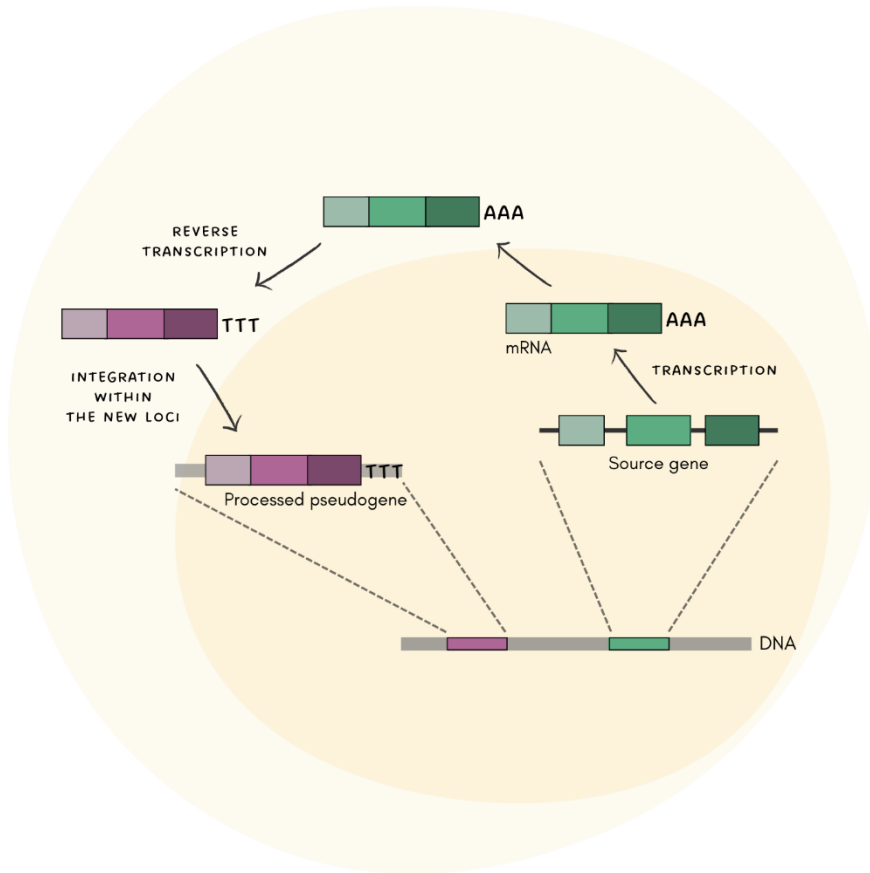


Figure 16. Representation of processed pseudogene formation. A fragment of DNA (green) is transcribed, and the resulting mRNA is retrotranscribed and inserted randomly in the genome (pink).

The functional equivalent gene of many pseudogenes are likely expressed in the germline cell. Hence, mRNA transcripts of functional genes are easily accessible for being potentially retrotranscribed during the next replication cycle and, subsequently integrated into the genome through repair and ligation (115). In a genome-wide study in 2003, among 20,000 pseudogenes identified in the human genome, 28% were due to segmental duplication whereas 72% arose through retrotransposition (116).

It is not clear if all human processed pseudogenes were formed recently in evolutionary time or many years ago. Even more, if these sequences have been highly mutated, they will not be detected as processed pseudogenes but completely different nucleotide sequences. In a review in 1985, the author claimed that all known pseudogenes arose after mammalian radiation, approximately 100 million years ago. In 2003, a comparative analysis between human PP and their orthologous region in the mouse genome was performed. Based on that, they could observe that pseudogenes align better to a different region in human than anywhere in the mouse genome. Therefore, the analyzed PPs were formed after the human-mouse split over 90 MY ago. The orthology criterion used in this study relies on the fact that retrotransposed mRNAs are randomly integrated, likely, far from their source gene. Actually, chromosomal localization studies revealed that PP and their functional genes are not syntenic, meaning they are not on the same chromosome. Additionally, it is known that their distribution does not correlate with the distribution of gene-rich regions within chromosomes. This statement argues against the idea that relaxed chromatin regions are more exposed to the integration of retrotransposed elements. It has been observed that PPs are more abundant near telomeres (115,116).

Processed pseudogenes are considered “dead on arrival” (112). Most PP acquire deleterious mutations to avoid them encoding functional polypeptides. Probably they are inactivated as soon as they are inserted due to missing promoters, frameshifts and truncation, and they cannot be transcribed by RNA polymerase II (115,116). Although genetic variants within them usually preclude their translation into a functional peptide equivalent to the active source gene, pseudogenes can affect genome function in diverse ways and influence evolution. First, their mobilization to another location can place the retrocopy in a novel regulatory context allowing the pseudogene to be transcribed and being an important source of material for new gene formation on evolution. In 2005, Harrison et al., identified about 4-6% of the known PPs expressed in the human genome (118). One remarkable example of new gene formation is the insertion of cyclophilin A (*PPIA*) into *TRIM5* in the owl monkey genome. This gene fusion confers resistance to HIV-1 infection (113). Transcriptional consequences can include the expression of UTRs or introns of target genes, as well as the production of antisense transcripts. When they are inserted within a gene, they could cause its disruption resulting on an aberrant and nonfunctional transcript, or not allowing it to be expressed. Finally, PP can also change the stability of the source transcript and compete with it for micro-RNA binding because of sequence similarity(110).

Germline L1s mobilization and processed pseudogenes formation have contributed considerably to the evolution of genes and genomes (111). If the event occurs in germ cells or during early embryonic development, it will be passed to the following generations and fixed in the population (112). Gene transcripts present as retrotransposed insertions in one or more individuals, but absent from the reference genome, are considered polymorphisms and are known as GRIPs (gene retrocopy insertion polymorphisms). Processed pseudogenes polymorphisms are present in many mammalian genomes including

mice, chimpanzees and humans and are an ongoing mechanism of mutation. In 2013, Ewing et al., explored GRIPs using available data from the 1.000 Genomes Project. They could ascertain whether a particular polymorphic PP occurred more frequently in one population than others. For example, insertions of *POLR2C*, *HSPE1* and *SNRPC* mRNAs appeared to be restricted in individuals with self-reported African ancestry. Moreover, they could report 22 human, 201 mouse and 9 chimp GRIPs in introns or exons that could lead to novel gene fusions, modifying their function (113). Recently, in 2021, sideRETRO was published as a mapping-based algorithm to identify retrocopies of genes, or PP in whole genome and exome sequencing data. Using this algorithm, the authors analyzed five individuals with WGS and WES data from 1000 Genomes Project. In the WGS data they could identify 20 retrocopies, whereas in WES from the same individuals only 6 candidates (117).

Processed pseudogenes are not only retrotransposed in germline cells, but also occur in somatic tissues including neural progenitor cells, stem cells, early fetal development, induced pluripotent stem cells and tumors. Evidence of somatic retrotransposition during early development has been observed in *Drosophila* and in humans, contributing to a variety of human diseases such as cancers and neuronal disorders (111,113,119). Among these events, somatic processed pseudogenes are also included as a product of the capacity to act on mRNA that LINE elements have. As an example, Boer et al. described an exonized retrotransposed *TMF1* gene inserted in the *CYBB* human gene, which knocked out the gene's activity. The PP insertion was identified in a Dutch man diagnosed from chronic granulomatous disease, an X-linked disorder. The newly created processed pseudogene linked with the disease, occurred during early embryonic development of the patient's mother and around 15% of her lymphocytes contained the insertion (120). As mentioned, somatic processed pseudogenes can occur in various cancers, so the estimation of de novo retrotransposition events

in normal and tumor cells is critical for understanding cancer formation and progression, as well as tumor heterogeneity (121).

2.4.1 Somatic retrotransposition events in cancer

In 1992, Miki et al. reported the first somatic retrotransposition event. In this case, they could observe an L1 insertion into the *APC* tumor suppressor gene of colorectal cancer (122). Before new L1 insertions were detected by next-generation sequencing, increased retrotransposition in tumors was predicted due to cancer-associated hypomethylation and elevated transcription of L1s. Although retrotransposition occurs at significant rates in normal somatic cells, they are more easily detected once the cell clonally expands as a tumor. In that case, the insertion would appear as a tumor-only event erroneously. However, somatic individual mutational events appear randomly on the genome and are later subjected to selective forces. Therefore, insertions proliferate preferentially in tumors than normal tissues since cancer cells divide more rapidly (94,111,119).

Each genome can have its own and unique active L1s, and they can vary between individuals in terms of activity having different “mutational power”. Consequently, retrotransposition occurs frequently in some tumors but differs greatly between cancer types, and individuals with the same cancer type. The disruption of mechanisms that usually suppress TE activity promote mutagenic retrotransposition in cancer. In 2012, by analyzing 43 WGS cancer samples Lee et al. identified 194 somatic insertions of transposable elements (TE). Authors developed a computational method (TE Analyzer or Tea) to detect the exact position and mechanism of TE insertions from paired-end WGS data. The evaluation of five different cancer types with Tea, reveals an average of TE insertions per tumor type ranging from 0 to 29. Colorectal tumors showed the highest frequency of somatic L1 insertions. In contrast, insertions were not identified in blood or brain cancer tissues (94).

Somatic mobilization of gene-derived transcripts has also been detected in cancer cells. Ewing et al. identified somatic processed pseudogenes by analyzing high depth sequences from The Cancer Genome Atlas. This study was the first comprehensive description of PP insertions in cancer. It included 85 pairs of tumor and normal genomes from acute myeloid leukemia (AML), breast cancer (BRCA), colorectal adenocarcinoma (COAD), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC) and ovarian carcinoma (OV). Comparing normal and tumor samples from each patient, three novel somatic processed pseudogene insertions were discovered in lung cancers (113). In a different study including 244 cancer patients, the percentage of somatic PP among all retrotransposition events was calculated. Of the total number of observed L1 somatic retrotransposition events about 2,3% cause mobilization of proximal exons or complete genes. Despite that, the range of genomic elements that can be targeted by transduction was known to be larger than just those near active L1 elements (111). The same year Cooke et al. published in greater detail the study of exclusively somatic PP insertions in cancer. Screening sequencing data from 660 cancer samples, they found 42 somatic PP in 17 samples (2,6%). These samples include 14 primary cases and 3 cell lines sequences. As an example, they described the insertion of all five exons of the gene *FOPNL*, into the eleventh intron of *SND1*. The somatic insertion identified in a lung cancer included a portion of the 5' UTR and the full sequence of the 3' UTR. Similar to the mentioned previous studies, acquired PP were present mainly in lung and colorectal cancer. These results correlate with high rates of somatic retrotransposition of LINE elements in these tumor types (110).

Somatic retrotransposition events have been mostly detected in cancers of epithelial cell origin with a rapid capacity to proliferate. Although many PP insertions seem to be early events in tumor formation, some of them have been shown to appear later during progression and not in all sections of the same tumor (113,119). Highly expressed transcripts are expected to be templates for somatic PP. In this way, the top expressed genes of a tumor tissue can be recurrently retrotransposed and inserted in a tumor genome (110,121). Many source genes seem to fall into similar functional categories. Gene ontology (GO) analysis of these genes includes terms like ribosomal function, metabolic processes, transcriptional regulation or signal transduction (113). Processed pseudogene insertions are more likely to occur in intergenic or heterochromatic regions than expected by chance. Also, in regions of the genome with a low exon density (111). Even so, insertions can also be located within annotated genes, and in that case tend to occur in genes frequently mutated in cancer including cancer drivers (94).

The disruption of target genes by PP insertions can have a significant impact on tumorigenesis. Despite the mutagenic potential of PP, it remains unexplored the extent of contribution to tumor formation they have. The majority of somatic PP are likely to be passenger mutations, but a few have oncogenic consequences. For example, PP insertions within cancer driver genes or the amplification of oncogene copy number may contribute to cancer development. Moreover, insertions in untranslated regions (UTRs) or introns can also alter cell's transcriptional activity, typically resulting in lower expression levels (110). The impact of retrotransposition events also depends on the orientation of the inserted sequence on the target gene, being antisense insertions less disruptive (94). Large scale studies across thousands of cancer genomes to identify somatic PP can help us to understand their impact on tumors.

2.4.2 Using NGS data to identify somatic retrotransposition events

The identification of somatic retrotransposition events, including processed pseudogenes, can have important implications for human cancer health. Diverse projects started working on the discovery of PP using next-generation sequencing data (110,111,113). Normal and tumor samples from initiatives such as 1000 Genomes, TCGA or the ICGC have been analyzed to identify both germline and somatic events. In this section we will focus only on the detection of somatic insertions.

Considering the mechanism PP are formed, there are various determining hallmarks to describe this event. First, it is important to identify 5' and 3' junctions of the sequence insertion within the target region. Paired-end reads spanning the insertion can be misinterpreted as balanced translocations (111), hence other features should be considered. As PP are the result of mRNA reverse transcription, a few sequencing reads should also cover exon-exon junctions of the source gene showing the absence of introns. Finally, the presence of a poly-A tail, or repeat sequences flanking the inserted sequence can be observed (119). To consider the event as somatic, this mentioned hallmarks should be observed on tumor but not on their matched normal DNA (110).

Massively parallel sequencing data, particularly WGS protocols, should help to explore the presence of somatic PP in cancer. However, sequence analysis pipelines usually lack sensitivity to detect rare insertions, especially if they occur late in tumor development (119). Heterozygosity and cellular and genetic heterogeneity of tumor samples can also result in lower frequency variants, adding a layer of complexity. On the other hand, when processed pseudogenes are flanked by repeat elements, their identification from short-read sequencing becomes a challenge. Nearly identical TE make difficult to differentiate the true

source or target regions (94). Therefore, and specific identification protocol together with manual inspection of the sequences and experimental validation of a significant number of candidates is needed to confirm their presence.

In 2014 Tubio et al., developed a bioinformatic pipeline named TraFiC (Transposome Finder in Cancer) (111). Their pipeline is capable of detecting various classes of retrotranspositions focusing on transposable elements, and it is not exclusive for processed pseudogenes. From paired-end sequencing aligned data, TraFiC inspects diverse read-pairs to identify insertions. Then, the pipeline uses RepeatMasker (www.repeatmasker.org) to identify TE-like sequences among unmapped reads with an aligned mate. Anchored reads with mates belonging to the same TE type, sharing the orientation are clustered. Reciprocal clusters represent both ends of one candidate TE insertion.

To specifically detect somatic processed pseudogenes in NGS data, Cooke et al., designed another bioinformatic method (110). The method was created to analyze targeted exome and genome-wide studies in cancer. In this case, paired-end reads were aligned to the reference genome and transcriptome. These alignments allow them to identify reads across canonical splice sites and between a pseudogene and its insertion region. However, their method required at least three exons from a single gene represented in the tumor DNA. To validate candidate somatic PP, they performed PCR on tumor and matched normal samples.

As presented in this thesis, we studied somatic processed pseudogenes using 2.589 tumor samples from the PCAWG dataset. After our results were published, other pipelines were created with similar purposes. SideRETRO, for example, detects somatic and polymorphic insertions of retrocopies and processed pseudogenes retroCNVs (117). This method is a mapping-based algorithm that uses WGS or WES to identify the mentioned events, and provide their genomic

insertion sites, zygosity, genomic context and parental genes. Comparable with the method developed by Cooke et al., sideRETRO requires aligned sequences, a reference genome and a reference transcriptome. Yet this pipeline was not able to identify insertions within highly repetitive genomic regions.

2.5 Translated small open reading frames: micropeptides

Around 20,000 human genes are annotated as protein-coding genes, covering less than 2% of the human genome (123). However, large-scale analysis and computational advances have revealed that a larger portion of the genome is transcribed and, at times, translated. Among this portion of the genome, a considerable fraction of genes produces transcripts with mRNA-like features but apparently without coding potential. These transcripts are long non-coding RNAs (lncRNAs) and are longer than 200 nucleotides (124–126).

The number of novel transcripts obtained from RNA-seq increased the attention paid to identifying the complete set of noncoding genes and protein-coding ones. Ji et al., showed in 2015 that 40% of lncRNAs and pseudogenes expressed in human cells were translated and could potentially be functional proteins (127). Not only within these ncRNAs, but also within 5'UTR or intergenic regions, a new class of genetic elements named small open reading frames (smORFs) has been discovered in the last years. These missing coding genes added complexity to the human genome annotation and proteome characterization. By definition, smORFs are sequences of less than 300 nucleotides and small proteins known as micropeptides can be directly translated from these short mRNAs. Micropeptides, which comprise a sequence of in-frame codons, may be of low abundance and can have tissue- and time-specific expression patterns. They differ from known bioactive small peptides as they are not the result of post-translational cleavage and modification of large pre-proteins, but are translated from smORFs (128,129). These novel genetic elements have been misunderstood since classical ORF-finding algorithms set a threshold length of 300 nt or 100 amino acids to detect them (125).

Analysis of smORFs coding sequences not only revealed that these genes had been discarded because of their short length, but also because of the classical assumptions and expectations about a canonical gene structure and sequence. The application of novel proteomic techniques has provided key findings regarding the use of non-AUG initiation codons in human translation, as well as in other eukaryotes and prokaryotes (128,130). In 2018, short and non-ATG-initiated open reading frames that express proteins were found in non-protein coding genes in mice (131). Diverse reports calculated that between 50 and 70% of smORFs detected do not initiate with canonical AUG start codon. Percentages differ depending on the experimental technique used for the study. The observed frequency of canonical AUG start codon occupancy by ribosome profiling is 49,76% in humans and mice, followed by CUG (15, 44%), GUG (7,17%) and UUG (4,17%) (132).

The identification of this hidden proteome opens the possibility to better understand human biology and disease. Although experimental validation of each peptide is needed to ultimately confirm their biological role, the function of several micropeptides have been characterized. It is known that they can act as regulators of larger protein complexes such as membrane-associated proteins (124), but also independently in different manners. The first functional encoded smORF in animals was described in 2007 by Galindo, M. I. et al. Their study was focused on the *tarsal-less (tal)* gene in *Drosophila*, which expresses a 1.5 kilobase (Kb) transcript previously classified as noncoding. Its classification was based on having no ORF longer than 100 aa and no known homologies. However, several candidate smORFs are present in the *tal* transcript and the peptides translated from ORFs of just 11 aa mediate the function of the gene, having an important role in development. *Tarsal-less* homologous genes were also identified in other species, defining a new noncanonical gene family in metazoans and of ancient origin (133). In 2008, 217 smORFs were identified using bioinformatics in

Escherichia coli, and 18 were found to be needed for bacterial growth (134). In humans, diverse studies have demonstrated that micropeptides are known to act as regulators of biological processes such as DNA repair, RNA decapping, calcium homeostasis, metabolism, stress signaling, myoblast fusion and cell death (125). An 84-aa-long conserved peptide named Protein myomixer that mediates myoblast fusion (135), or the *SPAAR* gene translated into a 90 aa micropeptide which regulates muscle regeneration (136) are two examples of these known functional micropeptides in humans. Linked to disease, Huang J et al. discovered in 2017 the HOXB-AS3 peptide translated from the human lncRNA *HOXB-AS3*. This micropeptide of 53-aa length suppresses colon cancer growth, and its loss is a critical oncogenic event in this tumor type (137). Micropeptides with a significant biological role are not only encoded by nuclear transcripts but also by the mitochondrial genome. Humanin is translated from a mitochondrial smORF and it is involved with programmed cell death (128). The functions of known micropeptides are very heterogeneous. A list of known functional micropeptides is provided in Table 1.

All these independent functional studies, together with the realization that hundreds or thousands of smORFs are translated and conserved across metazoans, demonstrated the importance of exploring micropeptides to understand many aspects of biology and medicine clearly. Understanding their origin, evolution and role is essential to clarify this underappreciated function of the genome.

Micropeptide	Conservation	Function	Size (AAs)	References
Polished rice (Pri)	Insects	Fly embryogenesis	11-32	(133,138)
Toddler	Vertebrates	Promotes cell migration	58	(139)
AGD3	Mammals	Involve in stem cell differentiation	63	(140)
Myoregulin (MLN)	Mammals	Calcium homeostasis	46	(141,142)
DWORF	Lamprey	Enhance muscle performance	34	(143)
Myomixer	Vertebrates	Involve in controlling muscle performance	84	(144)
MRI-2	Mammals	DNA repairing process	69	(145)
NoBody (NBDY)	Mammals	MRNA recycling	68	(146)
SPAAR	Human and mouse	Regulate muscle regeneration	90	(136)
Humanin	Different species	Involved in program cell death	24	(147)
MOTS-c	14 species	Metabolic homeostasis	16	(148)
Minion	Mammals	Muscle formation	84	(135)
HOXB-AS3	Primates	Suppresses colon cancer growth	53	(137)

Table 1. - Micropeptides identified in animals and their biological functions. Table extracted from (128).

2.5.1 Classification of small ORFs

New ORFs, and in this case small ORFs, are usually classified according to their relatively localization in known transcripts (Fig 17).

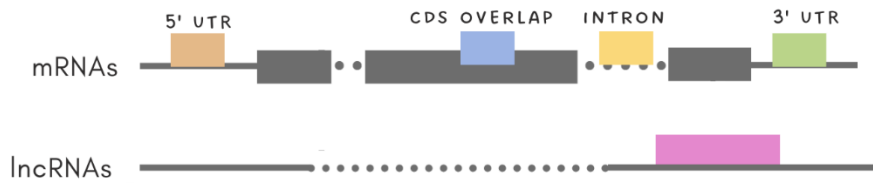


Figure 17. Classification of smORFs based on their location and considering known annotated genes (grey).

Considering gene structure, smORFs can be classified depending on whether they overlap with 5' or 3' UTRs, introns or exons of known transcripts, as well as with long non-coding RNAs or pseudogenes. They can be found in alternate CDS frames or starting from non-canonical codons (149). Small ORFs have been also identified among intergenic regions. However, there is no standard classification or labeling for small ORFs, and diverse classes are described depending on the study. Evidences of translation have been observed for all types of transcribed smORFs, with different translation efficiencies and chance of detection. Size, average rate of translation or the level of conservation differ among these classes (124).

Open reading frames within intergenic regions seem to be the most numerous in fruit flies and mammals and are known to have a median size of 22 codons according to a study done in 2017 (124). However, some studies considered that intergenic smORFs are randomly generated by our genomes, expecting not to be transcribed, nor functional. Therefore, the majority of the

studies working on the identification of functional smORFs or micropeptides do not consider this class of smORFs to avoid inflating the estimates.

It is known that a significant fraction of the translome maps to untranslated regions and sequences previously considered to be noncoding (150). Consequently, the second most abundant class of ORFs are identified within the 5' untranslated regions of mRNAs encoding canonical proteins. Upstream ORFs (uORFs) have been reported in many organisms including yeast, flies, zebrafish and mice. They commonly regulate the translation of the downstream canonical ORFs in their transcript, and their presence often produces a repressive effect on transcription or translation of the main coding sequence. Translated uORFs have also been shown to form protein complexes with the protein encoded from the main CDS of the same mRNA. Its pure cis-regulatory role fits with their low translation levels and low sequence conservation (151).

Long non-coding ORFs (lncORFs) are small ORFs found in putative lncRNAs, and the third most abundant class. Their size distribution is similar to that of intergenic ORF and uORFs, with a median of 23 codons. Several RNAs previously classified as lncRNAs have been shown to encode and translate peptides with biomedically important functions, and to be highly conserved in evolution. Although their amino acid usage is similar to random sequences, ribosome footprints have been also detected in this smORFs class suggesting translation. Since in lncORFs there is no downstream ORF encoding a functional protein, it is difficult to imagine they have regulatory functions. However, it has been hypothesized that they protect translation of downstream elements (152).

Lastly, smORFs found in exons of functionally monocistronic transcripts have a median size of 79 codons and seem to be translated as efficiently as canonical ORFs. Their amino acid composition resembles known protein coding genes and

differs significantly from randomized RNA sequences. Hundreds of them have been identified in humans.

2.5.2 Identification of micropeptides

Conventional gene annotation with ORF-finding algorithms has systematically discarded small ORFs as coding genes because of their level of uncertainty in terms of functionality, given their short length (128). Because they fall close to the transcriptional and translational noise of the cell, both in size and in expression levels, the validation and the functional characterization of micropeptides have been challenging and limited, even at experimental level. Computational and experimental approaches have been developed and implemented to deduce coding potential, examine transcription and translation of novel regions and identify putative protein products generated from sequences previously annotated as noncoding, including also UTRs, introns and intergenic DNA. Computational methods allow researchers to determine all possible ORFs, but their results will probably include ORFs that are not translated or functional. In contrast, experimental techniques such as ribosome profiling (Ribo-seq), mass spectrometry (MS) or western blot and immuno-cytochemistry, can directly discover protein products. However, these methods, especially the last two, are not sensitive enough to detect low abundant micropeptides (125). In addition, to complicate things even more, the expression and function of micropeptides is tissue and time dependent. Overall, tiny sizes, low abundances, rapid degradation and sample loss during preparation steps result in many technical challenges and difficulties to work with micropeptides.

Applying both computational and experimental approaches appears to be the best strategy for the study and identification of micropeptides. Combinatorial methods can identify ORFs actively translated, non-canonical or species specific. However, experimental data is needed, and often additional samples for low

expressed transcripts (126). Transcription and/or translation are two criteria for assuming that smORFs are functional even if they are within coding or non-coding regions. Therefore, proteogenomics, that is the combination of peptidomics and massively parallel RNA-sequencing seems to be an interesting field to discover novel coding regions (125). Advancements over the past few years in diverse technologies, allow scientists the discovery of a considerable set of putative coding smORFs. Below, a brief description of the computational and experimental methods most used in large-scale studies is given.

2.5.2.1 Computational annotation through in-silico evolutionary approaches

Several strategies have been used to systematically annotate small ORFs with coding potential (128). Based on *in-silico* translation of annotated transcript regions, a set of smORFs can be obtained. Transcripts should be converted into amino acids following the corresponding genetic code. Usually, ORF are identified using the most upstream canonical start codon (AUG) for each stop codon within the sequence. The translation could be done starting from the first, second and third nucleotide (3 in-frame), and for both forward and reverse strand (6 in-frame) (153). Diverse studies also include non-canonical start codons as translation origins. After translating the selected transcripts, sequences of 100 aa or less, are defined as putative smORFs and therefore, candidate micropeptides. These computational methods can identify all possible ORFs, even sequences are low expressed or tissue specific, and without needing experimental data. However, the results may include ORFs that are not translated and do not correspond to micropeptides (126).

In addition to experimental validation, which is explained below, conventional computational strategies have been invented and used to calculate the coding potential of small ORFs. These strategies evaluate codon content,

nucleotide composition, sequence homology, conservation between species, or secondary structure (125,126,128). As an example, Mackowiak et al., developed and implemented in 2015 a computational method to identify smORFs with high accuracy by using conservation features and codon and amino acid usage. Their identification started with *in-silico* translated sequences from an annotated transcriptome together with published lncRNA catalogs. They identify hundreds of previously unknown conserved smORFs in humans, mice, zebrafish, fruit fly and *C. elegans* (153).

Among all the mentioned features, evolutionary conservation is a key sign that a genomic region is functional. In gene prediction, cross-species comparisons are a powerful technique since most genes are subject to evolutionary pressure to preserve their function and, therefore, their amino acid sequence. Therefore, the conservation of putative coding sequences indicates purifying selection and can be used to infer function through the identification of similar proteins sequences with known function. (124,125,154). The term homology, used for proteins and genes encoding it, refers to two sequences that have a common ancestry. Two segments of DNA can share their ancestry because of speciation events (orthologs) or duplication (paralogs). Whereas orthologous genes generally conserve their main function, paralogs become different in sequence and function over time (155). Orthology-based searching among species, commonly based on sequence similarity, is performed to predict conserved biological functions to annotated novel genes, or in this case, micropeptides. Myoregulin, Phospholamban and Sarcolipin are some examples of micropeptides identified from homology-based characterization. This group of micropeptides share conserved peptide sequences and structure from flies to vertebrates, and they are involved in Ca²⁺ homeostasis (126,128).

True conservation and homology are difficult to establish considering the short length of smORFs. Compared with canonical proteins, smORFs have lower quantitative conservation scores. Moreover, they have a higher probability of obtaining low conservation scores by chance (128,156). It is also important to ensure that sequence similarity is not because of short divergence time between the species (150). However, diverse studies have shown that smORFs are widely conserved on the sequence level in human and other species (124,126,129). Sequence conservation rarely occurs far beyond the ORF and the absence of insertions or deletions within their sequence implies conservation of the reading frame.

Functional micropeptides also display a characteristic depletion of non-synonymous compared to synonymous mutations when compared to their orthologs (125,153). Generally, functional genes that are essential for cellular processes are subjected to selection pressures showing a reduction of non-synonymous variants, trying to preserve their amino acid sequence and their function. Therefore, mutations that result in changes to the amino acid sequence, are often selected against, and discarded through purifying selection. On the other hand, synonymous variants, which do not alter the peptide sequence, are less constrained and may be more tolerated and fixed within the population. These different levels of selective pressure acting on synonymous and non-synonymous substitutions in functional regions can be used as a signal for functionality. This can be calculated using the substitution ratio (dN/dS), which is defined as the ratio of non-synonymous to synonymous substitutions. The substitution ratio is therefore a useful measure of the strength and mode of natural selection action on protein-coding genes. When there are strong structural constraints on a protein there is little or no accumulation of non-synonymous changes. Therefore, the ratio for this sequence will approach zero. In contrast, if protein sequences are not under selection the ratio will be

approximately 1 (157–159). As an example, a program package for identifying smORFs with high-coding potential was developed in 2010. The analysis pipeline named sORF finder, is based not only on the hexamer composition of nucleotide sequences but also evaluates synonymous and non-synonymous substitution rates (160). Other computational identification methods are shown on Table 2.

Although conservation is useful to functionally characterize new smORFs, it is not applicable for all. For example, evolutionary analyses are not able to infer protein-coding or regulatory potential for “young” de novo proteins (152,156). It is known that up to 1% protein-coding genes could be species-specific and of recent origin. This idea is controversial and depends on the ability of computational approaches to detect homologues. Whereas some studies conclude some functional micropeptides are conserved, others support that most translations do not show signs of constraint as coding sequences (123,124,128,156).

Evolutionary conservation often suggests potential gene functionality. Nonetheless, the mere presence of a conserved and translated peptide does not inherently imply a critical or definitive biological function.

Prediction tool	Description	Website	References
PhastCons	Identification of evolutionary conserved elements in a multiple alignment, given a phylogenetic tree.	http://compge.n.cshl.edu/phast/	(161)
PhyloCSF	Determines whether a multi-species nucleotide sequence alignment is likely to represent a protein-coding region. Examines the frequency of synonymous codon substitutions and conservative aa substitutions, and low frequencies of other missense and non-sense substitutions.	http://compbio.mit.edu/PhyloCSF	(162)
miPFinder	Identifies and classifies potential microproteins, small single-domain proteins that act by engaging their targets into protein complexes. It takes into account protein size, domain origination, known protein interactions and evolutionary origin.	https://github.com/DaStraub/miPFinder	(163)
MiPepid	Machile-learning tool using logistic regression with 4-mer features. Predicts whether a sequence encodes a micropeptide based on its DNA sequence.	https://github.com/MindAI/MiPepid	(164)
SORF Finder	Program package for identifying smORFs with high-coding potential. Based on the hexamer nucleotide composition and the potential functional constraint at the aa level through evaluation of syn and non-syn substitution rates.	http://evolver.psc.riken.jp/	(160)

smORFunction	Provides function prediction for smORFs by analyzing their correlated genes with known functional annotations.	https://www.cuilab.cn/smorfunction/home	(165)
uPEPperoni	For 5'UTR smORFs, based on conservation.	http://u pep-scmb.biosci.uq.edu.au/	(166)

Table 2. - Current computational methods for smORF identification.

2.5.2.2 Ribosome profiling to monitor translation

Ribosome profiling is a deep sequencing method of mRNA fragments attached to ribosomes that provides a genome-wide snapshot of active translation (126,130). Ribosomes are complex molecular machines that link amino acids in the exact order within a transcript to produce a protein product by translating it (125). Stalling ribosomes on mRNA and protecting the portion of mRNA from nuclease digestion, ribosome-protected RNA fragments (RPFs) can be converted into DNA libraries for reading their sequence. For each RPF a ~30 nucleotide portion of mRNA is sequenced, producing a footprint fragment whose sequence can be mapped indicating its exact position on the reference genome and the mRNA it was translating. Ribosomes scan the coding sequences one codon at a time, showing a characteristic three-nucleotide periodicity of the translated region. Ribo-seq not only provides information about ribosome positions but also reports the amount of translation of a gene (130). Changes in protein expression that cannot be explained by transcript levels and translational regulation can be studied by combining ribosome footprint density and mRNA abundance measurements. Furthermore, since ribosome profiling requires only the nuclease footprint from ribosomes, it is less sensitive than RNA-seq to compromise RNA integrity of the sample (167).

Ribo-seq helped to extend the understanding of human genome translation and revealed thousands of open reading frames within noncoding and presumed untranslated regions (150,167). Mudge et al. recently published the first phase of an ongoing project. The aim of the consortium is to produce a standardized catalog of human Ribo-seq ORFs longer than 16 aa, to bring protein-level evidence into reference annotation databases. The presented catalog is the result of analyzing seven Ribo-seq ORFs datasets, however the consortium will incorporate a greater diversity of human cell types and tissues (156).

However, there is still a technical debate on whether low signal levels represent productive translation or not. It is also known that strong association and ribosome occupancy does not always guarantee active translation of the region (124,126,128,130). Considering that actively translating ribosomes have a discrete movement along the mRNA in three nucleotide steps, methods such as the ORFscore have been developed to quantify the biased distribution of RPFs and reduce noise in conventional analysis (129). They applied ORFscore to long non-coding RNAs and uncharacterized processed transcripts from Ensembl. By analyzing published ribosome foot printing data in HeLa cells, they could define 135 translated smORF.

Other algorithms and metrics have been created based on ribosome-profiling characteristics. RiboTaper, for example, exploits the subcodon resolution of the obtained sequencing reads to reconstruct the full set of ORFs in coding and non-coding transcripts. Applying this algorithm, Calviello et al., could identify 504 non-coding genes that harbor translated ORFs (130). Although some of the encoded ORF identified were shorter than 300nt, this study was not centered on micropeptides but in actively translated ORF. Researchers conclude that quantifying the presence of significant ribosome footprint reads in regions shorter than 20 amino acids becomes difficult.

It also needs to be considered that true coding potential and function at protein level is not certainly implied even though ribosome occupancy is observed. Translation can have regulatory consequences, for example modulating downstream ORF or peptides could be unstable.

2.5.2.3 Mass spectrometry to directly detect peptides

The gold standard in proteomics research is mass spectrometry, a powerful technique to directly detect and quantify proteins and peptides (125). This analytical tool measures the mass-to-charge ratio (m/z) of one or more molecules present in a sample (Fig 18). Using these measurements, the exact molecular weight of the sample components can also be calculated to identify unknown compounds, quantify known proteins and determine their structure and chemical properties (168,169) MS-based approaches help deciphering post-translational modifications and infer insights in biological functions and signaling pathways.

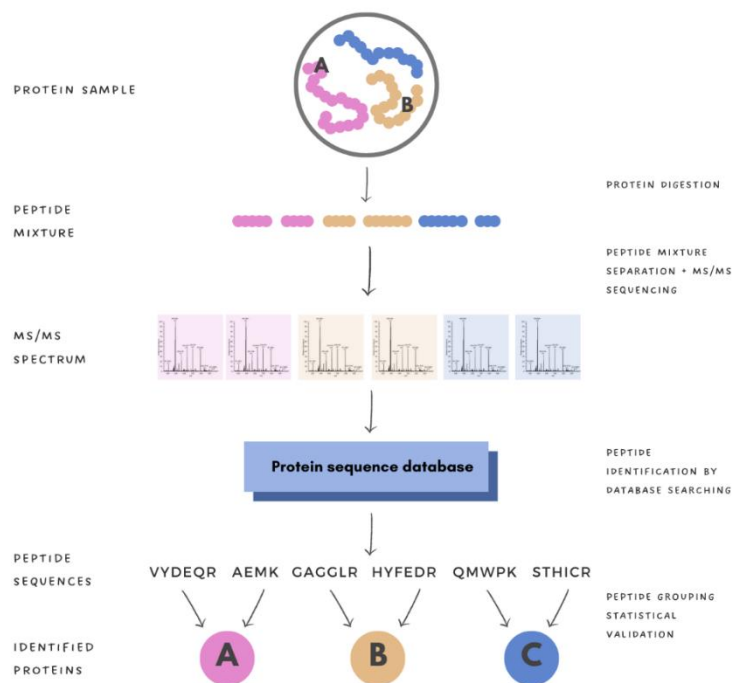


Figure 18. Schematic representation of MS/MS experiments.

High-throughput analysis can be done using MS as this application can be automated.

Usually, the protocol starts with the digestion of the complex mixture and consists in a combination of high-performance liquid chromatography (LS), used to separate the resulting peptides, followed by tandem mass spectrometry. This protocol, referred as shotgun proteomics, has been applied to identify and validate smORFs encoding micropeptides. However, since the detection of peptides depends on factors such as sequence length and abundance, novel micropeptides appear to be underrepresented when using shotgun proteomics and its detection using MS faces diverse challenges (125,128,129,170). With the aim of identifying micropeptides, LC/MS/MS protocols should be modified (171).

Detection is naturally biased towards the detection of more abundant proteins. The average tissue content of micropeptides is very low, and because of their instability are often subjected to rapid degradation or loss during sample preparation (125,128,154,171). This insufficient concentration in cells makes micropeptides detection difficult. Their identification can be likely benefit from an enrichment step during sample preparation. Therefore, the discovery protocol begins by enriching the proteome for low molecular weight peptides and small proteins. As an example, an study focused on improving the identification of encoded smORFs concluded that samples extracted in the lysis buffer detected most micropeptides, whereas acid extraction resulted in the fewest number (170). These efforts allowed them to identify 37 novel human micropeptides from non-annotated coding RefSeq regions in a lung cancer cell line.

Since complex mixtures are difficult to fully analyze by MS, enzymatic digestion is performed obtaining a large number of peptide products. Trypsin is the proteolytic enzyme generally used, which cleaves peptides between arginine or lysine and the adjacent amino acids. Yet, the small size of micropeptides and

their tendency to contain fewer arginine and lysine residues, results in a low number of tryptic peptides generated. Encoded smORFs do not generate large enough signatures and have to be typically identified from a single peptide (125,154,170). To improve its detection, alternative proteases can be used in combination with trypsin (171).

After obtaining m/z ratio from the mass spectrometer, the measurements have to be compared with a protein database to determine the sample compounds. Standard MS protocols generally utilize databases of known proteins and/or *in-silico* translated sequences using the canonical AUG start codon. However, micropeptides have been previously systematically missed by genome annotation because of their length and therefore, not included in these mentioned databases. Moreover, diverse studies have confirmed that non-canonical start codons can also initiate translation (154). To solve this problem, custom generated databases have been used to identify non-annotated proteins such as micropeptides. Protein sequence databases can be generated combining genomic and transcriptomic data, for example, by performing three-frame translation of the reference transcriptome, or RNA-seq data from a specific sample. Known proteins are then computationally excluded from the dataset. Proteogenomics, has enabled detection of missed gene products (128,170,171). Creating a custom database containing all short peptides that could be translated from the annotated transcriptome may result in a large set of peptides. Because of the inflated search space, this strategy suffers from reduced sensitivity and reliability. To avoid false positives peptide-spectral matches, expression level cutoffs, or cell- and condition-specific RNA-seq data should be curated for each experiment. By combining proteogenomics with RNA-seq experiments on K562 cells and restrictive filters, Slavoff et al., confirmed the presence of 37 micropeptides in this cell line (172).

MS proteomics offers direct evidence of encoded small ORFs, although the results can be biased depending on sequence composition. Numbers of identified micropeptides vary not only between organisms but also among cell lines and tested conditions. This information can also help reveal their biological significance. Data obtained from MS coupled with genomic, transcriptomic and translomic data provides an alternative validation.

2.5.3 Published databases to study micropeptides

Diverse numbers of studies using computational, experimental or combining both approaches for the identification of mp have been published during the last 8 years. Some of them provide public repositories and web tools to examine and download identified micropeptides.

HaltORF was the first web-based searchable database that allows the exploration of the human transcriptome of out-of-frame alternative open reading frames with a start codon located in a strong Kozak context. Products of out-of-frame alternative translation initiation result from distinct initiation codons located in different ORF in known human mRNA. Although they provide protein sequences of at least 24 amino acids long, it was not exclusively focused on smORFs (173).

In 2016, sORFs.org (174) was published as a novel repository of smORFs identified using ribosome profiling. Experimental results from ribo-seq data are combined with conservation analysis and MS rescanning. In their latest version (175), authors provide smORFs identified in human, mice, fruit fly, zebrafish, rat and *Caenorhabditis elegans*. They include 78 ribo-seq datasets, 34 of them from human cell lines. Through their website (www.sorfs.org) you can, by default, quickly lookup for smORFs. A BioMart interface is also provided for advanced query and data exportation. For each smORFs, sorfs.org includes their genomic coordinates, the transcript and amino acid sequence, the annotation depending

on its location (5' or 3' UTR, lncRNA, pseudogene, intronic, exonic, intergenic) and the cell lines where it has been identified. Based on ribosome profiling data, metrics such as ORFscore are calculated to indicate true coding sequences. Conservation evidence are examined using PhyloCSF and sequence variation is annotated from dbSNP, ClinVar and Cosmic databases.

Another smORF repository, specifically including small proteins identified in lncRNA was published in 2017. SmProt collects data from ribosome profiling and mass spectrometry experiments, known databases and literature mining. The first version of SmProt includes 255,010 small proteins from 8 species including human, and 291 cell lines or tissues (176). The new web server (<http://bigdata.ibp.ac.cn/SmProt/>) can be used for search, browser, download and submit information. Small ORFs are mapped to the genome and classified depending on their location on known transcripts. On their updated version, they improved the identification algorithm increasing its accuracy, predicted disease-specific translation events and variants in smORFs and included small peptides with non-AUG translation initiation. By analyzing 6,419 new ribo-seq datasets they upgraded the number of small proteins to more than 3.6 million records.

OpenProt was published and available in 2019 with the aim of offering a deeper and a more realistic and biologically relevant perspective of the proteome (177). Although it is not focused only on micropeptides, it includes all ORFs longer than 30 codons identified in transcripts, ncRNA and pseudogenes reported by Ensembl and RefSeq. OpenProt contains all possible ORFs within the mentioned sequences across 10 species. It also cumulates supporting evidence such as protein conservation, translation and expression.

Lately, and after starting our project in micropeptides, a repository of unique smORFs identified in human and mice was released to allow comparison from distinct original data sources. MetamORF (151) has been built collecting publicly available smORFs data, reprocessing, normalizing, homogenizing it and summarizing redundant information. However, MetamORF does not provide novel sequences. It gathers data from sORFs.org, OpenProt, SmProt, uORFdb, a comprehensive literature database on eukaryotic upstream open reading frames (178), TisDB, a website providing alternative translation initiation sites (179), and other RNA-seq and Ribo-seq or MS data repositories including RiboSeqDB, PITDB (180), TranslatomeDB (181) and RPFdb (182). MetamORF describes 664.771 unique ORFs, including small ORFs, in the human genome, providing information to locate them on the genome. Also in 2021, nORFs.org was publicly available, containing 194.407 ORFs curated from OpenProt and sORFs.org. The length distribution of ORFs in nORFs.org falls mostly below 100 amino acids and all sequences have translation evidence from MS or ribosome profiling experiments (149).

Small ORF have been usually identified within annotated coding and non-coding regions but not in intergenic sequences, being generally unexplored. However, the last two repositories as well as the updated version of SmProt and sORFs.org start including encoded smORFs in non-annotated sequences. All these databases can be useful to benchmark new smORF-finding algorithms as well as to, for example, add more experimental evidence on an *in-silico* obtained set.

3. Motivation and objectives

The general goal of this thesis is to expand the understanding of the genomic basis of the biology of tumors through the study of the potential contribution of concrete processes and elements, such as somatic processed pseudogenes and micropeptides.

The presented thesis can be divided into three main chapters covering different projects, and activity periods which resulted from the combination of prioritizing research opportunities, data availability and possibilities for publication. The chronology and the motivation behind the strategic plan of the presented thesis is explained in section 1: strategy and thesis trajectory.

For the sake of clarity, this thesis is divided and organized at thematic and conceptual level, without considering chronology, resulting in the three following blocks:

Chapter 1 - Analysis of somatic structural variants in CLL and their incorporation into subclonality studies.

As part of a wide study to understand the genomic and molecular basis of Richter Transformation in some CLL tumors (lead by Dr. E Campo from IDIBAPS), our first aim and final contribution was centered in the general characterization of SVs within these tumors (together with Dr. Royo from the group). In addition, and in the same context, I also aimed at exploring and designing strategies to characterize the distribution of SVs across the different subclones in these tumors, which remains as an unsolved challenge within the community. In particular, we here aimed at:

- 1) General characterization of SVs in CLL tumors (with Dr. Royo): identification and manual validation of somatic SVs through the analysis of short-read whole genome sequencing data.

2) Define and generate strategies and methodologies to classify and assign specific somatic SVs to previously defined tumor subclones of CLL tumors.

Chapter 2 - Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact.

In the context of the PanCancer Analysis of Whole Genomes project and, in particular, within a study of somatic retrotransposition events in cancer (lead by Dr. Tubio, Universidad de Santiago de Compostela) we had the opportunity to contribute with a study, also related to structural variation in cancer, but now focused on somatic retrotransposition events that generate processed pseudogenes across a wide range of tumor genomes. Here, we aimed, not only to identify and characterize somatic PPs at genomic level but also to assess their functional impact on tumoral cells through the analysis of gene expression data. Our specific goals are:

3) Develop and apply a methodology for the identification of somatic processed pseudogenes across multiple cancer types by using short-read tumor and normal genomic sequence data,

4) By using both genomic and transcriptomic data, we aimed at evaluating the potential contribution of somatic PPs to tumors at functional level, both through the disruption of functional elements (genes) in the genome, as well as through their impact in gene expression as fusion transcripts.

Chapter 3 – Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data.

Small open reading frames are a new class of genes, currently unexplored in cancer. Our main goal within this part of the thesis is the identification and characterization of previously unknown micropeptides across the entire human genome and to investigate their potential role in cancer. We did this at two different levels: as part of a collaboration with the groups of Dra. Abad (VHIO) and Dr. Hector Peinado and Dr. Javier Muñoz (CNIO) that covered experimental, bioinformatic and mass-spectrometry identification and validation of Pancreatic Adenocarcinoma (PACA) associated micropeptides; and internally in the group with the aim of finding and annotating, at genome-wide level, all detectable unknown intergenic micropeptides and to inspect their potential role in a wide range of cancer types. The specific goals are:

5) To define a new catalog of candidate micropeptide sequences for the mass-spectrometry searches, using transcriptomic data from pancreatic cancer samples.

6) To identify new candidate intergenic smORF in the human genome using comparative genomics and evolutionary conservation features and properties at DNA and protein level,

7) To evaluate these findings by assessing their expression levels in normal publicly available transcriptomes including diverse tissue types,

8) To identify candidate cancer driver smORFs by searching for somatic single nucleotide variants detected in The International Cancer Genome Consortium.

4. Methods

4.1. Analysis of somatic structural variants in CLL and their incorporation into subclonality studies

Chapter 1

The first chapter of this thesis summarizes the work we did in collaboration with Dr. Elias Campo and Dr. Ferran Nadeu from IDIBAPS, and Dra. Romina Royo (BSC). As part of a larger study which included CLL longitudinal samples with the aim of understanding the biological basis and evolution of this cancer type, we worked on the characterization of somatic SVs. Moreover, we evaluated strategies to infer tumor subclonality based on these somatic variants, that are usually excluded from ITH studies.

4.1.1 Chronic lymphocytic leukemia longitudinal study cohort

The genomic study of Richter transformation in chronic lymphocytic leukemia was approved by the Hospital Clinic of Barcelona Ethics Committee and lead by Dr. Elias Campo. This study includes a total of 19 chronic lymphocytic leukemia patients (9 female, 10 male) fulfilling the criteria of Richer transformation (RT). The complete change into this more aggressive cancer form was validated through pathological revision of all collected samples. Three out of 19 cases developed RT before therapy, whereas in the remaining cases the aggressive transformation occurred after chemoimmunotherapy or after multiple lines of treatment. Almost all patients (17) transform into a diffuse large B-cell lymphoma-type, one developed a plasmablastic lymphoma transformation and one had a prolymphocytic leukemia transformation. Within this cohort, 15 tumors had unmutated IGHV (U-CLL) and 4 had mutated IGHV (M-CLL).

For all except one case longitudinal samples (range 2-8 samples/case) were collected at different time points of the disease. Purity and tumor contamination were considered to discard samples. The complete dataset encompassing germline, CLL and RT samples was available for 12 patients, while 6 patients lacked germline material and 1 case had not the previous CLL sample but only the RT.

This study cohort including 19 CLL patients was widely analyzed, described and published (3). The structural variants identification pipeline, including the merge of the results across variant callers and the rescue of SVs explained in this chapter was applied to all CLL patients (13) where both germline and tumor material was available. However, during all the work we have done on this project and particularly when analyzing SVs subclonality, we mainly explored case 63. Other cases including 365 and 1669 were also evaluated.

4.1.1.1 Disease course of one pilot CLL case

From this longitudinal cohort, we mainly worked with one case. We used case 63 (male, unmutated IGHV patient) as a pilot to explore and define the strategy designed for quantifying variant allele frequencies of somatic structural variants (Fig 19).

Tumor (4) and normal (1) WGS were available for case 63. Three different time points were explored (T1, T2 and T3), having two tumor samples from distinct tissues (peripheral blood – PB and lymph node – LN) for the first time point. Time points one (T1) and two (T2) corresponded to samples diagnosed as CLL, whereas the third time point (T3) was collected after Richter transformation. Regarding its type of transformation, case 63 had a diffuse large B-cell lymphoma-type (RT-DLBCL), as most of the studied cases in this cohort. Its chronic lymphocytic leukemia tumor transformed into Richter after ~ 10 years from diagnosis.

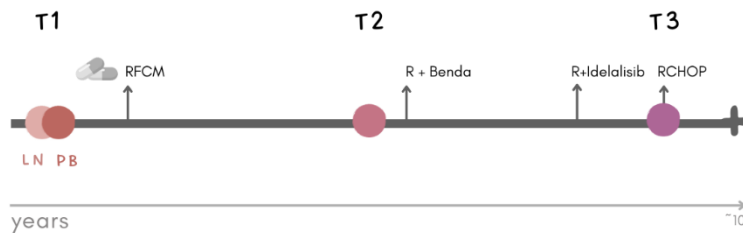


Figure 19. CLL case 63 follow-up. Longitudinal samples collected in time points 1, 2 and 3 are represented as circles. Two samples from different tissues (LN and PB) are collected in T1. Sample in T3 corresponded to RT. Treatments are shown above the arrows indicating when and how the patient was treated. Samples are named based on the time point and tissue.

Cases within this CLL cohort were grouped based on the therapy received prior to RT. For case 63, its CLL tumor transformed into Richter after receiving diverse targeted therapies. Concerning its treatment course, tumor samples (2) corresponding to the first time point were collected before any treatment was given to the patient. Between the first and second time points, a combination of rituximab, fludarabine, cyclophosphamide and mitoxantrone (RFCM) was given to the patient. Two more different target therapies (rituximab in combination with bendamustine, and idelalisib together with rituximab) were used as treatments after the second time point and before Richter transformation (third time point).

4.1.2 Whole genome sequencing and alignment

Whole-genome sequencing (~30x coverage and ~126/151 bp paired-end reads) was performed for all patients including all the available tumor and normal samples. After sequencing, a collection of reads encoded with the 4-letter alphabet (C, G, T and A) referring to DNA nucleotides (cytosine, guanine, thymine and adenine) were stored in FASTQ files. Quality alignment scores for each nucleotide were also supplied in FASTQ files.

Paired-end reads were mapped to the human reference genome (GRCh37) using BWA-MEM (v.0.7.15, <https://github.com/lh3/bwa>). After alignment, for each sequencing read its location on the human genome, mapping quality values and mate read information were outputted in SAM files. The obtained SAM files were converted into BAM and sorted using biobambam2 (v.2.0.65, <https://gitlab.com/german.tischler/biobambam2>). FastQC (v.0.11.5, www.bioinformatics.babraham.ac.uk/projects/fastqc) and Picard (v.2.10.2, <https://broadinstitute.github.io/picard>) were used to extract quality control metrics including the mean coverage for each sample.

4.1.3 Somatic structural variants identification

Although somatic single nucleotide variants, short insertions and deletions and copy number variation were also identified for the tumor genomes within this cohort, the work presented in this thesis is only focused on somatic structural variants. Therefore, only the variant calling analysis of SVs is described below.

A huge variety of variant callers have been designed by the community to identify or “call” variants through genomes. Each of these tools have been generated considering specific rules and criteria and therefore, variants detected can differ among them. For this reason, filtering the results from a set of callers and combining them can improve the identification of variants and remove false calls. A brief description of the tools used for analyzing somatic structural variants and the strategy designed for merging the results is explained in the following sections.

4.1.3.1 Variant caller programs

Somatic structural variants were not identified for the six patients lacking the germline sample but only for those 13 patients we could compare tumor versus normal genomes. In patients who underwent allogenic stem-cell transplant (case 1523 and 4675) tumor versus patient’s germline and tumor versus donor’s germline variant calling were performed. For these patients, we only considered those variants that intersected between both analyses. Variant callers were run by Romina Royo from the INB Computational Node 2 group at Barcelona Supercomputing.

Four different variant caller programs were used to extract somatic structural variants, including SMuFin (v.0.9.4), BRASS (v.6.0.5), SvABA (v.7.0.2) and DELLY2 (v.0.8.1).

- SMuFin (Somatic MUtation FINder) (183) is a reference-free method able to detect somatic variants including multiple types by comparing tumor samples with their matched normal samples.
- BRASS (184) examines paired-end sequencing reads marked as improperly paired to identify rearrangement breakpoints by clustering their mapped locations and performing an assembly.
- SvABA (structural variation analysis by assembly) (185) also performs local assembly to create groups of sequence reads that deviate from the reference genome including unmapped or discordant reads and compares them to the reference to annotate SVs and indels.
- DELLY2 (186) works as a prediction method based on read-depth, paired-end and split-read information to discover all kinds of structural variants (deletions, tandem duplications, inversions and translocations). Diverse optional parameters were modified when running DELLY2. We allowed 5% of tumor contamination in normal (-c 0'05) and at least 5% of alternate reads in the tumor sample (-a 0'05). Moreover, the minimum size for deletions and insertions was 15bp (-m 15) and 400bp (-m 400) for inversions, intrachromosomal translocations and duplications.

4.1.3.2 Variant validation through manual inspection of aligned sequencing reads

Results obtained from each variant caller algorithm were manually inspected to determine whether they were true somatic structural variants or not. Variant validation was also used to define the parameters and criteria strategy to merge SVs identified by multiple variant callers.

Before defining our merging strategy, we not only inspect sequencing reads of patients included in this Richter study (case 63, 365 and 1669) but also CLL patients (cases 16, 48, 64, 373 and 853) analyzed by Puente et al. in 2015. These five CLL cases were finally not included in this Richter's research. From the set of published structural variants (187) identified in CLL samples, we manually inspected 35 SVs counting 15 experimentally validated.

Variant validation was done based on aligned tumor and normal sequencing reads. To manually inspect the sequences, we used Samtools (v. 1.5) view mode. Using BAM files, we searched for two different supporting read categories including paired-end and split reads.

Considering structural variants are formed by two breakend regions involving one (insertions, deletions, inversions, duplications or intrachromosomal translocations) or two chromosomes (interchromosomal translocations), supporting paired-end reads were those where each paired read was aligned within one breakend region. Moreover, the observed insert size between them differed from the expected (around 300bp) (Fig 20.A). On the other hand, split reads were broken, and some nucleotides aligned through one breakend whereas the remaining ones correspond to the second breakend location (Fig 20. B-C). If needed, reads observed across the variant region were realigned using web Blastn (nucleotide Basic local alignment search tool) (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) and the human reference genome (GRCh37).

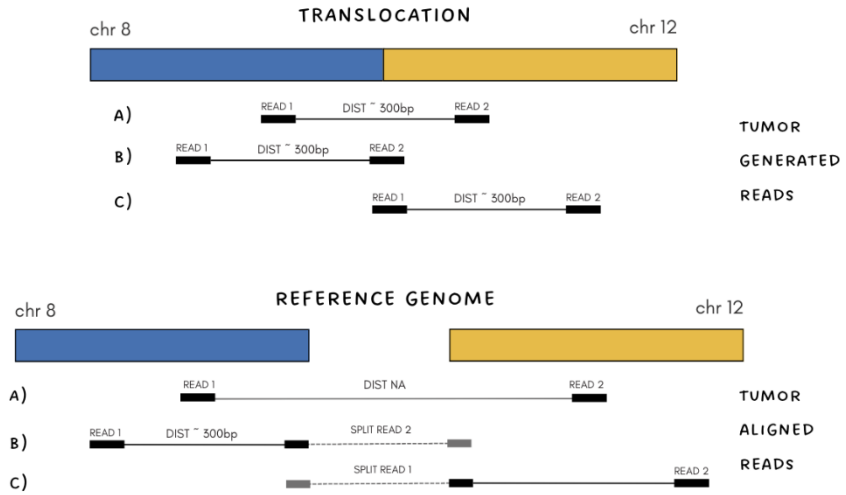


Figure 20. Representation of tumor generated (up) and aligned (down) reads supporting a translocation. A) Paired-end read where each mate read is aligned in one chromosome. B and C) Reads generated at the breakend position are splitted after the alignment.

For each manually validated structural variant, we counted the number of supporting PE (paired-end) and split reads. We then analyzed whether the variant was detected or not and if it was considered good or low quality by the variant callers.

4.1.3.3 Filtering, merging and consensus variant calling results

To end with a list of somatic structural variants for each tumor sample based on a multi-variant calling approach, the results obtained for the mentioned algorithms were filtered and merged. Through manual validation of identified structural variants we defined the final criteria to select consensus and conservative data set of variants describing the somatic landscape of tumor samples.

First, structural variants shorter than 100bp were removed from this analysis and labeled as indels (small insertions and deletions). As the detection of the breakend position of SVs is less precise and algorithms are usually not able to

determine them with base pair resolution, we intersected variants considering a window of 300 bp around break points. However, we only kept for downstream analysis those SVs detected by at least two programs if a minimum of one algorithm called the variant with high quality (MAPQ \geq 90 for BRASS, MAPQ = 60 for SvABA and DELLY2). We use Integrative Genome Viewer (IGV) to visually inspect all structural variants.

4.1.3.4 Rescue of somatic structural variants from longitudinal samples

Based on the information obtained from longitudinal samples of the same patient, we rescue genomic alterations. Therefore, we could increase the number of detected somatic SVs.

Those structural variants identified in one sample after the filtering and merging step, were automatically added in the additional time point(s) of the same patient if any of the variant callers detected the variant, independently of the filters.

After all these steps, we ended with a list of conservative somatic structural variants for each tumor sample included in the Richter's cohort. Somatic structural variants identified in case 63 were used to continue with our study of subclonality in Richter's transformation.

4.1.4 Inferring structural variant allele frequencies to analyze intratumor heterogeneity

The intratumor heterogeneity of a sample is characterized based on the variant allele frequency of somatic tumor mutations. Variants of similar frequencies are clustered together representing a specific cell population. Moreover, the analysis of samples collected at different time points allows us to reconstruct how these cell populations evolve during time and therefore, to observe clonal dynamics of the tumor. Variant allele frequency is calculated using aligned sequences and dividing the number of mutated reads by the total number of reads covering the mutated position. It is usually calculated for single nucleotide variants, or small insertions or deletions, but not for structural variants including large indels, inversions, duplications, intra- and interchromosomal rearrangements. This is due to the complexity of identifying all supporting reads aligned through the reference genome and the variability of the coverage within these large, mutated regions. For this reason, we first explored both supporting reads and coverage variability among diverse selected structural variants and on different CLL samples, including cases 63, and 365 from the Richter's study to understand the nature of the region. Coverage distribution in healthy and tumor samples from CLL cases (29, 48 and 723) finally excluded for the publication, was also evaluated.

Intending to design a strategy to calculate the variant allele frequency of somatic structural variants identified in Chronic Lymphocytic Leukemia patients, we started exploring somatic SV within an *in-silico* tumor sample. Then, we applied this strategy to SVs identified in the CLL cohort and we focused our analysis on case 63 using it as a pilot.

4.1.4.1 Analysis of aligned tumor reads in an *in-silico* sample

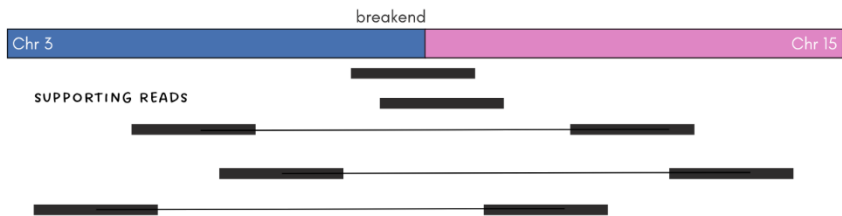
As structural variants involve large genomic regions, reads covering the variant are usually challenging to align through the reference genome by the algorithms. Therefore, few supporting reads can be unmapped and not detectable on the BAM files. This loss of supporting reads directly influences the obtained variant allele frequency.

So as to explore if all supporting reads are usually aligned and consequently, can be identified from the BAM file, we started analyzing somatic structural variants within an *in-silico* WGS sample (Fig 21.1). This artificial sequenced sample was created by Dr. Jordi Valls in the context of his PhD thesis and in our group. To generate a sample simulating a real genome sequence, human variants from the 1000 Genomes Project ADD REF and the PanCancer project were inserted. All the artificial reads supporting each of these variants were known and searchable in both FASTQ and BAM files created for this *in-silico* sample.

Manual inspection of *in-silico* SVs including a deletion (chr3:173048887-chr3:173050455), and one inversion (chr20:53484361-chr20:53485620), both bigger than 1000 bp, and one interchromosomal translocation (chr21:18877844-chrX:131913425) was done to determine if all supporting reads were aligned. To do this, we used Samtools (v.1.5) view mode to export and analyze aligned sequencing reads. If needed, reads were realigned using the BLAT (Basic local alignment tool) from the UCSC (<https://genome.ucsc.edu/cgi-bin/hgBlat>) and to the GRCh37 reference genome.

We also explored mutated regions and a range of different window sizes (10bp, 50bp, 100bp and 150bp) from each breakend (Fig 21.2) of the structural variants to determine where supporting reads were aligned.

1) To determine if all supporting reads were aligned



2) To define the region where supporting reads are aligned

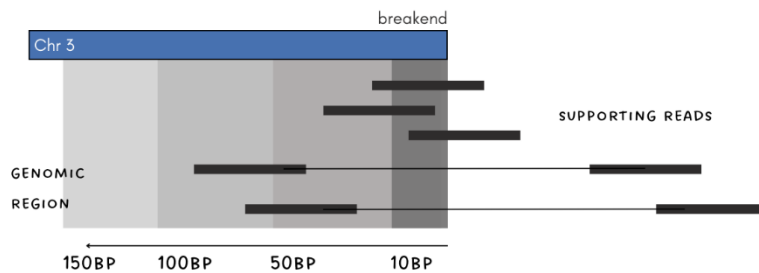


Figure 21. Schema showing the evaluation of *in-silico* structural variants. 1) A determined genomic region including the SV is analysed to search for aligned supporting reads. 2) Different windows sizes are defined to identify the region where supporting reads are usually aligned.

4.1.4.2 Calculating the variant allele frequency for *in-silico* structural variants to define a strategy

All somatic SVs within the artificial *in-silico* sample were heterozygous, clonal and not within copy number variants, thus their expected variant allele frequency was around 0,5. That means half of the aligned reads should support the mutated allele, whereas half covered the non-mutated allele. For this reason, the *in-silico* sample was also used to define the strategy to calculate the variant allele frequency of structural variants. Results obtained were compared with the expected VAF (0,5) to adjust the strategy.

To calculate the variant allele frequency of each structural variant, we counted mutated reads and the total number of reads covering each position within a region of 100bp from each breakend position (up or downstream depending on the SV type) (Fig 22). Sequencing reads including mutated and non-mutated were extracted from the BAM file using Samtools (v. 1.5). Mutated reads included paired-end where each mate aligned within a breakend and do not have the expected insert size (~300bp) and split reads, defined as broken reads aligned through each breakend. Moreover, the number of mutated reads was corrected by adding one more read in all the base pair positions where part of a split read should be aligned, even if it does not directly appear on the BAM file. We then calculated the VAF of each breakend (two per SV), dividing the mean of mutated reads in 100bp by the mean of the total number of reads covering the same region.

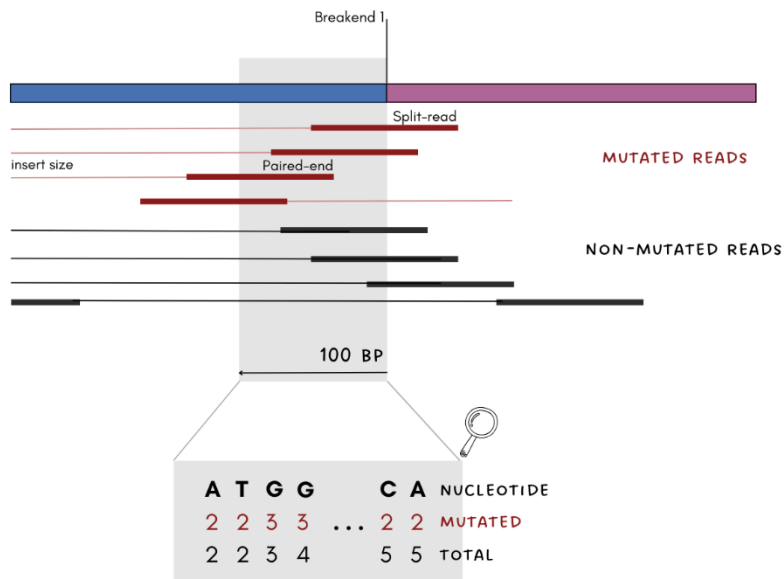


Figure 22. Schema of the strategy applied to calculate the variant allele frequency of a breakpoint based on aligned reads.

Following the described strategy, we analyzed the variant allele frequencies of a set of structural variants present in the *in-silico* sample. This includes deletions (n=45), intrachromosomal rearrangements (n=97) and inversions (n=20) longer than 1000bp.

4.1.4.3 Applying the designed strategy to CLL longitudinal samples

Following the strategy designed using the *in-silico* sample and mentioned on the previous section, we calculated the variant allele frequency for the somatic structural variants identified in CLL samples, including variants within the same chromosome larger than 1000bp and interchromosomal variants. For this analysis, to avoid an automatic misidentification of mutated reads, we removed those structural variants identified in a sample that clustered together or nearby (considering a windows size of 100 bp).

As for the *in-silico* sample, for each structural variant we calculated two frequencies representing both breakends. We then compared the variant allele frequencies of breakends corresponding to the same variant to analyze whether they were similar or not and how to adjust them.

4.1.4.4 Deducing cancer cell fraction of structural variants and clonal dynamics for one pilot CLL case

To study intratumor heterogeneity and the evolution of different cell populations coexisting in one tumor sample, the variant allele frequency is translated into a cancer cell fraction. This value represents the fraction of tumor cells where the somatic variant is present. Similar CCF are clustered together to represent cell populations. For this reason, to continue exploring intratumor heterogeneity in CLL samples through somatic structural variants, the VAFs obtained from the pilot CLL case (63) were translated into cancer cell fractions.

Variant allele frequencies are converted into cancer cell fractions following the equation (188):

$$CCF = VAF * 1/p * (p * Ntot_t + (1-p) * Ntot_n),$$

where p represents the purity of the tumor sample, meaning the fraction of tumor cells within the sample, and $Ntot$ is the number of chromosome copies in tumor cells ($Ntot_t$) and in normal cells ($Ntot_n$) at the mutation locus. Usually, $Ntot_n$ is 2 considering no copy number variation has occurred on normal cells. In those cases, a breakend was identified within a somatic copy number variant, the number of chromosome copies in tumor cells ($Ntot_t$) was calculated from the CNVs previously identified by variant callers and defined as the mean of total number of alleles in each position within the 150bp mutated region.

We calculated the cancer cell fraction for each breakend identified in case 63 separately to avoid variability due to CNVs affecting just one region of the SV. Finally, the CCF for somatic structural variants of this CLL case were obtained from the mean of the CCFs of each breakend. The values obtained were compared between longitudinal samples of the same patient (63) to observe the evolution of somatic structural variants during time.

Results of chapter 1 starting in section 5.1 (page 161).

4.2. Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact

Chapter 2

Within this chapter, and in the context of the Pan-Cancer analysis of Whole Genomes, we analyzed more than 2.000 tumors and their matched normal genomes to identify processed pseudogenes acquired somatically and explored their potential functional impact in tumors. This work was published in Nature Genetics in 2020, within a larger study of retrotransposition in cancer.

4.2.1 Genomic and transcriptomic cancer data

In order to study the landscape of somatic processed pseudogenes in cancer genomes, we used the PCAWG international cohort. From a set of 40 different tumor types and subtypes, six were removed because of having less than 19 donors. In total, we explored 2589 donors distributed on 34 tumor types and subtypes (Fig 23).

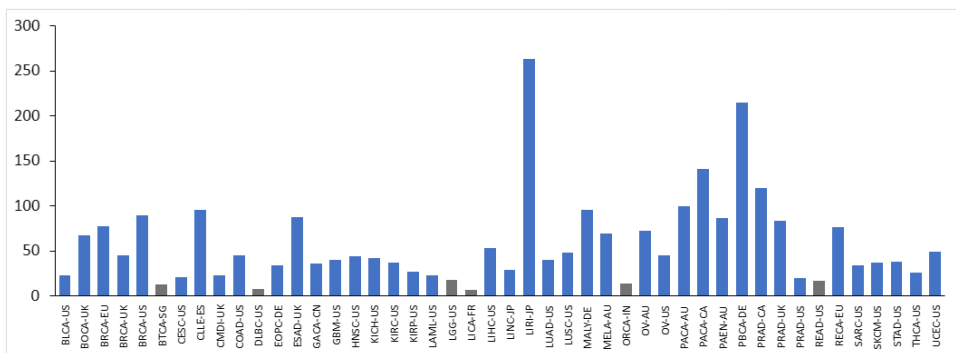


Figure 23. Number of donors (y axis) in each PCAWG project, including 34 different tumor types for diverse countries. Bars colored in grey correspond to discarded sets of genomes.

For the identification of somatic processed pseudogenes, we analyzed tumor-normal pairs. The normal sample for each donor was essential to identify somatic events, since they are only present in the tumor genome but not in its normal mate. Therefore, we downloaded for each tumor-normal pair, whole genome sequences formerly aligned using the GRCh37/hg19 reference genome. This data was downloaded in BAM format files. We also used the PCAWG catalog of somatic structural variation. This catalog was previously obtained by the consortium after applying the three official PCAWG variant calling pipelines (Sanger, Broad and DKFZ) and merged the results into VCF (Variant Caller Format) files.

After the identification of somatic PPs, we look for expression signals by interrogating tumor RNA-seq data available for 144 samples containing the event. RNA aligned reads in BAM files were downloaded from the PCAWG cohort.

4.2.2 Somatic processed pseudogenes identification

Due to the lack of standard protocols for the identification of somatic PPs, we first explored different bioinformatic strategies with one donor with the aim of generating an automatic protocol that could be extended to all PCAWG samples.

We based our examination on recursive steps combining automatic searches for somatic structural variants through VCF files that could point to PPs, with manual inspection of the results by evaluating aligned tumor and normal reads from the same donor. Using this approach, we came up with a combination of some basic rules that provided candidate PPs. This set was then validated manually resulting in a more restrictive list of somatic PPs.

Although the data analysis to identify somatic processed pseudogenes was the same in any case, one or multiple donors, the workflow varied in order to analyze 2589 donors automatically. Some statements were added, and the final criteria was defined once the results in pilot candidates were observed. A description of the genomic data analysis and the generation of the automatic protocol is given below.

4.2.2.1 Genomic data analysis

4.2.2.1.1 Candidate PP selection through VCF files

We adapted the protocol described by Cooke et al. (110) to identify somatically acquired processed pseudogenes. As PPs are the result of the reverse transcription and integration of an mRNA, the absence of introns and the presence of exonic sequences in unexpected locations were used as main characteristics to define an identification strategy. The combination of both features was necessary to avoid misclassification of somatic events, including translocations or deletions involving genes. Also, to prevent signals derived from mRNA contamination that can be found in DNA samples.

Based on these criteria, we expected to see on the somatic structural variants VCF files, a) mutations joining an exon of this gene and any other part of the genome, the insertion, and b) point mutations denoting exon-exon junctions within the same gene, the candidate pseudogene. This is summarized in figure 24.

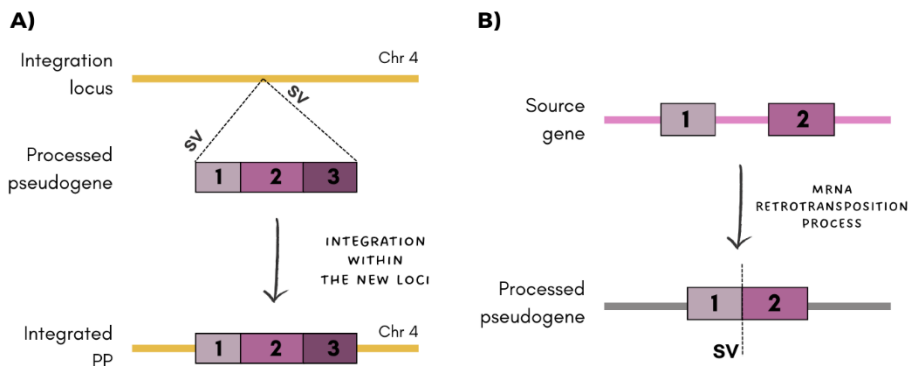


Figure 24. Structural variants representing the main characteristics of PPs. A) Dotted lines indicate two SVs pointing to an insertion of a PP (pink) in chromosome 4 (yellow) resulting in the somatic fusion of a DNA within another genomic region. B) SV pointing to a splicing event. Dotted line (SV) representing the deletion/splicing on an intron as a result of the reverse transcription of an mRNA.

Considering each structural variant is formed by two breakpoint positions, we annotate both genomic locations, +/- 100 bp because of unprecise given coordinates, using the RefSeq gene database (GRCh37/hg19) (Fig 25) Structural variants where none of the immediate flanking breakpoints mapped on an exon were removed as they do not represent any of the mentioned PP features.

This SVs annotation provided us with information to select candidate pseudogenes and to continue with the evaluation.

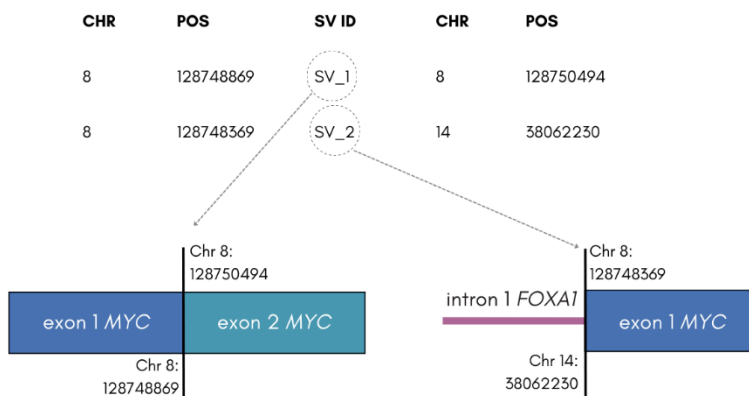


Figure 25. Annotation of structural variants from a VCF file. Flanking positions of each breakend mapped to the reference genome in order to identify exon-exon breakends (left) on the same and exon-new loci breakend (right). In this example, MYC is the candidate pseudogene and appears inserted within intron 1 of FOXA1.

4.2.2.1.2 Manual validation: inspection of tumor sequencing reads

The low reliability of automatic rules derives from the number of breaks not included in the official PCAWG VCF files because of their doubtful identification, as well as the number of false breakpoints obtained from these algorithms at the time we were using this data. Moreover, precise genomic coordinates for the structural variants are not consistently given since it is a challenge to define them. For this reason, candidate pseudogenes selected from the VCF file were confirmed with manual inspection of the tumor genome. BAM files including genome sequencing reads were visualized using Samtools (v.1.5).

On top of the main features used (absence of introns and evidence of insertion into new loci), two reads-based conditions were evaluated on the tumor genome: i) paired-end reads and ii) split reads. A description of how these conditions were used is explained below.

- i) Paired-end reads. As explained before, the term is used when both ends of the DNA fragment are sequenced and distance between them (i.e. insert size) is known. In WGS data used for this study, the insert size was around 300 bp.

PE reads where one end maps to an exon of the candidate pseudogene and its mate into the new integration loci support the insertion of the PP. PE reads mapping, each, a different exon of the candidate pseudogene with an insert size larger than expected, highlight splicing events on the source gene fig 26. We rely on the fact that these exons will be together in the tumor genome, as part of the same DNA fragment. However, they are aligned to the reference genome with larger distances than expected because intronic sequence separation. We could only identify these second

PE reads when the size of the introns in the source gene was large enough.

- ii) Split reads are sequences that break when mapping to the reference genome. That is when only some bases of the read map somewhere on the reference genome, whereas the remaining nucleotides are unmapped. Split reads add further evidence to SVs events and usually, when realigned provide the precise coordinates for each breakpoint. Therefore, both the insertion site within a new loci and exon-exon junctions of the PP can be observed accurately if the two halves of the split read align across the structural variant fig 26.

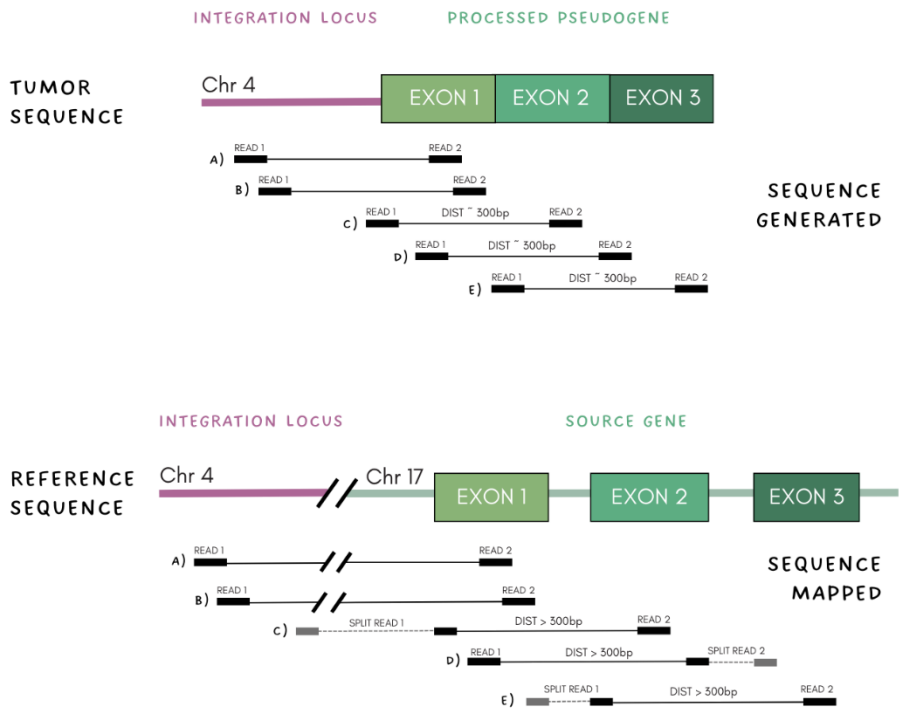


Figure 26., Tumor sequencing reads supporting PP formation. A , B) Paired-end reads supporting the insertion site. One read aligned in the insertion loci and its mate within an exon of the source gene. C) Split read mapping the insertion point. D, E) Split reads mapping splice junctions of the same transcript (source gene). When these reads (read 2 D and read 1 E) are aligned to the reference genome, sequences are broken, as these exons are separated by an intron sequence in the reference genome. Distance between these PE is larger than the expected (insert size 300bp) as on the tumor sequence these pairs are closer than on the reference sequence.

To evaluate both paired-end reads and split reads, aligned sequences from the tumor BAM file were obtained using Samtools. We mainly extracted reads aligned (+/-300bp) through the candidate insertion site as well as reads within the genomic coordinates of the source gene. Reads were then aligned to the reference genome (GRCh37/hg19) using UCSC Blat (189) (default parameters) to validate the insertion site, and to the reference RefSeq transcriptome using Blastn (190) (default parameters) to confirm splice junctions.

Manual inspection of the tumor sequence verifies structural variants pointing to candidate processed pseudogenes formation are real. Moreover, this evaluation allows us to observe the presence of poly A tails, characteristic at the 3' end of the PP sequence.

Finally, reads supporting the somatic variation were also evaluated in the matched normal genome, to confirm their absence and therefore define the event as somatic.

4.2.2.2 Generation of an automatic protocol

4.2.2.2.1 Pilot exploration of one candidate PP

Before developing a protocol to identify somatic processed pseudogenes in all PCAWG cohort at once, one donor was explored to define and calibrate our strategy. We applied the genomic data analysis explained above, implementing both steps. This also allowed us to understand the characteristics of processed pseudogenes.

From the PCAWG cohort, we randomly selected one sample (submitter donor id: 9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca) from 48 patients included on the lung squamous cell carcinoma (LUSC) subcohort. We decided to start with this group considering that other previous studies confirm the highest number of PPs somatically acquired on this tumor type-subtype (110).

First, the somatic structural variant landscape of the patient was analyzed retaining those SVs with at least one breakpoint position (+/- 100bp) corresponding to an exon. Considering we expected to observe SVs supporting the insertion point and the absence of introns, *CNIH4* was selected as the source gene producing a processed pseudogene in this tumor sample. We relied in *CNIH4* as a candidate PP since multiple structural variants involving exons from this gene were identified.

We decided to reconstruct *CNIH4* PPs insertion using directly the positions provided by the structural variants identified and their corresponding location on the gene using the human RefSeq database (GRCh37/hg19). Next, we inspected its corresponding tumor BAM file to verify the automatic VCF-based predictions. We added the data obtained from tumor PE and split reads to the manual PP reconstruction.

To further verify this processed pseudogene was acquired somatically, we looked for these reads on the normal pair genome.

4.2.2.2 Protocol development for the complete analysis

To scale up our search for somatic PPs to all 2589 PCAWG tumor-normal genome pairs, different workflows were tested. In this step of the study, we used the entire subcohort from lung squamous cell carcinoma (LUSC) encompassing 48 donors. Our idea was to define a protocol as automatic as possible based on two types of somatic structural variants (exon-exon and exon-new loci). We selected candidate pseudogenes coupling this type of data, obtaining diverse datasets with different levels of sensitivity. We applied in-house scripts to obtain these datasets combining the following criteria:

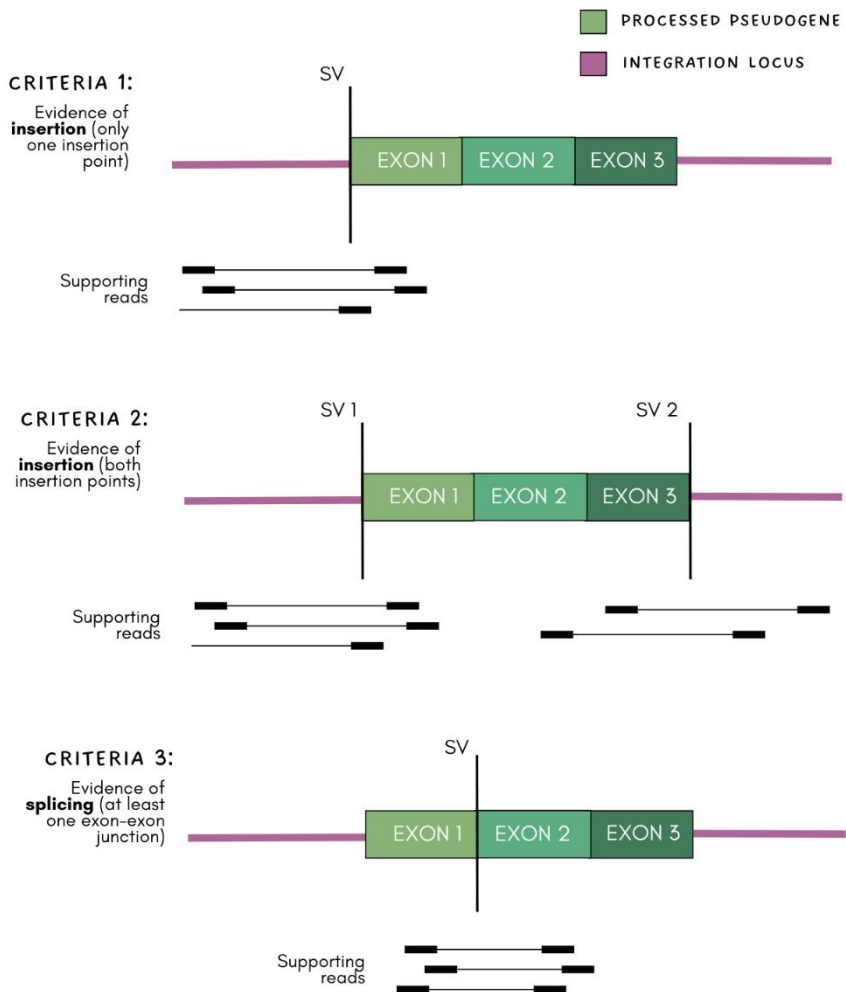
Criteria 1 – Evidence of insertion: at least one structural variant between an exon and any other region of the genome, representing one insertion site of the PP. The insertion locus could be an exon of another gene.

Criteria 2 – Evidence of insertion: two structural variants joining the same source transcript and the same integration locus. It differs from criteria 1 since here we require evidence for both insertion sites.

Criteria 3 – Evidence of splicing: at least one structural variant involving two exons of the source gene, likely indicating a minimum of one splicing event.

These criteria were combined to generate four different datasets: Dataset 1: criteria 1, Dataset 2: criteria 1 and 3, Dataset 3: criteria 2 and Dataset 4: criteria 2 and 3. We count as a candidate pseudogene, each time a source transcript was detected inserted on a chromosome. Therefore, if the same source transcript was identified inserted in, for example, chromosomes 7 and 19, we considered them as two candidate pseudogenes. If the same transcript appeared on many SVs always inserted within the same chromosome, only one candidate pseudogene was counted. Figure 27 summarizes criteria and datasets generated.

For each candidate set, we manually evaluated a subsample to determine the type and rate of false positives included and thus, the level of accuracy of each of the criteria used. We applied new specific rules while analyzing the results from this manual validation, until we defined the final identification criteria.



DATASET	CRITERIA
Dataset 1	criteria 1
Dataset 2	criteria 1 + criteria 3
Dataset 3	criteria 2
Dataset 4	criteria 2 + criteria 3

Figure 2723. Visual representation of the criteria used to define candidate PPs and its combination to generate datasets with different levels of sensitivity. Criteria were defined based on the two types of SVs we expected to detect when a PP was acquired.

4.2.2.2.3 Final PP searching strategy

From the previous analysis we could generate different collection of criteria that provides the diverse set of candidate somatic processed pseudogenes, as well as a validated collection. Results obtained for each criteria combination are explained in the Results section 5.2.2. Considering these observations, we ended up with the final PP searching strategy. Then, we applied it to the entire PCAWG cohort cited before, using its catalog of somatic structural variation.

As PPs are the result of reverse transcription and integration of an mRNA, the identification of the presence of exonic sequences in unexpected locations within the tumor genome was used as the main criteria to define our final searching strategy. Structural variants supporting the insertion of a PP were defined to be flanking (+/- 100 bp) an exon sequence on one side (defined using the NCBI RefSeq gene coordinates; GRCh37/hg19) and any genomic region on the other. To avoid events such as intrachromosomal translocations or deletions involving other genes, structural variants affecting the same chromosome with a distance between both breakpoints lower than 100Kb were excluded. We considered insertions of the source gene less likely to occur near its location.

Finally, we labeled as candidate PPs those integrations supported by both insertion points with a distance flanking these sites of less than 350bp. Moreover, to avoid nucleotide insertions, at least 50 bp of the same source gene exon, the candidate PP, must be inserted (Fig 28).

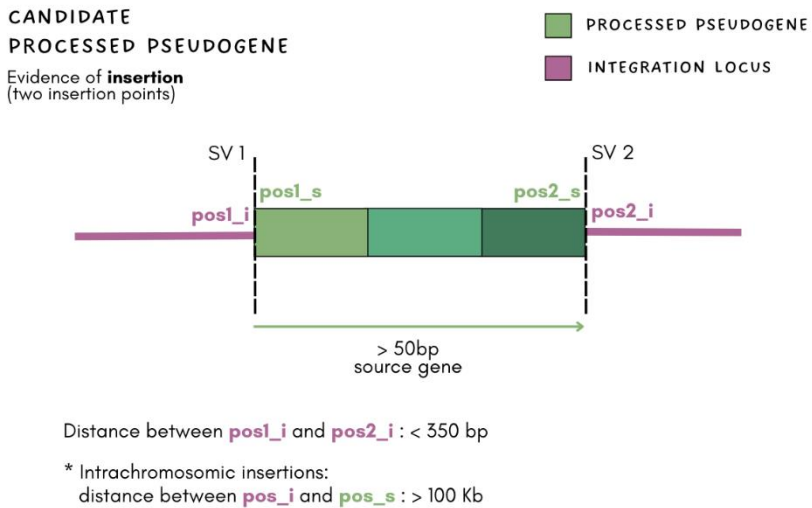


Figure28. Schema of the strategy applied to identify candidate processed pseudogenes. Two somatic structural variants (SV1 and SV2) must be identified on the VCF file representing both insertion points with a distance flanking these sites (*pos1_i* and *pos2_i*) shorter than 350bp. At least 50 bp of the same source exon must be inserted (green). Events affecting the same chromosome (intrachromosomal) must be inserted more than 100Kb farther than the source gene coordinates.

Candidate pseudogenes identified using the criteria mentioned above were evaluated through manual inspection of the tumor sequence. This resulted in a validated collection of somatic processed pseudogenes. We considered a candidate as validated, if PE or split reads covering both insertion sites (insertion loci - source gene, source gene – insertion loci) were identified on the tumor genome. Manual inspection allowed us to also identify sequencing reads supporting splice junction sites. However, evidence of splicing was not as determinant criteria since it could derive from mRNA contamination of genome samples.

Although our identification strategy was based on the somatic structural variant landscape, manual inspection was also done for the normal sequence of the donor, to confirm the PP was acquired during tumor development.

4.2.3 Expression evaluation of acquired PPs

Evidence of chimeric RNA processed pseudogene – receptor loci were required to distinguish expression of the PP from the transcription signals of native mRNAs derived from the source gene. To do so, we interrogated RNA-seq data, if available for the donor that have acquired the PP, involving retrocopies with part of the integration region.

For each candidate PP, we extracted sequencing reads from the BAM file aligned to the source gene coordinates. We also selected reads aligned to the receptor gene or to the intergenic sequence (\pm 5Kb from the insertion site) and unmapped reads. All these selected reads were used as query to perform two independent Blastn analysis. On one side, reads were aligned against a database with all cDNA transcript forms described in RefSeq (NCBI) database corresponding to the retrotransposed mRNA sequence of the source gene. The second alignment was done against the complete reference DNA sequence of the receptor gene or the genomic region 5Kb upstream and downstream the integration site for those candidates inserted within an intergenic sequence (Fig 29).

Positive expression of the candidates and/or the formation of fusion transcripts was confirmed with at least two paired-end reads. In each pair, one end must align to the cDNA sequence of the source gene, and its mate into the DNA sequence of the integration loci. Both with more than 98% identity. Identification of split reads across one insertion site was also considered as supporting evidence of expression.

Expression signals were used to predict and manually reconstruct the resulting fusion PP-host gene transcript. Also, to infer the fusion gene coding

potential through *in-silico* translation starting from the initial codon (ATG) for each of the host gene mRNAs.

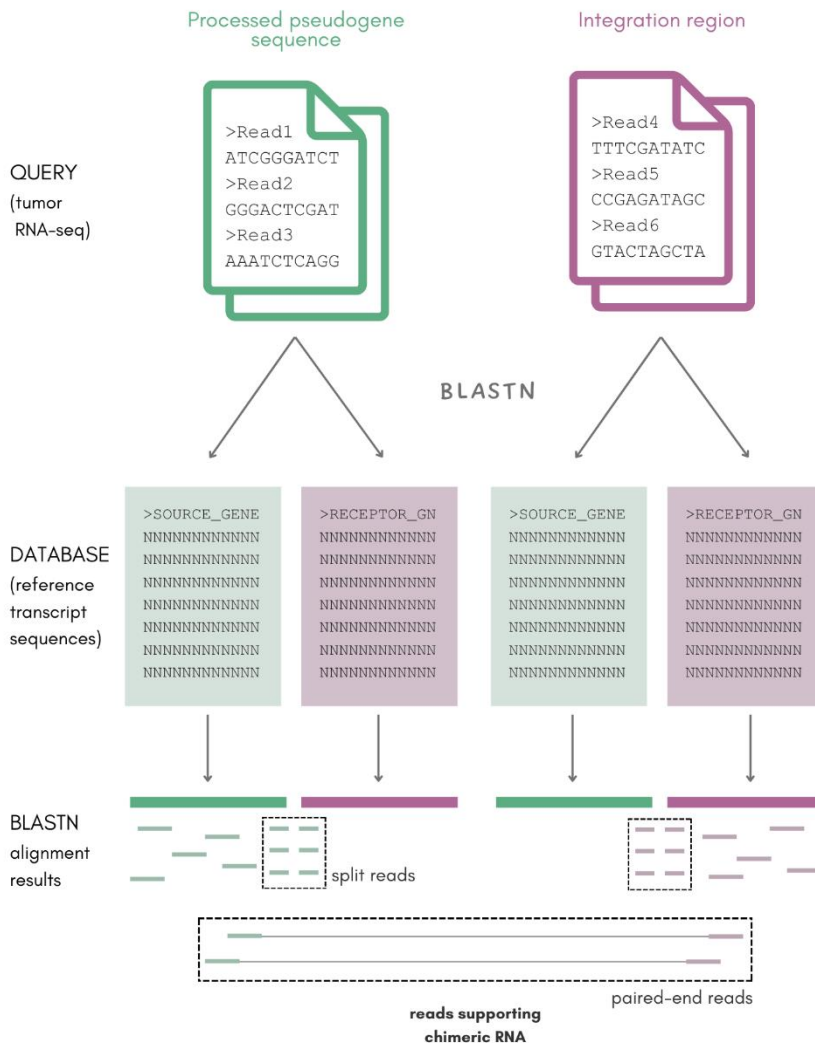


Figure 29. Evaluation of positive somatic processed pseudogene expression. RNA-seq reads were extracted from the source gene (query green) and integration region (query pink). All these sequencing reads were aligned (Blastn) against reference cDNA corresponding to the retrotransposed mRNA (database green) and reference DNA of the integration region (database pink). Aligned PE and split reads within dashed rectangles represent sequences supporting a fusion PP-host gene transcript confirming the expression of the somatic PP.

Results of chapter 2 starting in section 5.2 (page 185).

4.3. Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data

Chapter 3

Study 1:

In this first study included in the third chapter of the present thesis, we describe the results obtained from our efforts in the context of the identification and characterization of new functional micropeptides in human and their application in mass spectrometry studies of cancer micropeptides. This work was in collaboration with Dra. Maria Abad and Marion Martínez from VHIO and Dr. Javier Muñoz, Dr. Hector Peinado and Pilar Ximénez de Embún from CNIO.

Study 2:

The knowledge acquired from this first study opened the possibility of searching for unexplored micropeptides and determine their relationship with tumorigenesis. Therefore, this second study is focused on the identification of novel candidate micropeptides located in intergenic regions, using comparative genomics and evolutionary conservation features at DNA and protein level. We search for evidence of expression in healthy tissues and evaluate their role in cancer by searching for clusters of mutations within them.

Study 1: Catalog of candidate micropeptides for MS/MS searches

4.3.1 Transcriptomic data from pancreatic adenocarcinoma

Transcription and translation are tissue-specific biological processes meaning different phenotypes are generated from the same genome sequence among tissues. Therefore, tissues are distinguished by gene expression patterns, resulting in distinct regulatory programs controlling the function of each specific tissue type. These processes can also vary between normal and tumor cells. Since the presented study is focused on the identification of micropeptides in pancreatic adenocarcinoma through mass spectrometry, we based our search of novel candidate small ORFs in the transcriptomic analysis of the same tumor type.

We randomly selected six pancreatic adenocarcinoma adult patients (3 female, 3 male) from The International Cancer Genome Consortium. All of them were provided by the Australian project PACA-AU. For each patient, we used BAM files of aligned RNA-seq samples (126 bp length, paired-end reads). Sequencing reads were previously aligned by the consortium with STAR (v.2.4.0i, <https://github.com/alexdobin/STAR>), using GRCh37 reference genome (<https://github.com/ICGC-TCGA-PanCancer/pcawg3-rnaseq-align-star>).

4.3.2 De novo transcriptome assembly

To end with a set of novel candidate micropeptides within non-annotated transcripts, de novo transcriptome assembly was done for the 6 pancreatic adenocarcinoma patients. The assembly of RNA-seq reads without a sequenced genome to guide can, in theory, reconstruct transcripts even from regions missing from the reference.

De novo transcriptome assembly was done using StringTie (v.1.3.6, <https://ccb.jhu.edu/software/stringtie/>), a computational method to assemble complex data sets into transcripts. Starting with aligned RNA-seq paired-end and spliced reads, StringTie groups them into clusters and creates a splice graph for each cluster. Later, the approach identifies transcripts from these clusters of reads and estimates their expression levels simultaneously (Fig 30).

Since we aim to also obtain low-expressed transcripts, the minimum number of reads per bp coverage to consider for transcript assembly was 2.5 (default parameter). Moreover, to assemble short transcripts the minimum length allowed for the predicted sequences (-m parameter) was set to 50 bp.

One more optional parameter was tested when running StringTie for the 6 RNA-seq samples. A brief description of it is given below.

- Maximum fraction of multiple-location-mapped reads that are allowed to be present at a given locus (-M). StringTie was tested with default parameter (1.0) and 0.1.

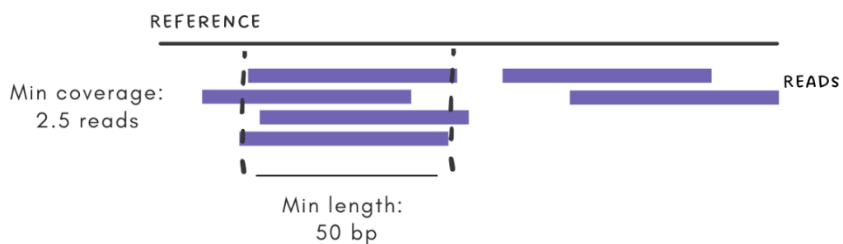


Figure 30. Schema showing StringTie identification of transcripts. A minimum coverage of 2.5 reads and a minimum length of 50bp is needed to define the clustering RNA-seq aligned reads as a transcript.

We compare the results obtained after testing both options (default and 0.1) by manual inspection and we define the best approach to continue with the identification of novel transcripts.

After running StringTie, we obtained for each patient one GTF (General Feature Format) file containing all the identified transcripts, their genomic coordinates, strand and expression values in terms of coverage. Transcripts identified in chromosome Y in the three female patients were removed since are false positives due to previous alignment errors.

4.3.3 Transcriptome combination of multiple samples analyzed

A list of transcripts was obtained for each patient analyzed. However, a consensus set of sequences detected in diverse samples was needed to continue with the identification of novel candidate micropeptides in pancreatic adenocarcinoma. The merging step will allow us to remove false positives obtained from de novo assembly because of the inclusive search of low expressed and short regions.

We explored two approaches to combine the results and get a representative set of expressed transcripts for pancreatic adenocarcinoma. We tested the merge StringTie function available for the program, as well as we defined our strategy to determine consensus transcripts.

4.3.3.1 StringTie transcript merge mode

StringTie provides a usage mode different from the assembly function to merge and assemble transcripts from diverse RNA-seq analysis and to obtain a non-redundant set of consensus and filtered sequences.

To run the merge mode, we used as inputs the six GTF files obtained from de novo assembly performed across RNA-seq samples. Even different options can be modified on this function, StringTie merge was run with default parameters.

4.3.3.2 In-house strategy to obtain a consensus set

The criteria used by StringTie merge to combine the results and afford a consensus set is not described in the documentation. Accordingly, we explored the results obtained from de novo assembly and searched for a merging strategy applying diverse criteria.

Our protocol was divided into a merging step of the transcripts identified in RNA-seq, the definition of a consensus sequence including the clustered transcripts, and the subsequent selection of representative groups.

4.3.3.2.1 Merging step through transcript clustering

To merge transcripts and isoforms identified on different samples, we first explore diverse requirements to consider two transcripts as the same one. Therefore, we started exploring the overlap between their genomic coordinates. We examined a range of window sizes (0, 150, 250, 500, 750, 1000, 1500, 1750, 2000bp) to define the best criteria for considering both start and end coordinates represent the same transcript in diverse samples. We later analyzed whether strand, and the number of exons were also necessary to consider for merging these sequences. Manual inspection of the clustered transcripts obtained depending on the applied requirement was done to define the strategy. An extended description of this decision-making process is explained in the results and discussion sections.

Finally, we merged isoforms from different samples in case they shared both start and end coordinates within a window size of 500bp, regions overlapped between them, and had the same strand and number of exons (Fig 31). Merged isoforms were outputted in a tsv (Tab-separated values) format file. For each clustered group of isoforms, we recalculated their start and end coordinates as the average among all the samples where it was identified.

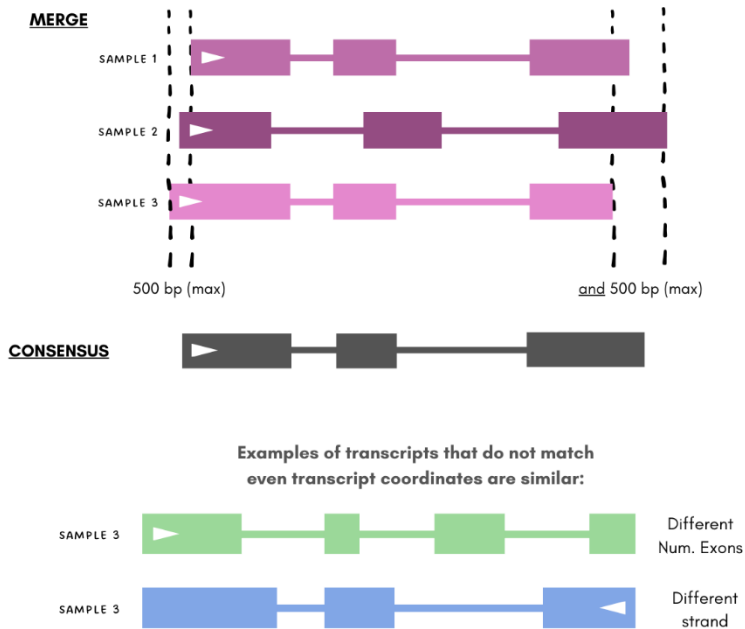


Figure 31. Schema summarizing the merging step. Transcripts detected in different samples are clustered if they share start and end coordinates (ws 500bp), and the number of exons and strand are the same. The consensus sequence obtained after considering clustered transcripts is shown in grey.

4.3.3.2.2 Definition of a consensus sequence and selection of representative transcripts

After the merging step, we aim to get a consensus sequence for each group of merged transcripts considering the variation among exon genomic coordinates detected in samples.

To define the best approach, diverse measures were proposed and studied through exploring the sequences and the manner their exons overlapped. After this, start and end transcripts were calculated as the mean between all clustered transcripts. Exon coordinates were defined based on those more represented within samples, or randomly selected if all of them appear the same number of times (Fig 31). We automatically validated all the coordinates were continuous.

For some cases, manual inspection was needed to adjust and get the consensus sequence.

Due to this recalculation of all genomic coordinates, we removed redundancy if two or more consensus sequences shared their start and end coordinates, strand, number and exon coordinates and were identified in the same patients. Consensus sequences sharing only their start and end coordinates, but not the strand, the number or the coordinates of their exons, were considered different isoforms of a transcript.

Those merged isoforms identified in at least two different samples were selected to remove false positives from the analysis and were defined as representative for the pancreatic adenocarcinoma transcriptome. Single-exon isoforms must be present in all samples to be more restrictive because they are easily detected by StringTie.

4.3.4 *In-silico* 3-frames translation of de novo consensus transcripts

We continued the analysis of small ORFs identified in pancreatic adenocarcinoma patients translating consensus sequences to obtain a set of candidate short amino acid sequences.

First, the DNA sequence (GRCh37) of each consensus transcript was downloaded from the REST (REpresentational State Transfer) API data interface of UCSC. This interface allowed us to get all nucleotide sequences from start to end bp coordinates through a command line and in a JSON (JavaScript Object Notation) format file. An in-house script was run based on the identified exon coordinates for each isoform to get only their coding sequence (CDS). Whenever the strand was not characterized by StringTie due to the lack of split reads across the region, the coding sequence in both forward and reverse was considered.

With the aim of obtaining a list of potential micropeptides, i.e. open reading frames with a maximum length of 100 aa, all the CDS were translated *in-silico*. We performed 3-frames translation for each CDS, meaning codons were defined starting from the first, second and third nucleotide of our candidate genomic sequences. We not only considered the canonical start codon (ATG) as the origin of translation but also the five most abundant non-canonical codons (CTG, GTG, TTG, ACG and ATT) (132). After an origin was found, translation was extended until the first stop codon (Fig 32). Sequences between 7 and 100 amino acids (both included) were characterized as candidate micropeptides identified in pancreatic adenocarcinoma transcriptomes.



Figure 32. *In-silico* translation of coding sequences. Micropeptides corresponded to sequences between 7 and 100 amino acids length.

Additionally, to analyze the type of genomic region where we identified candidate micropeptides, we annotate their location in the reference genome. To do so, we downloaded from the Biomart data mining tool (<https://grch37.ensembl.org/biomart/>) the genetic coordinates of all the human annotated genes (GRCh37), including UTR regions, exon and introns and non-coding regions. We then locate separately and through an in-house script the start and end positions of each candidate smORF, and therefore the translated candidate micropeptide. Each position was labeled as CDS Exon, No-CDS Exon, Intron, 3' UTR, 5' UTR or NA if it was identified within a non-annotated region. Since diverse isoforms can overlap between them on the reference genome, we

prioritize the region categories in the mentioned order. As an example, if one genomic coordinate was present in the 3' UTR of the isoform A, and within a coding exon of the isoform B, we labeled it as CDS Exon.

4.3.5 Local alignment search to remove overlap with annotated CDS

Because we aim to identify novel candidate micropeptides present in non-annotated transcripts we filter out micropeptides translated from nucleotide sequences overlapping with annotated CDS.

To do so, we first performed a local alignment search using Blastn (v. 2.6.0, <https://blast.ncbi.nlm.nih.gov/>). We intended to remove micropeptides located within known and annotated coding sequences, but not within UTRs, introns, non-coding genes or intergenic regions. Therefore, we compared candidate smORFs nucleotide sequences (query) together with all the human coding sequences (database) from Ensembl (GRCh37). We applied default parameters except for the word-size. This value represents the minimum length to find and give a perfect sequence match. We selected a length of 7, since our query sequences corresponded to small ORFs (7-100 aa) and in order not to lose reliable hits. Moreover, to limit the search and provide a more efficient analysis, the maximum number of target sequences (-max-target-seqs) was limited to 2.

Considering the results obtained from the local alignment, candidate smORFs were filtered and as consequence, candidate micropeptides. We defined as good local alignments those results from the Blastn with an e-value lower than 0'001. Small ORFs were finally selected based on the percentage of sequence overlap (<30% or <60%) between their nucleotide sequence and an annotated CDS.

4.3.6 Candidate micropeptides selection based on expression for MS analysis

Mass spectrometry was the analytical tool used in this study to measure and detect micropeptides in pancreatic adenocarcinoma samples. Mass spectrometers can identify peptides through the comparison of the mass-to-charge ratio obtained for each molecule present in the sample with a peptide sequence database.

For these proteogenomic searches, databases are constructed with peptide sequences inferred from genomic or transcriptomic evidence. Although this enlargement of sequences has potential to identify novel peptides, it raises concerns on reliable identification. A consequence of this inflation may result in an underestimated false discovery rate and a decrease in the sensitivity of identification because of the increased number of high-scoring random hits (191).

To reduce the number of entries in our dataset and obtain better results from MS analysis, candidate micropeptides were selected based on the expression values of their host-transcript. Note that we consider as host-transcript the entire transcript sequence identified by StringTie and not only the region defined as smORFs (nucleotide sequence) or micropeptide (amino acid sequence). The transcript sequence could include non-coding regions such as UTRs or introns that may appear covered by RNA-seq reads, whereas the smORFs corresponds to the potentially coding sequence and must be smaller than 100 codons.

Expression values were calculated for all the consensus host-transcripts previously obtained and in the 6 RNA-seq samples separately. We used StringTie (v. 1.3.6) applying the abundance optional parameter (-A). We used a maximum fraction of multiple-location-mapped reads allowed in a locus (-M) of 0.1. Given a list of transcripts coordinates (GTF file), this approach calculates expression in

coverage, FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Million) values. One tabular file containing expression values for all host-transcripts was outputted for each patient.

We then calculate the median expression value in TPM for each pancreatic adenocarcinoma sample. Finally, candidate micropeptides were selected if their host-transcript had an expression value (TPM) higher than the median in each of the 6 samples.

4.3.7 Strategy and final parameters to build candidate micropeptides datasets

Following the pipeline described, we constructed two independent datasets. Based on some lessons learned from the first dataset (dataset version 1 or DS1) creation which are explained in the results and discussion sections, as well as the need for including more smORFs, we redefined the parameters and steps for obtaining a second dataset (dataset version 2 or DS2). However, both are being used on the MS analysis since they are the result of combining restrictive and permissive steps and requirements. Moreover, the resulting amino acid sequences are only defined as candidates, and we probably be able to identify true micropeptide on both through MS.

For each set of candidate micropeptides, host-transcript genomic coordinates, including chromosome, start, end and exons, coding start and end, nucleotide and amino acid sequence, start codon class (canonical or non-canonical), expression values, and alignment results were annotated.

A summary of the steps and parameters used for each dataset definition is described in Table 3 and Fig 33.

Results of chapter 3 (study 1) starting in section 5.3.1 (page 208).

Database name	STEPS				
	TRANSCRIPT PREDICTION	SAMPLE COMBINATION	FROM DNA TO AA SEQUENCE	FILTER	
	De-novo transcriptome assembly	Merge and consensus of multiple samples	<i>In-silico</i> translation	Overlap with annotated CDS	Expression of the host transcript
DS1	Stringtie, -m 50, -M 1,0	Stringtie transcript merge mode	ATG	< 60 %	Not used
DS2	Stringtie, -m 50, -M 0,1	In-house strategy	ATG and CTG, GTG, TTG, ACG, ATT	< 30 %	TPM > median expression

Table 3 - Steps and parameters used for each dataset.

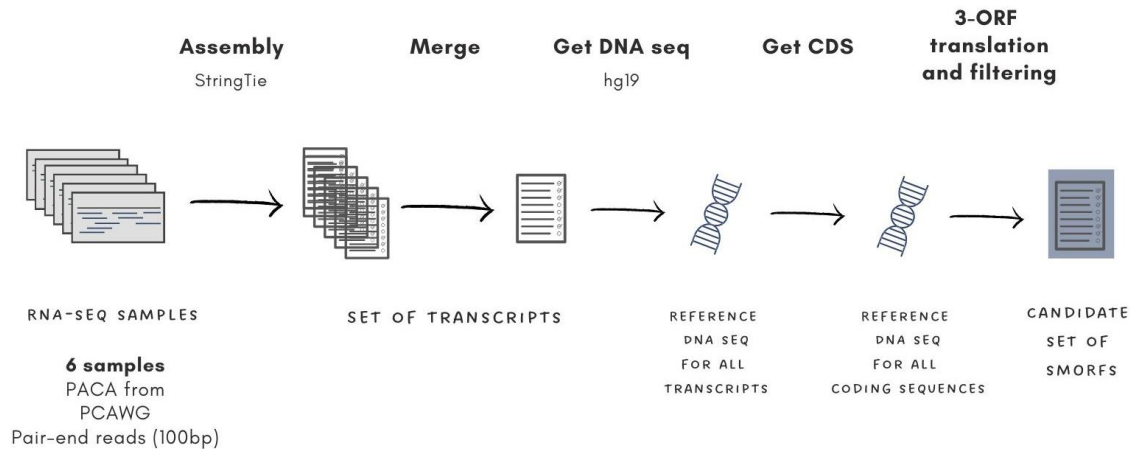


Figure 33. Schema of the general strategy used to define both catalogs of smORFs.

Study 2: Identification of candidate highly conserved micropeptides in intergenic regions

4.3.8 Collection of known and conserved intergenic human regions

The majority of known and published small ORFs have been identified considering annotated regions, including protein coding genes, pseudogenes or noncoding RNAs. However, little is known regarding candidate smORFs in intergenic DNA since it is not supposed to be transcribed, neither translated. Therefore, internally in the group and as a second study in the context of micropeptides, we intended to explore genome-wide intergenic sequences to identify novel micropeptide candidates.

We based our search on known and conserved intergenic regions from the Zoonomia Project (5). The project is an international collaborative effort focused on the discovery of the genomic basis of shared traits in mammals to understand remarkable phenotypes and the origins of disease. Through the comparison of diverse mammals, they provided genome assemblies for 131 species including humans. Moreover, the alignment of the genome of 240 species allowed to increase the power to detect sequence constraints at individual bases.

Among other public data, they furnished a list of unannotated intergenic constrained regions (UNICORNs) defined as non-coding regions on the genome that lack annotation in ENCODE3 (192). UNICORNs show high evolutionary constraint (nucleotides with a PhyloP score $> 2,270$, FDR 5%), and therefore suggest function (Fig 34). We downloaded a bigBed format file containing genome

coordinates (GRCh38) for a list of 424.179 UNICORNS. This was our starting collection of known and conserved intergenic regions.



Figure 34. Genome browser view of Zoonomia UNICORNs (shown in green) identified in the human chromosome 20. PhyloP scores can be seen above UNICORNs in grey.

4.3.9 *In-silico* translation of intergenic constrained regions

Considering the genomic coordinates obtained from The Zoonomia Project, we downloaded from the REST API data interface of UCSC the DNA sequence (GRCh38) of each UNICORN.

To identify potential candidate small ORFs within these intergenic regions, we performed a 6-ORF *in-silico* translation of each DNA sequence associated with a UNICORN (Fig 35). Since splicing had not been previously studied or identified in these sequences, and considering their high and similar conservation throughout, we directly translated the complete DNA of each UNICORN assuming they are intronless and their entire sequence is coding

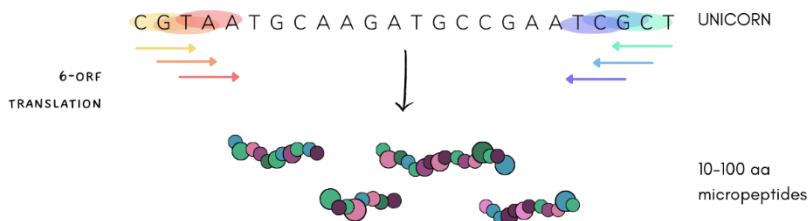


Figure 35. *In-silico* translation of UNICORNs. Nucleotide sequences are translated starting from the first, second and third nucleotide in forward and reverse (6-ORF). Only peptides between 10 and 100 aa are selected.

For the *in-silico* 6-ORF translation, codons were defined starting from the first, second and third nucleotide of the DNA sequence in forward and reverse strands. To encompass a wide range of candidates and account for the diverse codons that can initiate translation in humans, we considered not only the canonical start codon ATG but also all possible trinucleotide combinations. Translation was extended until the first stop codon was encountered, or until the end of the UNICORN sequence was reached. We retained amino acid sequences with lengths ranging from 10 to 100 aa.

4.3.10 Searching for orthologs on *Mus Musculus* using Reciprocal Best Hit approach

Once we obtained a list of *in-silico* short amino acid sequences from conserved intergenic regions, we went for more evidence to assume or suggest they could be translated into micropeptides in nature and have a functional role.

Thus, we evaluate the obtained candidate micropeptides searching for orthology on *Mus Musculus* genome (GRCm39/mm39). Orthologs are genes in different species that have evolved through speciation events only, generally assuming they have similar biological functions in these species. We used the Reciprocal Best Hit (RBH) approach to define pairs of orthologs between human and mouse. RBH assume two sequences are orthologs when each in a different genome find each other as the best hit in the other genome (Fig 36).

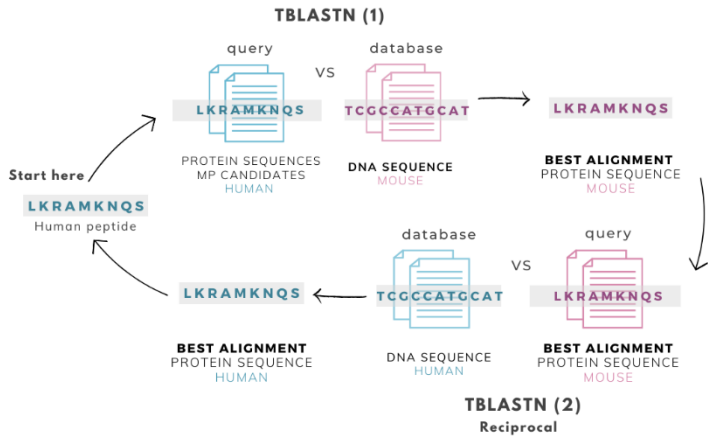


Figure 36. Reciprocal Best Hit approach.

With the aim of identifying pairs of orthologs, TBLASTN (version 2.6.0) was used in both directions to compare amino acid sequences with an entire genome. We applied default parameters except for the length of initial exact match (word size = 3). The soft masking option was also enabled and therefore repeat sequences were identified and masked for finding the initial matches.

We first compared the human amino acid short sequences obtained from the *in-silico* translation of UNICORNs with the mouse reference genome (GRCm39/mm39) (TBLASTN1). We selected those alignments with an e-value (expected value) lower than $1e-05$, showing a significant match, and an overlap between human and mouse sequences higher than 50%, meaning more than a half of the sequence matched with the sequence on the other species. Gaps among the sequence were not considered to calculate the overlap.

After applying these filters to elect only good local alignments, we extracted from TBLASTN1 results the mouse amino acid sequences. These mouse short peptides were then considered as the query for the second TBLASTN (TBLASTN2) analysis and aligned through the human reference genome (hg38).

To determine an automatic criterion to consider alignments as the best reciprocal hit in both directions (human vs mouse and mouse vs human), and therefore define orthologs, we manually inspected the obtained results. Particularly we contrasted the peptide sequence and the genomic coordinates (chromosome, start and end) of the human candidate micropeptides (query for the TBLASTN1) to the human sequences resulting from the TBLASTN2.

We define orthologs if the resulting alignment on the TBLASTN2 had an e-value lower than $1e-05$, and the aligned human amino acid sequence and their genomic coordinates were identical or overlap with the initial human candidate smORFs. We also checked whether mouse amino acid sequences were also identical or overlapped when comparing both TBLASTN analysis. Gaps within any sequence were not considered to evaluate the similarity between them. To remove duplicated genes in any species, only one-to-one (1:1) pairwise orthologs (Fig 37) were retained meaning that both genes in the pair have only one ortholog and therefore, one best hit in the other species.

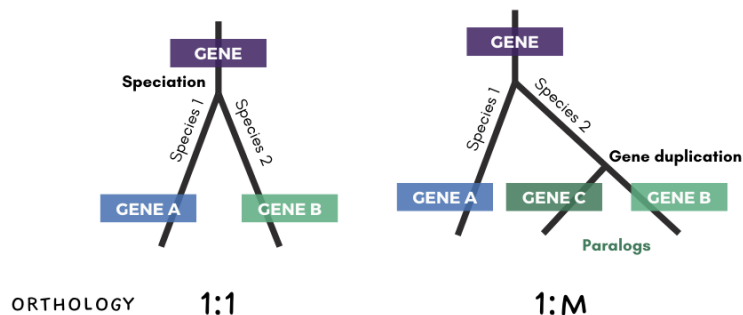


Figure 37. Representation of one-to-one and one-to-many orthologs. In 1:1 pairs, both genes have only one pair in the other species, whereas in 1:M, the gene of interest (Gene A) has more than one pair in the other species, which are paralogs between them (Gene B and Gene C)

4.3.11 Inference of purifying selection based on dn/ds ratio

To end up with a list of candidates novel micropeptides and add more information indicating functionality, we calculated the ratio of non-synonymous to synonymous (dn/ds) variants on 1:1 ortholog sequences (Fig 38).



Figure 38. The ratio of dn/ds variants is calculated for 1:1 ortholog sequences, considering only the coding regions aligned and obtained through tblastn.

Usually non-synonymous changes, that is nucleotide variants resulting in a different amino acid, negatively alter the structure and function of a protein, and may be deleterious. Purifying selection acts to remove these deleterious mutations in genes that are essential for basic cellular functions, resulting in a higher rate of synonymous variants compared to non-synonymous substitutions. Accordingly, when there are structural constraints on a functional protein and it is under strong purifying selection to maintain their role across species, the dn/ds ratio is close to 0 when compared to its orthologs.

To calculate the dn/ds ratio for each ortholog pair we used the Codeml function of the PAML (Phylogenetic Analysis by Maximum Likelihood) package (v 4.9j). We enable the pairwise option (runmode = -2) to perform a comparison between two species, human and mouse, so a phylogenetic tree was not needed for the calculation.

4.3.11.1 Expected dn/ds ratio on known protein coding genes

We expected to have a much higher rate of synonymous substitutions, which do not alter the protein sequence compared to non-synonymous on conserved genomic regions. Therefore, the dn/ds ratio should be close to 0.

To define a dn/ds threshold to select candidate functional micropeptides, we analyse and calculate the dn/ds ratio of a set of known protein coding genes annotated in Gencode (version 38, GRCh38.p13).

As we are evaluating micropeptides, short coding regions, and to ensure that the calculated dn/ds ratio on known protein coding genes was not affected by the size of the sequence, we first pick all the coding exons shorter than 1000 bp. From this subset we randomly selected a total of 300 short coding exons. Moreover, since we aim to explore candidate micropeptides that can have a role in cancer, our subset included 100 coding exons from known cancer genes based on a list provided by COSMIC database.

We downloaded the nucleotide sequence of these 300 short coding exons from the REST API data interface of UCSC (GRCh38) and we translated to get its known amino acid sequence. Following the methodology explained above, we looked for 1:1 orthologs in mouse based on the RBH approach. Finally, we calculated the dn/ds ratio using PAML for all the human-mouse pairs of orthologs. We analyzed the obtained results and determined the expected dn/ds ratio. To avoid ratios closer to 0 due to low numbers of synymous variants, we also define a threshold value for the number of silent substitutions (ds) that have occurred in the coding exon.

4.3.11.2 Selection of candidate functional micropeptides

The results (explained on section 5.3.7) obtained from the analysis of known protein coding genes, allowed us to define the expected dn/ds values for coding short regions and therefore to consider only candidate functional micropeptides.

In order to accomplish this, we first downloaded human (GRCh38) and mouse (mm39) nucleotide sequences for all candidate micropeptides previously identified as 1:1 ortholog pairs. Thereafter, we calculate the dn/ds ratio running Codeml and as explained above. Note that this value was derived by comparing nucleotide variants between the human and mouse genomes, specifically within the previously aligned region. It is important to consider that the aligned sequence may be shorter than the *in-silico* translated candidate smORF.

Considering the analysis done on known protein coding genes, candidate micropeptides were selected if they had a dn/ds ratio lower than 0,32 and a ds value higher than 0,1 to ensure variation within both sequences even their short length.

This was our final set of candidates and novel micropeptides.

4.3.12 Expression analysis of candidate functional micropeptides in normal tissues

The list of candidates novel micropeptides we provided was based on nucleotide conservation among 240 species, and preservation of amino acids when compared to their mouse orthologs. However, conservation does not always and directly imply functionality of the peptide. For this reason, we continue the evaluation of smORFs in intergenic regions through the analysis of transcription data on normal samples. Signals of expression among the regions defined as candidates show the sequence is at least transcribed and could allow us to suggest if its function is tissue-specific or not.

Considering expression values are generally calculated for known and annotated transcripts, we downloaded raw data, and in particular aligned RNA-seq data from the GTEX project (Genotype-Tissue Expression v8, dbGaP Study accession phs000424.v8.p2) (6), a resource database and associated tissue bank available for the scientific community to study genetic variation and gene expression in human tissues. We randomly selected 135 samples from a diverse range of 27 different tissues, with 5 samples per tissue. We ensured that samples from the same donor, even if obtained from different tissues, were excluded.

Available algorithms designed to calculate expression values are usually restrictive in terms of minimum number of reads aligned through the region. Furthermore, they are generally built for inspecting larger genes.

Hence, expression was evaluated by inspecting the number of aligned reads covering each candidate micropeptide, directly analyzing RNA-seq aligned bam files. Paired-end reads were extracted using Samtools (v.1.5) view mode.

Obtained paired-end reads were filtered to discard multi-mapped sequences and low-quality alignment scores (mapping quality value = 255). Moreover, paired-end reads where one of them align to a known transcript including non-coding RNAs were also excluded.

We finally evaluate counts of paired-end reads to analyze potential expression signals through our candidate micropeptides.

4.3.13 Exploring somatic cancer SNVs within candidate micropeptides to assess their role in tumorigenesis

Abnormal clustering of mutations is complementary to other signals to detect driver cancer genes. Intending to explore the potential role of micropeptides in cancer disease and tumorigenesis, we therefore started analyzing the recurrence of somatic single nucleotide variants in smORFs.

We downloaded somatic SNVs from The International Cancer Genome Consortium (ICGC). We then excluded hypermutated samples, deemed when the mutation count was greater than 1,5 times the interquartile range length above the third quartile ($> Q3 + 1,5IQR$) in their respective tumor dataset. After filtering them out, we get genomic variants from 5.807 donors, including 68 ICGC projects, 21 primary tumor types and 44.401.585 SNVs. However, only 12 different ICGC projects representing a set of donors sharing the same tumor type and collected from a specific country were selected for the analysis. Projects included and the number of SNVs identified in each are shown in figure 39.

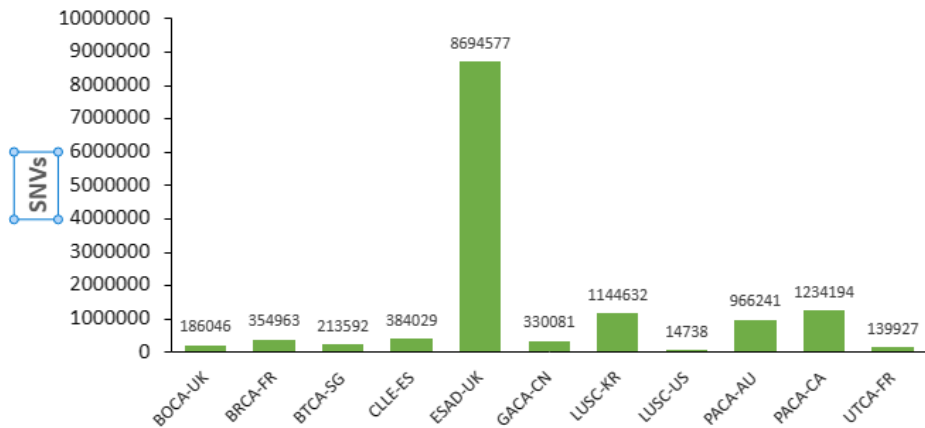


Figure 39. Number of somatic SNVs included in each ICGC project.

We used OncodriveCLUSTL (7), a driver discovery algorithm, to look for significant clustering signals of SNVs within smORFs (Fig 40) This computational method is based on a local background model, determined from the simulation of mutations accounting for the composition of tri-nucleotide context substitutions observed in the cohort under study.

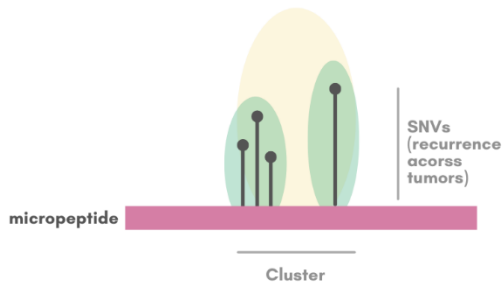


Figure 40. Identification of clustering signals of SNVs in micropeptide sequences. Figure adapted from Arnedo-Pac. et al, 2019 (7).

4.3.13.1 Applying OncodriveCLUSTL to published small ORFs

Before running OncodriveCLUSTL to explore our candidate micropeptides identified in intergenic regions, we tested the algorithm analyzing previously identified and published smORFs.

Accordingly, we take as micropeptides all the small ORFs from the SmProt database (176). SmProt contains micropeptides identified using mass-spectrometry or ribo-seq techniques and complies with other databases and literature sequences. We only selected mp identified in humans that do not share their amino acid sequence with other human micropeptides from the dataset. As micropeptides are short, it is probable to get the same translated sequence from diverse genomic regions.

Mutational processes contribute distinct depending on the region type. Therefore, we separated exon and intron regions for each micropeptide, and we only looked for clustering signals within their coding exon regions. A total of

49.065 micropeptides published in SmProt and their respective exon coordinates were used for the first analysis of cancer driver smORFs.

For each set of SNVs (classified depending on the ICGC project), we tested multiple parameter combinations for the smoothing window (sw), cluster window (cw) and simulation window (simw). We decided to elect and apply 6 parameter combinations for the 12 ICGC projects including 9 primary types. In all the analysis, default options for signature calculation and simulation mode were changed to region normalized and region restricted respectively. These options allowed to restrict the background model calculation to the given genomic regions. All the parameter combinations are specified in Table 4. OncodriveCLUSTL calculated a q-value for each smORFs indicating a significant (q-value < 0,01) signal of clustered mutations within it (Fig 41).

ID	SW	CW	SIMW
51-51	51	51	31
71-71	71	71	31
91-91	91	91	31
101-101	101	101	31
101-91-101	101	91	101
101-101-101	101	101	101

Table 4. Parameters (sw, cw and simw) tested in OncodriveCLUSTL. ID refers to the name of the combination.



Figure 4124. Schema summarizing OncodriveCLUSTL input and output files. Small ORFs and micropeptides with a q-value < 0,01 were considered potential drivers.

To choose the most adjusted combination for each set of variants, we calculated the Kolmogorov-Smirnov (KS) test (two-sided option). This statistic is used to decide if two sets of samples, in this case, the expected and the observed, have a similar probability distribution and therefore, the observed probability is not inflated. Only those p-values obtained from OncodriveCLUSTL higher than 0,01 were used to calculate the KS statistic. The enrichment in cancer genes was also considered. To do so, smORFs identified within known cancer genes were defined and counted as cancer related. Finally, the qq-plot obtained from OncodriveCLUSTL was manually inspected for this selection step.

To assume similar probabilities and low inflation, the KS value should be around 0, while the observed p-values were closer to the expected ones on the qq-plot. The enrichment should be higher, meaning we were identifying known-cancer related genes as expected when analyzing cancer driver genes.

Micropeptides with significant q-values ($< 0,01$) identified on the most adjusted combinations were evaluated for each ICGC project and could be considered potential drivers.

4.3.13.2 Evaluation of recurrent variants within novel candidate micropeptides

To evaluate the presence of somatic single variants acquired in diverse tumor types and within candidate novel micropeptides identified in intergenic regions, we first annotated them. To do so, we looked for somatic SNVs identified in the 12 ICGC selected projects and present within candidate novel micropeptides.

Following the strategy tested with published smORFs, we run OncodriveCLUSTL to analyze clusters of variants within conserved intergenic regions defined as candidate micropeptides. *Results of chapter 3 (study 2) starting in section 5.3.4 (page 229).*

5. Results

5.1. Analysis of somatic structural variants in CLL and their incorporation into subclonality studies

Chapter 1

5.1.1 Identification pipeline for somatic structural variants

The accurate identification of somatic structural variants in cancer is essential for understanding the complex genomic landscape and the underlying mechanisms contributing to tumorigenesis. The necessity of implementing advanced computational algorithms arises from the inherent complexity and heterogeneity of cancer genomes, where SVs can play pivotal roles in driving oncogenic transformation. Through the combination of bioinformatic analysis and manual inspection of the sequencing data, we provide a detailed account of the identified SVs in Chronic Lymphocytic Leukemia. This will allow us to shed light on the potential implications of somatic SVs in these cancer types. Furthermore, we considered using diverse algorithms was a crucial strategy for several reasons such as the varying sensitivity and accuracy between variant callers that might result in missing mutations or fail to detect low frequency variants when using a single program.

5.1.1.1 Evaluation of the structural variant identification pipeline

There exists a wide variety of variant callers developed by the community for the identification of somatic structural variants. In our study of CLL tumor genomes, we employed four different algorithms for this purpose. Prior to our evaluation of the results produced by each algorithm, we tested various filtering options available within the variant callers. We then compared the detected variants across the different algorithms. Moreover, we conducted a manual inspection of previously obtained variants in tumors also included in this cohort, some of which were experimentally validated in published studies.

5.1.1.1.1 Fine-tuning specific parameters used by DELLY2

As mentioned in the methodology, we allowed 5% tumor contamination in normal samples when running DELLY2 for the detection of SVs in CLL samples. This decision was taken considering we were working with a liquid tumor where both normal and tumor samples are collected usually from blood, where healthy and cancer cells coexist. Although samples were filtered before sequencing and purity was inspected, allowing a slight tolerance of tumor contamination enabled us detecting somatic variants that might otherwise be missed because of the presence of few tumors reads in the normal sample.

For DELLY2 we also evaluated the optimal percentage of alternate reads in tumor samples to be observed to identify a variant. To do so, we compared the results obtained when running the algorithm setting this parameter to 0,05 (at least 5% of alternate reads in the tumor sample), 0,2 (default value; 20% of the tumor reads must be alternate to identify the variant) and 0,5 (50% of alternate tumor reads). As an example, we counted the number of structural variants identified on a CLL patient (case 16), by using 0,05 and 0,5 values. We could observe a decrease of 1963 translocations, 15 inversions, 65 duplications and 96 deletions that were not detected when using 0,5. We considered using higher values such as 0,5 was unrealistic when analyzing tumor samples, as in light of sample heterogeneity and the difficulty of mapping structural variants, it is highly unlikely to find that number (50%) of sequenced reads confirming each of the acquired variants. Moreover, less stringent percentages allow us to detect more subclonal variants, since genome studies of CLL tumors had revealed the high subclonal heterogeneity of the tumors (193).

We then used variants identified and experimentally validated and published by Puente et al. (187) to check whether these large variants were detected by using 0,2 or 0,05 values. For a total of 35 SVs distributed across 5 tumor genomes, five were not detected with high quality when using default value (0,2). However, all except one were identified when a less stringent filter was used (0,05). Based on these results, we decided to continue using DELLY2 expecting at least 5% (0,05) of alternate reads in tumor to identify structural variants.

5.1.1.1.2 Comparative analysis of somatic structural variant callers

Integrating results from multiple algorithms increase confidence in identified variants. Before merging the results obtained through different variant callers (DELLY2, BRASS and SvABA), we compared their grade of concordance for structural variant detection. We also evaluated their performance based on published and validated structural variants.

We started comparing the SVs detected by Brass and DELLY2 in cases 63, 365 and 1669. At this point, we considered all SV types together and a windows size of just 50bp to define variants detected by different VCs as the same. We observed that around 30% of the SVs identified by BRASS were also obtained when using DELLY2. However, due to the large number of variants detected and provided by DELLY2 and including those considered as high or low quality by the algorithm itself, DELLY2 detected between 50 and thousands of SVs more than BRASS across different tumor genomes.

We then compared results obtained for SvABA against DELLY2 and BRASS. As we already knew both SvABA and DELLY2 provided more low-quality variants than BRASS, we only considered variants defined as high quality by each algorithm. Around 4% of the total number of SVs identified in a tumor genome by any of

these three algorithms were detected by all variant callers, and 7,5% by at least two.

We went back again to the selected 35 structural variants published in Puente et al. that were identified in CLL tumor genomes analyzed but not finally included in the present longitudinal study. For these SVs we checked for concordance between the results obtained for DELLY2, BRASS and SvABA using the specific mentioned parameters (see Methods section; 4.1.3.3). Manual inspection through the aligned sequencing reads was also done in 14 out of 35 SVs. Variant callers (DELLY2, BRASS, SvABA) could clearly detect 33, 31 and 27 of these SVs respectively. Only one of these missing SVs was not identified even as a low-quality variant by DELLY2 and BRASS. The remaining missing SVs in DELLY2 (1), BRASS (3) and SvABA (8) were identified but not considered high quality by the algorithm, meaning they could not achieve the minimum number of supporting reads and mapping quality values required. All the variants (14) that were inspected by tumor sequencing reads were validated, including the non-identified SV by DELLY2 and BRASS in case 853. This structural variant was not detected due to supporting reads in the matched normal genome.

Finally, we manually inspected 57 structural variants detected in any of the four tumor genomes from case 63. Structural variants identified by more than one VC, with high mapping qualities (>60 DELLY2, > 90 in BRASS and >60 in SvABA) were clearly detectable by paired-end and split reads. Those SVs identified by more than one VC, but with high mapping quality by just one algorithm tended to had a smaller number of supporting tumors reads. However, their presence could be confirmed.

We did not include SmuFin in this evaluation since the algorithm was included later in the pipeline. Moreover, SmuFin did not add more new structural

variants to the final consensus collection but generally supported some of the already validated.

This comparative analysis allowed us to define the structural variant identification pipeline (see methods 4.1.3.3 and 4.1.3.4), setting up specific parameters for DELLY2 and defining mapping quality thresholds for each VC. Also, to determine the number of algorithms we must support the variant after the merging step to include it in our conservative list of somatic SVs.

5.1.2 Exploring intratumor heterogeneity from structural variant allele frequencies

Single nucleotide variants, and short insertions and deletions were mainly used to characterize the intratumor heterogeneity. Based on the frequency of these alterations and their clusterization, cell populations of a tumor sample can be described. However, structural variants are usually not included in these studies since inferring their frequency from sequenced tumor genomes is challenging.

5.1.2.1 Sequencing coverage variation in normal and tumor genomes

Variant allele frequencies are generally measured as the ratio of tumor reads supporting the variant across the number of reads covering it. Therefore, taking into account that structural variants do not imply just one nucleotide position as SNVs, we first wondered how variable the sequencing coverage across any region of the genome was.

To do so, we explored coverage distribution across four randomly selected genomic regions (4.000bp each) in five healthy samples (30x coverage). This first exploration supported the idea that the coverage of a sample was constantly and

significantly changing across the genome, even though no large variants were identified there. These changes were not correlated with complex genomic regions, such as repeat sequences, that can usually be challenging during the alignment process (Fig 42). We could also detect variation when comparing different sequencing samples, even though they are the same genomic region. These differences could be due to technical variability since not all of them were sequenced using the same platform, as well as to sequencing artifacts. Generally, comparing these five healthy samples we could determine more peaks of high coverage in case 29. Although this case was previously sequenced for the CLL-ICGC project, it was not unique across these five samples.

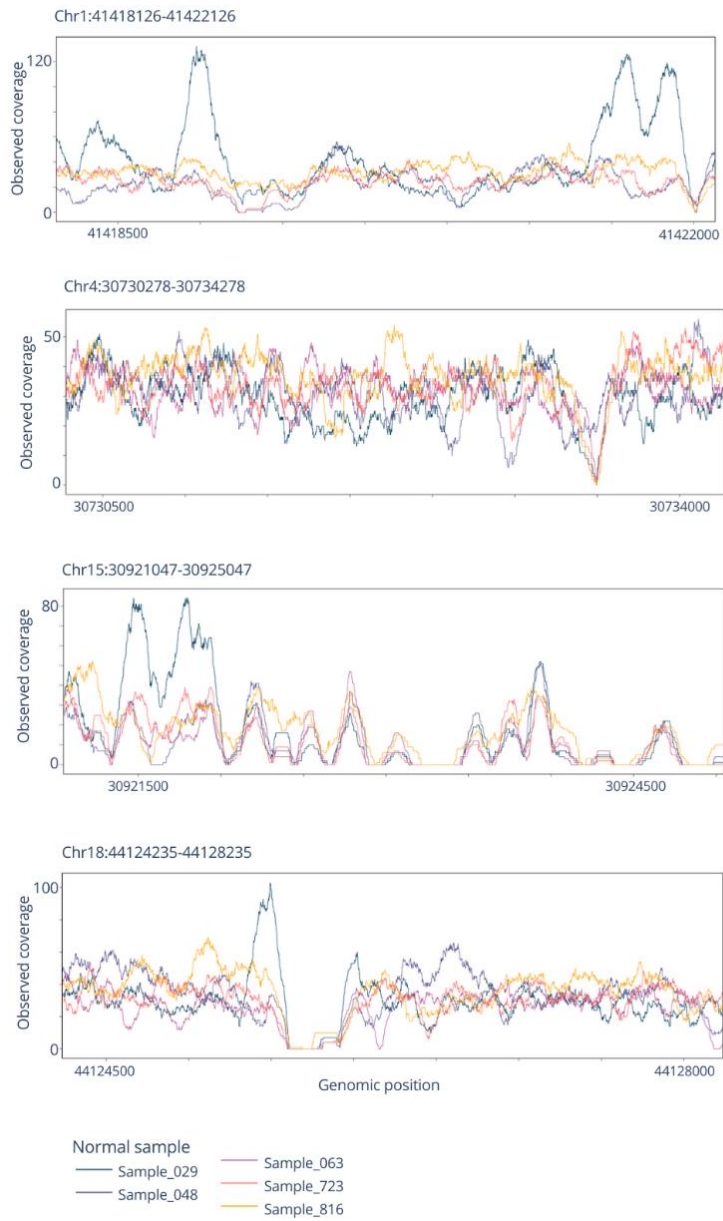


Figure 42. Coverage distribution in five healthy genomes and across four randomly selected regions.

Focusing on case 29, we also inspected its coverage distribution in 4 new regions with no SVs identified but comparing normal and two tumor genomes (Supplementary figure 1). Differences were also seen between sequenced samples from the same case. In particular, we saw a significant increase in the coverage ($> 1000x$) of tumor sample 2 (S2) in chromosome 21. Repeat genomic sequences were located within this specific region but did not have the same effect in all sequenced samples.

Lastly, we started estimating how to detect tumor reads supporting structural variants from a tumor WGS. This was the first preliminary exploration done in a few identified somatic SVs in cases 63 and 365. Tumor supporting reads were determined based on unexpected insert sizes, paired-ends joining different chromosomes, and split reads. We could start noticing that supporting reads were mainly aligned in a window around 300 bp from each breakpoint and in the up- or downstream nucleotides depending on the SV type. Coverage differences were also noticed in regions with SVs in both tumor and healthy samples (Fig 43).

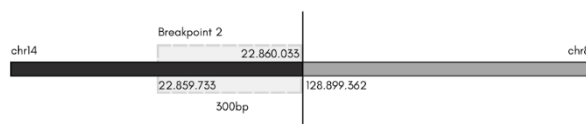
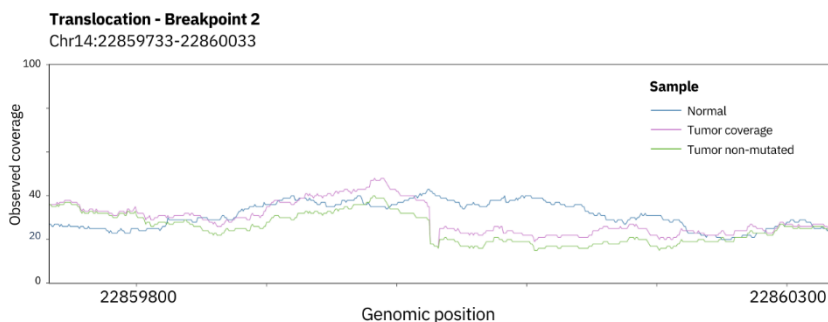


Figure 253. Coverage distribution observed across one breakpoint of a translocation identified in case 365. A representation of the translocation (chr14:22.860.033-chr8:128.899.362) is shown below the plot. A region of 300 bp upstream the bkp2 (light grey discontinuous box from 22.859.733 to 22.860.033) was analyzed. Number of reads is shown for the normal sample (blue), and for the tumor sample (pink and green). The observed total coverage (pink) is higher than the number of non-mutated reads (green) because of the presence of aligned reads supporting the structural variants.

5.1.2.2 Identification of variant supporting reads in an in-silico sequenced sample

After examination of the coverage variability across genomic regions and tumor and normal genomes, we started estimating the strategy we needed to apply to calculate variant allele frequencies for structural variants. Although we first explored tumor reads aligned in few SVs identified in CLL cases, we thought that the most accurate approach to define this strategy was comparing observed and expected frequencies of diverse large variants. For this reason, we decided to examine an *in-silico* sequenced sample generated artificially where all structural variants were heterozygous. Moreover, this artificial sample was homogeneous and did not represent many cell populations but just one clone, so the expected variant allele frequency for all the somatic mutations inserted was 0,5. Finally, reads supporting each variant were known so based on their identifier we could directly look for them in the BAM file and evaluate their alignment. The observations and messages learned from this data allowed us to define the strategy to infer structural variant allele frequencies.

Among 150 studied SVs identified in the *in-silico* tumor, we widely evaluate three somatic structural variants including one deletion, one inversion and one translocation involving two different chromosomes. Artificial sequencing reads (12, 36 and 39) supporting the deletion, inversion and translocation respectively were detected in the BAM file, and therefore correctly mapped against the human reference genome. As an example, 11 out of 12 paired-end reads supporting the deletion were completely mapped meaning both reads of the pair were aligned across at least one breakend of the deletion. For the remaining (1) supporting PE, only one read was identified aligned across a breakend whereas their matched read was mapped but not within the analyzed genomic region ($\pm 300\text{bp}$) (Table 5). This variant was not only supported by mutated PE but also split reads. From these

12 PE, all except one had at least one read of the pair splitted beyond the variant. However only four of them had both half of the split read mapped in each breakend whereas only a region of the remaining reads was aligned.

ID PE read (<i>in-silico</i>)	BKP1	BKP2
	ALIGNMENT	ALIGNMENT
chr3.b-_al1_5908595	85M15S	100M
chr3.b-_al1_20603417	77M23S	95M5S
chr3.b-_al1_25246355	58M42S	100M
chr3.b-_al1_14110103		20S80M 100M
chr3.b-_al1_28760847	100M	4S96M
chr3.b-_al1_7235689	100M	8S92M
chr3.b-_al1_14449787	100M	23S77M
chr3.b-_al1_25382709	100M	32S68M
chr3.b-_al1_19369799	100M	41S59M
chr3.b-_al1_8788975	100M	45S55M
chr3.b-_al1_13445961	100M 70M30S	
chr3.b-_al1_25415471	100M	-

Table. 5 - *In-silico* generated paired-end reads aligned across both breakends (BKP1 and BKP2) corresponding to a deletion. Each pair was aligned within a breakend except for PE chr3.b-_al1_14110103 and chr3.b-_al1_13445961, where both pairs were identified within the same breakpoint. Read chr3.b-_al1_25415471 had its pair outside the analyzed genomic region.

Regarding the insertion evaluated, this was supported by 36 split reads. We could detect all of them mapped over at least one of the breakends. Moreover, we calculated the fraction of split reads detectable depending on the nucleotide region inspected. When we looked for these split reads evaluating only the breakends positions (windows size ± 1) 36% of these split reads were not detected. However, once we increased the inspected region up to 5 nucleotides, all of them were identified as aligned in the BAM file. Although at some point we proposed to calculate the VAF based on split reads and just looking into the breakpoint

position (one nucleotide), we did not identify split reads for all SVs, neither perfectly aligned across one precise genome nucleotide.

Finally, also on these three SVs, we investigate the size of the genomic region over each breakend where mutated reads were aligned. We evaluated different sizes up (right side) and down (left side) of both breakend positions in a window of ± 1000 nucleotides. We measured the number of reads (total coverage) in each nucleotide position within these genomic regions as well as the number of non-altered reads, those not supporting the variant, and compared their values. For all three structural variants, we could observe difference values between both measures in a window of around 300 bp up and down the breakend genomic positions (Fig 44). We assumed that the discrepancy in the number of total and non-altered reads correlated with the number of aligned reads supporting the structural variant detected. In fact, when analyzing other *in-silico* and real structural variants, we could mainly identify tumor altered reads across this windows size.

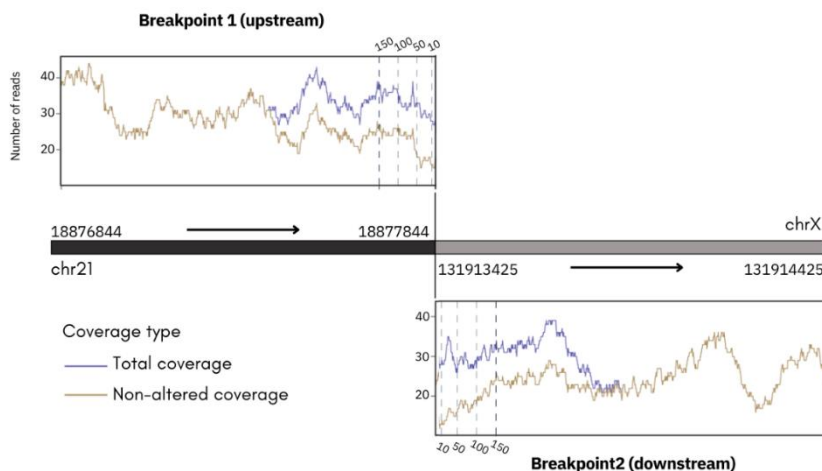


Figure 44. In-silico generated tranlocation between chr21 1887784 and chrX 131913425 0 genomic coordinates. Total number of reads aligned (blue) and reads not supporting the variant (yellow) are represented in each plot, corresponding to one breakpoint or side of the variant. A range of 1000bp upstream (left) or downstream (right) the structural variant was analyzed.

Based on this observation, we estimated the variant allele frequency for both breakends of a variant and considering diverse window sizes. As an example, Table 6 describes all genomic region analyzed for the *in-silico* selected translocation. For this artificial variant we expected a frequency of 0,5 as it was clonal and heterozygous. We estimated decreasing VAF values from 0,6 to 0,3 calculated on regions ranging between 10 and 300 bp nucleotides up or down each breakend. Therefore, results obtained on diverse window sizes suggested that the inspection of larger regions increase coverage variability and noise underestimating the observed frequency.

BKP SIDE	Windows size (bp)				
	10	50	75	150	300
BKP1	0,5553	0,4859	0,4295	0,3593	0,3171
BKP2	0,6841	0,6219	0,5772	0,4552	0,3547
mean	0,6197	0,5539	0,5034	0,4073	0,3359

Table 6 - Variant allele frequency calculated in each breakend (BKP1 and BKP2) of an *in-silico* translocation considering different windows sizes. Mean value considering both breakends is calculated in the last row.

Although expected frequencies (0,5) were obtained when analyzing smaller regions, we decided to adjust and define an intermediate size of 100bp up or downstream the breakends that matched with read length. Even the variability seen, we did not encounter nucleotide positions within this window size with outlier read counts compared to other positions on the same breakend, since all counts were between the confidence intervals.

5.1.2.3 Variant allele frequency estimation of artificial structural variants

Based on studying coverage variability and the alignment of altered reads through 162 somatic structural variants generated in the *in-silico* tumor sample, we determined and proposed a strategy to calculate mutated allele frequencies for large variants. Details regarding the decisions taken to end up with this strategy are summarized below. From the results we got from the analysis of structural variants in *in-silico* samples, we mainly considered three assumptions:

- 1) We could generally identify all mutated reads aligned across SVs in a windows of 300 bp correlating with the insert size of the sequence fragments, up or down the breakpoint depending on each variant and therefore we discarded the idea of looking for unmapped reads supporting the variant,
- 2) Split reads partially unmapped could underestimate the count of mutated reads in few nucleotides, so the number of reads aligned in these positions should be corrected even the fragments were not directly identified in the BAM file and,
- 3) Expanding the genomic region under analysis increases total coverage variability, introduces noise, and not many mutated reads were added when increasing the analyzed region. We decided to adjust and define an intermediate size of 100 bp, which matched with read length.

We theoretically reconstruct structural variants depending on the type (deletion, inversion, duplication or interchromosomal translocation) and based also on the observed coverage distribution, we define the breakend side (up or down the nucleotide position) where both total and mutated reads should be counted.

We then calculated the variant allele frequencies for both breakends of each *in-silico* generated SVs. The strategy we followed is described in the methods section (see 4.1.4.2). Variant allele frequencies calculated for each breakend of a structural variant were slightly different. Variation between VAF breakends was around 0,05 (median) for *in-silico* deletions, 0,056 and 0,13 in inversions and translocations. Larger differences were observed in translocations were in few variants, the frequency observed in each of their breakends was more than 0,4 divergent (Supplementary Figure 2).

Finally, we inferred variant allele frequencies for *in-silico* structural variants as the average between its breakends frequencies. Variant allele frequencies range from 0,097 to 0,5 even if the expected value was 0,5 for all of them (Fig 45). The median VAF observed when analyzing all *in-silico* deletions (n=45) was 0,322 and 0,336 for intrachromosomal translocations (n=97). Frequencies calculated for inversions (n=20) seem more underestimated as the median observed was 0,252.

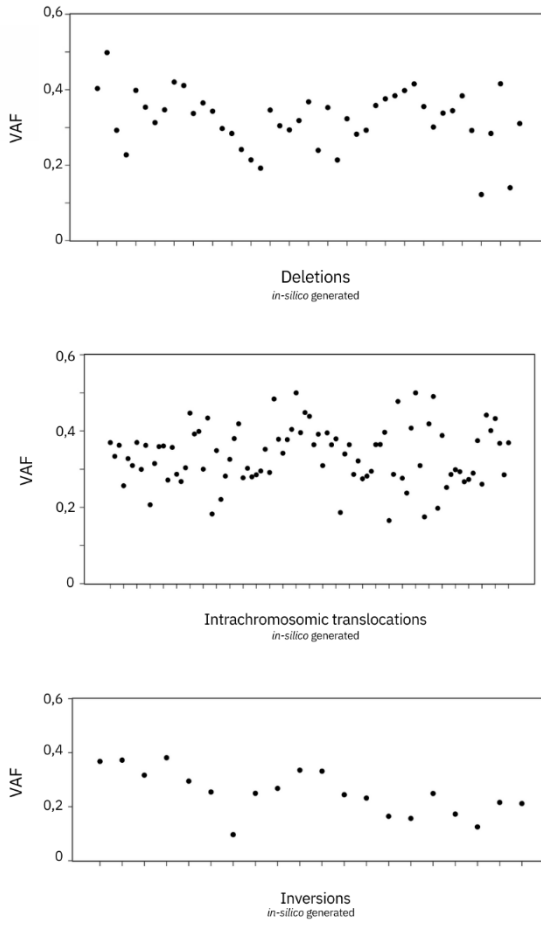


Figure 4527. Variant allele frequency calculated for each in-silico structural variant evaluated. VAFs are obtained as the average between *bkp1* and *bkp2* of each SV. Ticks in X axis correspond to variants.

Our strategy seemed to work on an *in-silico* sample where all structural variants were clonal and heterozygous. We could calculate VAF around 0,5, expecting variation as it has been observed for clonal single nucleotide variants. Although the generation of this artificial sample was performed considering typical sequencing and alignment issues, real tumor genomes are likely more complex. Furthermore, we could not test this strategy on known subclonal variants.

5.1.3 Applying the define methodology to longitudinal CLL samples: case 63

Although the entire CLL longitudinal cohort was studied, we mainly focused and described the analysis on case 63. Normal and tumor samples collected for this case had good purity and quality values. Moreover, we had longitudinal samples of three different time points including pretreatment, post treatment and Richter's transformation. Therefore, it was an interesting case to use as pilot.

5.1.3.1 Somatic structural variant landscape

The strategy defined in the present study to identify somatic structural variants across CLL longitudinal samples (see methods 4.1.3.3 and 4.1.3.4) was applied to the entire cohort, which comprises 13 cases. Results were published by Nadeu et al. in 2022 together with an extensive characterization of the genomic, transcriptomic and epigenomic profile of chronic lymphocytic leukemia (3). In the presented work, case 63 was used as a pilot case to continue with the subclonality study. Structural variants were identified for all its tumor genomes (Supplementary Table 1).

In the two samples collected at the first time point (T1-PB and T1-LN) four inversions were identified (See Fig 46, circos representation at T1). Samples T1-PB and T1-LN corresponding to pretreatment and collected at the same time point, had the same somatic SVs even though they coincided to different topographic tumor sites including peripheral blood and lymph node respectively. Contrary to other leukemias (194,195), CLL cells are known to reciprocally recirculate between the PB and LN to favor their maintenance and proliferation through the crosstalk with nonneoplastic cells on the lymph nodes microenvironment, so genomic similarities between distant CLL cells are expected. Minimal spatial diversification seems to occur between PB and LN suggesting the genomic profile of CLL remains relatively stable in diverse topographic sites before treatment. Samples collected at T1 from CLL case 63 confirmed this low genetic variability across different CLL cell locations.

Seven SVs, including all the inversions detected in T1 were identified in the tumor genome corresponding to the second time point (T2), and 27 SVs along with the previous seven were identified in the last collected time point (T3) (See Fig 46, circos representation at T2). One of the somatic inversions was detected with high quality in T3 and then rescued in T2 even it was low supported by tumor reads. However, in T1 any tumor read was alighted through variant. In summary, we observed an increase of somatic mutations acquired during time and correlating with the progression of the tumor. The significant increase observed in T3(See Fig 46, circos representation at T3) correlated with the assumption that Richter transformation might result from the accumulation of novel genetic lesions that drive clinicopathologic shift and change the course of the disease. Richter transformation is known to be marked by a profound genomic instability (196–198).

Considering the 27 unique and identified structural variants, 8 of them involved breakends in intergenic regions whereas for the remaining SVs at least one break was located within a protein coding gene. Interestingly both inversions detected in chromosome 11 and in all tumor samples, involved the *ATM* cancer gene, which is not only a driver in chronic lymphocytic leukemia but also in DLBCL, the tumor type in which CLL is transformed when developing RT. The presence of DLBCL driver genes mutated at diagnosis and prior to treatment could suggest predisposition of this CLL tumor to develop Richter transformation.

Two more inversions in chromosome 13 acquired in T2 and T3, as well as one deletion in chromosome 9 only detected in T3 involved DLBCL driver genes (*FOXO1*, *BRCA2*, *CDKN2A*, *CDKN2B*, *MTAP* and *PTPRD*)(Supplementary Table 1). (199–201).

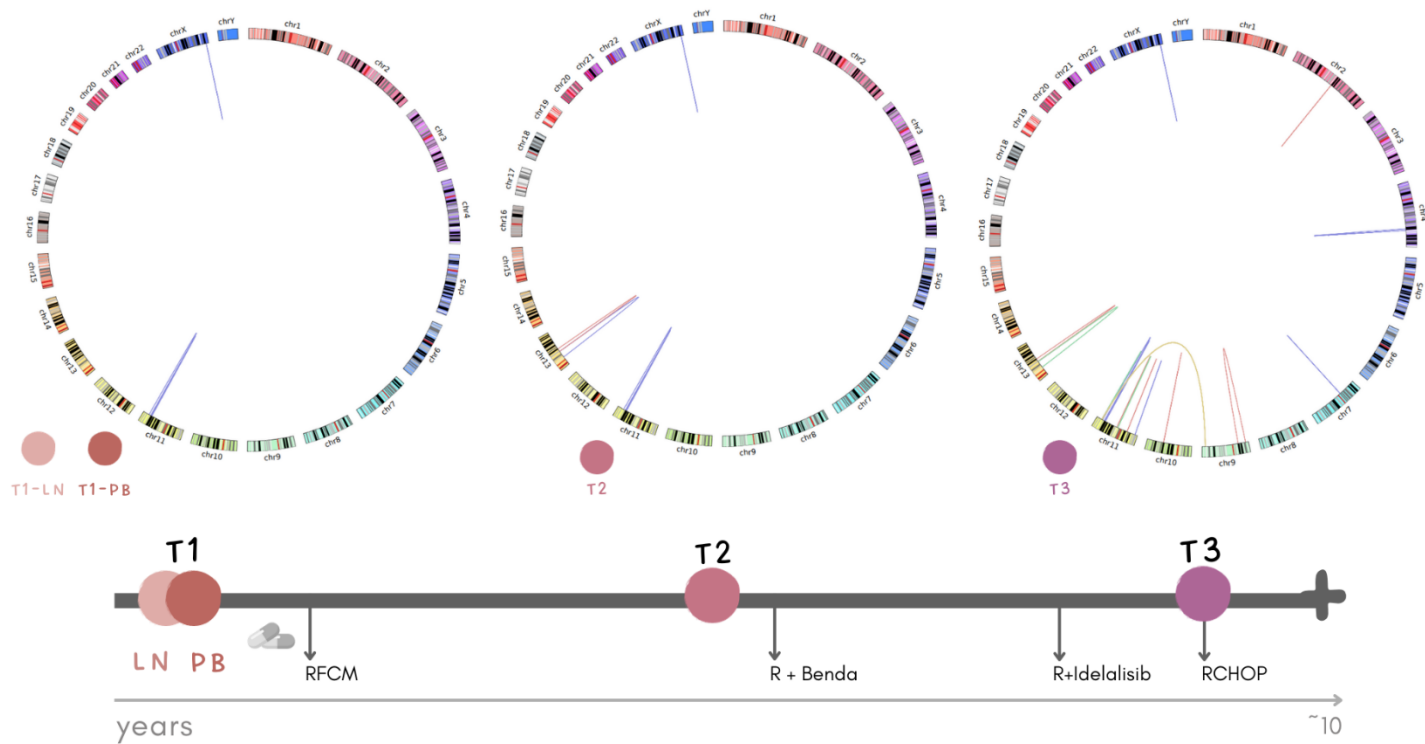


Figure 46. Somatic structural variants identified in case 63. Above, three circular representations of the human genome indicating (inside) the location of SVs detected in each tumor type. Lines in blue represent inversions, red are deletions, green are duplications and yellow translocations. Below, an schema of the patient follow-up of.

5.1.3.2 Frequency and evolution of structural variants during tumor progression

Through manual inspection of the genomic regions affected by structural variants in both *in-silico* and tumor genomes, and curated detection of mutated reads across the aligned sequences, we suggested a strategy to infer the variant allele frequency for SVs. Therefore, we applied this strategy (see Methods 4.1.4.3) on 36 previously identified somatic structural variants in tumor genomes from pilot CLL case 63. Across all these variants, four were identically detected in all samples, three were found in tumor genomes collected after treatment (T2 and T3), and the remaining were only present in the last tumor sample (T3). Note that we did not infer the frequencies of six SVs including inversions and deletions, identified in samples T2 and T3 due to their short length (< 1000bp).

Variant allele frequencies were calculated separately for both breakends of each structural variant. After that, we could see slight differences lower than 0,446 (average 0,0962) (Supplementary Table 2, column Difference VAF) between frequencies calculated on each side of the structural variant. The clonal inversion acquired in chromosome 11 across 106.417.594 and 110.207.731, detected in all tumor genomes of this case, had the highest differences when comparing both breakend frequencies in all four longitudinal samples. Interestingly, this variant linked two genomic locations covered by repeat sequences, a LINE-1 together with an Alu. Structural variants such as SV_309, SV_85 and SV_86 identified within regions affected by different copy number alterations (one breakend (bcp) within a deletion and the other bcp in a duplication) tended to have higher differences. Contrary, breakpoints composing SVs identified within regions not affected by CNVs or equally affected, had similar VAFs. Therefore, differences between breakends of an SV could be due to a challenging and difficult genomic region for performing the sequencing alignment of a large variant. The number of SVs

evaluated was too low to characterize and confirm an enrichment of a specific group of SVs with higher differences across their breakends.

Although we expected variant allele frequencies of around 0,5 for heterozygous and clonal variants, the frequencies (Supplementary Table 2, column Average VAF) obtained for the four inversions detected in all tumor genomes varied between 0,20 and 0,68. Moreover, their frequencies also changed over time. The presence of these inversions in all tumor samples collected at different time points suggested that these structural variants were clonal and present in all tumor cells expecting a VAF around 0,5. However, inversions in chromosome X had VAFs of 0,3 suggesting subclonality on the first's samples (T1-PB, T1-LN) but increase up to 0,68 after the first treatment was given (T2 and T3).

Finally, considering the VAF, the purity of each sample (0,977 T1-PB, 0,96 T1-LN, 0,97 T2 and 0,952 T3) and copy number alterations previously detected in these tumor genomes, we calculated the cancer cell fraction of each breakend (Supplementary Table 2). We expected values around 1,0 for those clonal structural variants present in all tumor cells. Few structural variants identified in chromosome 4 and 11 in the T3 tumor sample were located in duplicated genomic regions where three copies of the DNA were detected. Large deletions (CNV = 1) were also identified in other genomic regions involving chromosomes 9, 11 and 13. We obtained the cancer cell fraction of 36 structural variants present in one or more longitudinal tumor samples from case 63. Similar CCFs were obtained for SVs identified in both T1 samples corresponding to different tissues. The two inversions involving *ATM* gene in chromosome 11 appeared as clonal since they had a CCF of around 1,0, suggesting their presence in all tumor cells and in all samples (Supplementary Table 3). Contrary, the number of cells acquiring the two inversions on chromosome X had an increased in the cancer cell fraction during time and treatment exposure. Whereas on the first T1 samples these inversions

seem to appear only in half of the cells, after chemotherapy and Richter's syndrome development, their CCF was around 1,3. Structural variants acquired once the patient developed Richter's syndrome, had lower CCF and therefore they were supposed to be subclonal and just acquired by a group of cells.

As for single nucleotide variants, VAF and CCF combined with longitudinal samples collected at different time points during the development of the disease, were used not only to reconstruct tumor heterogeneity but also the evolution of cell populations during time (Fig 47).

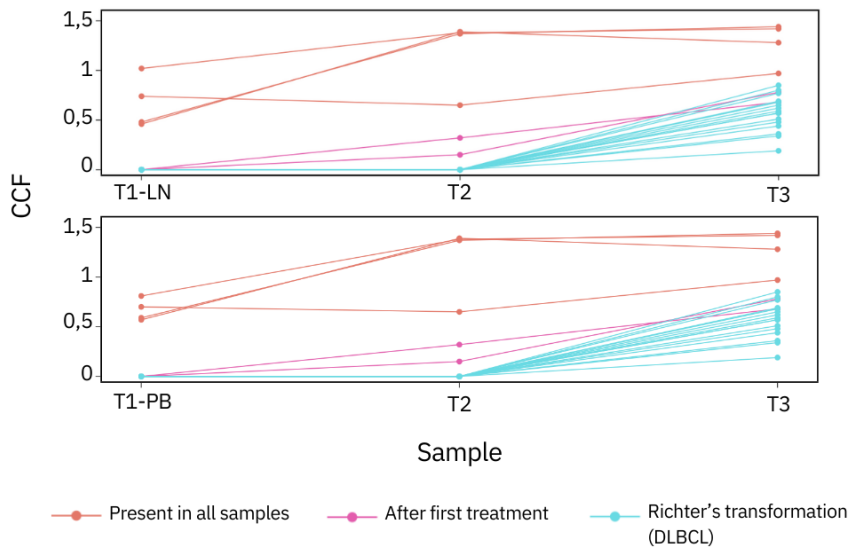


Figure 4728. Cancer cell fraction obtained for SV identified in case 63. Evolution is shown from T1 (LN above, PB below) to T3. Each dot-line represents one structural variant, and each CCF calculated in every sample. Variants are colored depending on the samples were identified; all tumor genomes (orange), those (T2 and T3) collected after the first treatment was given (pink) or only once the tumor transformed into DLBCL (blue).

Discussion of chapter 1 starting in section 6 (page 251).

5.2. Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact

Chapter 2

5.2.1 Analysis of a lung squamous cell carcinoma genome

As mentioned, somatic PPs are formed through retrotransposition and random integration within the genome, generating complex structural alterations. These variants are not uniformly and precisely detected by variant calling algorithms. Therefore, the identification of PPs from SVs is challenging and demands new strategies that can distinguish these events from other alterations captured in NGS data.

To start with the identification of somatic processed pseudogenes we first explored the structural variation landscape of one tumor genome from a patient diagnosed with lung squamous cell carcinoma. To ensure the presence of processed pseudogenes and to calibrate our protocol, we selected among those LUSC patients with more somatic structural variants detected, expecting a higher probability of identifying SVs supporting processed pseudogene formation.

5.2.1.1 Identification of somatic structural variants supporting PPs formation

The selected tumor genome was previously analyzed with the official PCAWG variant calling pipeline (202). This analysis allowed the detection of 515 somatic SVs acquired on the tumor genome and therefore not present on its matched normal DNA.

Among this set of 515 SVs, we then looked for mutations where at least one breakpoint position corresponds to an exon, suggesting the insertion point of a candidate processed pseudogene (see Methods section 4.2.2.1.1). A total of 164 SVs fulfilled this requirement after removing 6 variants annotated within long-noncoding RNA genes. This set of variants involved 97 different coding genes, and 63,9% of them were affected by more than one structural variant. Genes including *DAPL1*, *NTS*, *CNIH4* and *TOP1MT* not only had BKPs supporting an insertion point

but also variants between different exons of the source gene suggesting splicing events across their coding region, as expected when mRNA copies are retrotranscribed.

5.2.1.2 Reconstruction of *CNIH4* pilot processed pseudogene

After the first genomic exploration of one LUSC tumor genome to select structural variant supporting PPs formation, we proceed with the reconstruction of one candidate PP observed.

Among those source genes affected by more than one structural variant, *CNIH4* had two breakpoints between an exon of the source gene and another gene (receptor gene), and two more BKPs joining two different exons of *CNIH4*. Evenmore, breakpoints supporting the insertion points were located in both the first and the last exons of the source transcript isoform.

Using the genomic coordinates provided by the four breakpoints affecting this gene (Table 7) we reconstructed the candidate processed pseudogene. To do so, we looked for the exact location of each genomic position within all (6) transcript isoforms of *CNIH4* and through manual exploration with the UCSC Genome Browser. We expected to reconstruct an intron-less sequence joining all the reference exons of at least one specific transcript and inserted within the receptor region. However, independently of the gene isoform we inspected, one or two different exons appeared to be deleted (Fig 48). As an example, *CNIH4* isoform 2 was reconstructed and inserted on chromosome 7 but its second exon was not present within the cDNA sequence (Fig 49).

SV_ID	CHR	POS	DIST	GEN	EXON	SV_ID	CHR	POS	DIST	GEN	EXON
25_1	1	224544530	22	CNIH4	YES	25_2	7	158934807	-	-	NO
269_1	1	224544660	0	CNIH4	YES	269_2	1	224553578	3	CNIH4	YES
435_1	1	224553694	-1	CNIH4	YES	435_2	1	224563495	2	CNIH4	YES
455_1	1	224563738	0	CNIH4	YES	455_2	7	158934829	-	-	NO

Table 7. Breakpoints mapping *CNIH4* candidate pseudogene. Four somatic structural variants were identified by variant callers affecting *CNIH4*. Two of them (SV_ID: 25 and 455) suggested two insertion points since they involved an exon of the source gene (*CNIH4*) and a new loci (chr 7). The remaining SVs (269 and 435) represented the absence of intron sequences.

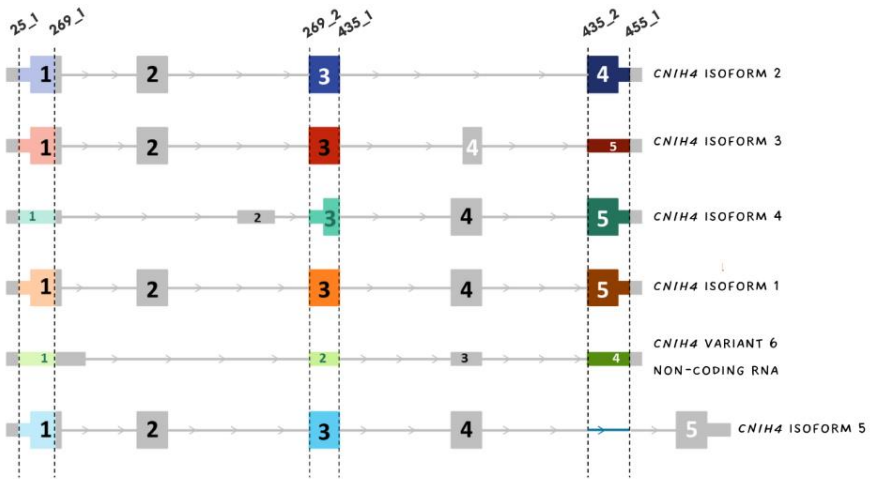


Figure 48. Breakpoints location considering six RefSeq Isoforms of *CNIH4*. Dashed lines represent each breakpoint (*_1* and *_2*) corresponding to a somatic structural variant (shown in Table 7). Grey exons and introns are those that seemed to be deleted due to the presence of an SV joining two exons of the source gene.

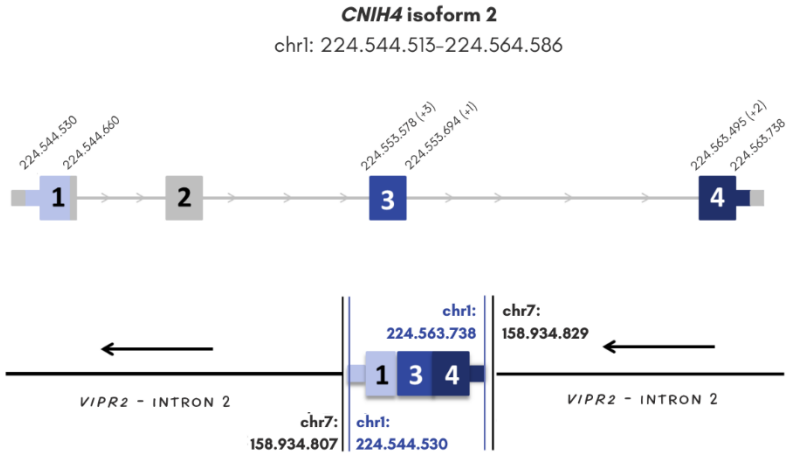


Figure 4929. Proposed reconstruction of a candidate pseudogene identified in one LUSC tumor based on detected somatic SVs. The source gene *CNIH4* (isoform 2) is retrotranscribed and inserted within the second intron of *VIPR2*. Not only intron sequences are missing but also the intermediate exon 2.

Considering low precision on identifying the exact genomic position of a structural variant by variant calling algorithms, we next inspected the DNA sequence manually. In order to verify the automatic search based on the VCF predictions, we looked for supporting sequencing reads by analyzing the tumor BAM file corresponding to this LUSC patient. Tumor reads were first extracted from the source gene region and realigned to reference RNA sequences. This step allowed us to observe split reads mapping all splice junctions from exon one to exon five of the *CNIH4* isoform 1, showing the absence of all introns and the presence of the full transcript, which is what we expect for processed pseudogenes. Moreover, split reads aligned across the first *CNIH4* exon and chromosome 7, together with paired-end reads where one read mapped *CNIH4*, and its mate mapped the same receptor location confirmed the insertion of the cDNA. The fact that we observed this retrotranscribed mRNA inserted into the genome, allow us to confirm the formation of the processed pseudogene, and refuted RNA contamination in our genome sample (Fig 50 and Fig 51). Lastly,

reads aligned to the 3' end of the source gene showed the presence of a poly-A tail, another feature of a processed pseudogene.

To further verify that this processed pseudogene was acquired somatically, we looked for split reads and paired-end reads on the normal genome and as expected, no evidence was found, confirming that the cDNA sequence obtained from the reverse transcription of *CNIH4* was inserted during tumor development.

This first and detailed analysis of a particular processed pseudogene allowed us to later define and calibrate our protocol to identify PPs on all ICGC-PanCancer genomes.

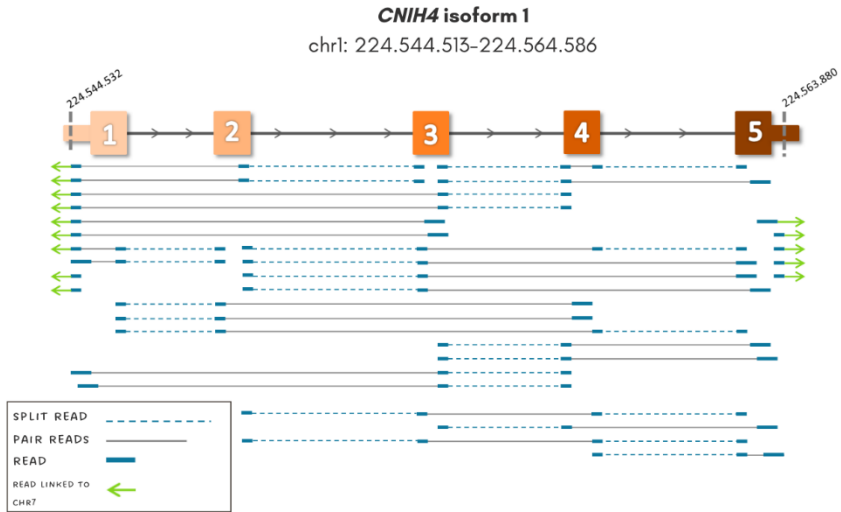


Figure 50. DNA reads from tumor WGS. Paired-end reads (grey lines) and split reads (dotted blue lines) reveal all exon-exon junctions of CNIH4 isoform 1 together. Tumor aligned reads also support the PP insertion in chromosome 7.

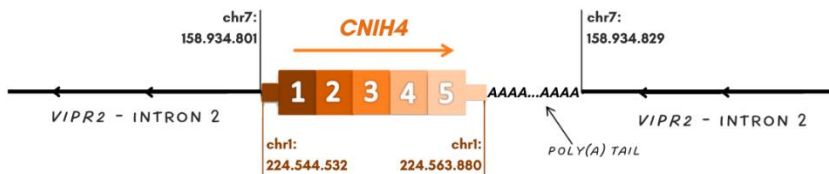


Figure 5130. Somatic PP on a LUSC genome. Reconstruction of CNIH4 pseudogene using WGS data. All five exons from transcript isoform 1 of the source gene are inserted within the second intron of VIPR2. It also includes part of the 5' UTR and the polyA tail.

5.2.2 Automatic search of PPs across all LUSC tumors based on diverse criteria combinations

To define an automatic protocol based on the search of somatic structural variants, we considered three different criteria, and we combined them to get diverse datasets of candidate processed pseudogenes. Then, we manually inspect them to evaluate each criteria combination and redefine the final search strategy. In this step, we inspected 48 tumor genomes corresponding to the entire PCAWG subcohort of patients diagnosed with lung squamous cell carcinoma (203).

5.2.2.1 Dataset 1: evidence of one insertion point

Following the criteria described on the methods section, we obtained for the first dataset (dataset 1.A) a list of 1291 candidate pseudogenes including all 48 LUSC patients (See Dataset 1.A on Table 8). For this dataset, only one structural variant suggesting the insertion of a PP was necessary to count for a candidate PP. A total of 827 candidates among this dataset were observed inserted within the same chromosome of the source gene and many of them, also near its genomic location. As an example, we manually inspected few of these candidate PP including *LPHN3*. For this candidate, we could not identify split reads confirming splice junctions on the source gene were joined. Furthermore, the breakpoint representing its insertion involved an exon with a noncoding region of the same source gene suggesting a partially deletion of *LPHN3* instead of the insertion of a processed pseudogene. To avoid mistaking PP by intrachromosomal translocations, evidence of PP insertion was only considered if the SV affects an exon and any other region of the genome, with a distance larger than 100Kb if it only involves one chromosome.

We applied this new condition to the previously defined criteria and recalculated the number of candidates processed pseudogenes for this first dataset (dataset 1.B). The number of candidate PP decreased to 806 in this

dataset, but the same number of patients (48) still retained at least one processed pseudogene. However, identifying PPs based only on one insertion point was overly permissive, and therefore not exclusive enough to define their formation. This criterion includes in the search so many other genomic events acquired in the tumors such as translocations.

5.2.2.2 Dataset 2: evidence of one insertion point and splicing events

After analyzing the first dataset, we kept applying additional criteria as evidence for somatic PP, by including the analysis of splicing events within the insertions. To do so, we combine criteria 1 and 3 (see Methods 4.2.2.2.2) to select supporting structural variants. Moreover, considering the results obtained for dataset 1, intrachromosomal translocations suggesting the insertion of a cDNA were also filtered out if the distance between both genomic coordinates was shorter than 100Kb.

For this second dataset (dataset 2) we obtained a list of 50 candidate PPs distributed across 21 out of 48 donors (See Dataset 2 on Table 8). We manually inspected 17 candidates and only 6 were confirmed as processed pseudogenes, including *C6orf48*, *C12orf57*, *DYNLL1*, *NUFIP2*, *PLEKHA5* and *RSL1D1* as the source genes. For the remaining 11 events, we could not find split reads between splices junctions of any of their transcript sequences. Therefore, the structural variants previously selected from the VCF suggested recombination between two exonic regions of the gene but could not confirm the presence of an intron less cDNA sequence. On the other hand, structural variants joining this same gene with any other genomic location suggested the recombination event was also translocated (See *BTF3* example on Table 8). Moreover, we could not identify poly-A sequences together with the candidate PP within the tumor aligned sequences.

5.2.2.3 Dataset 3: evidence of two insertion sites

The third dataset of candidates we explored (dataset 3) was based on the search of two different structural variants for each candidate PP indicating both insertion sites of the cDNA (criteria 2). To pair supporting SVs as the insertion sites of one candidate, both variants must be between an exon of the same source gene, and any other genomic location. Following this criterion, we counted 135 candidate PP acquired in 39 different donors (See Dataset 3 on Table 8). Among these candidates, 75% of the events were insertions of only one exon of the source gene. Therefore, although both insertion sites were validated in most, we could not confirm these cases were due to retrotransposition or the translocation of a particular single exon gene (See *PPT1* example on Table 8).

Without considering one exon candidates, we selected 12 events and manually inspected them. Only three (*CAPN2*, *CCDC47* and *NOL7*) out of 12 were confirmed as processed pseudogenes by looking at the tumor sequence. Split reads across splice junctions of the source genes were also found.

5.2.2.4 Dataset 4: evidence of both insertion sites and splicing events within the source genes

Combining the most conservative criteria we end up with a reliable set of candidates processed pseudogenes (dataset 4). These PP show both, evidence of insertion represented by two insertion sites (criteria 2) and evidence of splicing events between at least one exon-exon junction of the source gene (criteria 3). Since one splicing event was needed to select candidate PP, the minimum length considered for this dataset would cover at least two exon-long PP. This conservative dataset was represented by 26 candidate PP identified among 14 LUSC patients (see Dataset 4 on Table 8). From this collection, we could not find supporting reads for only one candidate (*CD177*) since a high number of possible insertion sites due to repetitive sequence were observed on the tumor aligned

reads. Therefore, manual inspection could confirm 25 out of 26 events as true somatic processed pseudogenes.


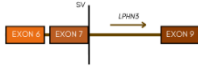

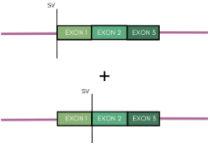



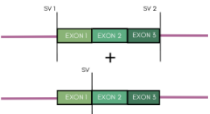
DATASET	CRITERIA	CANDIDATES	FALSE POSITIVES
Dataset 1A CRITERIA 1	One structural variant suggesting one insertion point. 	1291 candidates 48 LUSC tumor genomes	Structural variant between an exon of <i>LPNH3</i> (suggested source gene) and a non-coding region of the same gene. 
Dataset 1B CRITERIA 1	One structural variant suggesting one insertion point. *If the variant is intrachromosomal , distance between the exon and the insertion region must be larger than 100Kb . 	806 candidates 48 LUSC tumor genomes	
Dataset 2 CRITERIA 1 + CRITERIA 3	Structural variants supporting one insertion point and at least one splicing event. 	50 candidates in 21 LUSC tumor genomes	Reads supporting splicing events were not identified, and SVs suggested recombination between two not contiguous exonic regions. 
Dataset 3 CRITERIA 2	Two structural variants supporting both insertion sites of a candidate PP. 	135 candidates in 39 LUSC tumor genomes	Only one exon of <i>PPT1</i> (source gene) was inserted. 
Dataset 4 CRITERIA 2 + CRITERIA 3	Two structural variants representing both insertion sites and one SV supporting one exon-exon junction. 	26 candidates in 14 LUSC tumor genomes	

Table 8. Dataset definition based on diverse criteria combinations. Summary and schema of the searching rules. Number of candidates identified in each dataset are shown, together with examples of false positive or doubtful results (column 4).

5.2.3 Identification of somatic processed pseudogenes in all PCAWG tumor genomes

The results obtained after analyzing LUSC patients through diverse automatic searches combined with manual inspection of selected candidates were used to define the final identification strategy. As mentioned on the methods section, we applied a combination of re-defined criteria across the somatic structural variant catalog of all PCAWG patients to get a set of candidates processed pseudogenes. In general terms, this final criterion defined candidate PPs if they were supported by both insertion points. Later, we manually validated these candidates to identify somatic processed pseudogenes in a more conservative manner.

The application of this final protocol (see Methods section 4.2.2.2.3) across 2589 PCAWG tumor-normal sample, resulted in evidence for 433 somatic retrotranscription and integration events of coding mRNAs across 250 tumor genomes and 248 patients, ranging from complete mRNA copies of the source gene to fragments of different sizes, with 260 of them only consisting in one exon copy. Based on the genomic coordinates observed from their supporting somatic structural variants, 51% of candidate PPs appear to be inverted cDNA sequences compared to the strand of the host gene or insertion region.

Candidate processed pseudogenes were identified in 28 out of 34 tumor type-subtypes that were studied and were not equally distributed across them. Notably, most of these candidates (74 of 433) were acquired in pancreatic adenocarcinoma (PACA) samples sequenced in Canada and Australia. Although we studied 240 tumor genomes corresponding to PACA, only 22 of them had acquired at least one candidate PP, being the donor PCSI_0231 the PACA tumor with the highest number, 47. Pancreatic adenocarcinoma was followed by lung squamous cell carcinoma (LUSC) (69 candidate PPs), head and neck squamous cell

carcinoma (HNSC) with 46 identified candidates, esophageal adenocarcinoma (ESAD) (45), breast cancer (BRCA) (33) and ovarian cancer (OV) (32). However, seeing the number of patients analyzed for each tumor type, LUSC, HNSC, ESAD and STAD show the highest frequency among patients (50%, 36%, 34%, 34% respectively) and within them, a higher rate of PP formation (Fig 52).

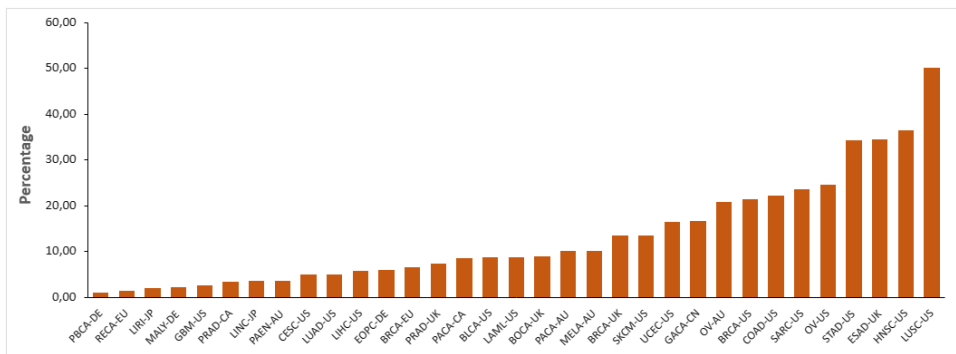


Figure 5231. Percentage of donors with at least one somatic PP identified in its tumor genome across all PCAWG tumor projects.

In most of the analyzed donors (188/248) we could only identify one candidate processed pseudogene. However, the top 5 patients with the highest number of acquired candidate PPs range between 10 and 47 somatic events. These patients were diagnosed with PACA (1 donor), LUSC (2) and HNSC (2).

Whereas all 433 source genes producing somatic PPs were not significantly enriched in any cancer related function, we could identify up to 26 of them generating PPs in different samples, and across different tumors, including the Beta-2-Microglobulin (*B2M*) and the Myosin heavy chain 9 (*MYH9*), both described as cancer driver genes (www.intogen.org). Interestingly, the *TRMT10C* gene which codes for TRNA Methyltransferase, was found to generate up to six different PPs across six LUSC, HNSC and GACA (gastric cancer) genomes.

As it was observed for germline and somatic PPs, the integration of somatic candidate PPs found across all tumor types, appears to be enriched in more accessible parts of the chromatin, like intronic regions, where we identified 49% of these cases. Insertions were allocated in all chromosomes in different proportions being chromosome 2, 3 and 1, the chromosomes with more candidate PP insertions and, chromosome Y the one with only one candidate event. Only 20% of the somatic candidate PPs identified were intrachromosomal insertions.

5.2.3.1 Manual validation of candidate PPs previously identified across PCAWG cohort

To finally get a conservative and validated set of somatic processed pseudogenes, we looked for tumor supporting reads analyzing the genomes of all 433 candidates obtained after applying the final strategy explained in section 4.2.2.2 and across all PCAWG tumor genomes.

We could identify supporting reads for both insertion sites for 69 out of 433 candidate pseudogenes. For 45 of them (Table 9), evidence of splicing was validated, this being the last set the most accurate collection of somatic processed pseudogenes.

Validated processed pseudogenes (45 in total) were distributed across six tumor types including ovarian cancer (1 PP insertion), pancreatic adenocarcinoma (2), colon adenocarcinoma (4), esophageal adenocarcinoma (7), head and neck squamous cell carcinoma (7) and lung squamous cell carcinoma (25). Similarly, to the results obtained after evaluating candidate PPs, 40% of these somatic PPs were inserted within an annotated gene and only *B2M* was identified as the source gene in two different events. Five processed pseudogenes result from cancer genes including *B2M*, *MAX* and *MYH11* that are tumor suppressor genes, and *KTN1* which is an oncogene. Insertions were only identified in 17 different

chromosomes, chromosome 7 being the sequence with the highest number of insertions, followed by 2 and 3. Only two of these retrotransposed insertions were intrachromosomal.

PCAWG Project	Donor ID	Source gene	Cancer gene	Insertion site	Gene	RNA evaluation
COAD-US	613aa3e8-a70b-45a9-9c08-0c2346c8bf00	<i>C2orf69</i>		Chr5:98157631-98157718		Non conclusive
COAD-US	613aa3e8-a70b-45a9-9c08-0c2346c8bf00	<i>HNRNPM</i>		ChrX:99988609-99988620		Non conclusive
COAD-US	613aa3e8-a70b-45a9-9c08-0c2346c8bf00	<i>LDHB</i>		Chr5:133665049-133664911	<i>CDKL3</i>	Non conclusive
COAD-US	613aa3e8-a70b-45a9-9c08-0c2346c8bf00	<i>PPP1CA</i>		Chr6:82636007-82636017		Non conclusive
ESAD-UK	OCCAMS-AH-096	<i>B2M</i>	TSG	Chr2:65746571-65746585		No RNA-seq DATA
ESAD-UK	OCCAMS-AH-047	<i>CLUAP1</i>		Chr15:91625252-91625267		No RNA-seq DATA
ESAD-UK	OCCAMS-PS-012	<i>DDX18</i>		Chr2:70601348-70601364		No RNA-seq DATA
ESAD-UK	OCCAMS-AH-091	<i>LRRC31</i>		Chr2:165637601-165637674	<i>COBLL1</i>	No RNA-seq DATA
ESAD-UK	OCCAMS-ZZ-009	<i>LYZ</i>		Chr5:147527887-147527897		No RNA-seq DATA
ESAD-UK	OCCAMS-WG-019	<i>RPS27L</i>		Chr3:168507166-168507153	<i>EGFEM1P</i>	No RNA-seq DATA
ESAD-UK	OCCAMS-RS-024	<i>SH3KBP1</i>		Chr11:12104723-12104737		No RNA-seq DATA
HNSC-US	64bb5550-2735-4401-a0db-58ec1020a32d	<i>ALDH1A1</i>		Chr3:188900757-188900716	<i>TPRG1</i>	Non conclusive
HNSC-US	8c238d30-df8e-4e6b-98fc-21696269a294	<i>ANAPC13</i>		Chr8:89993268-89993312		Non conclusive
HNSC-US	d89b1fd6-bef4-4803-8ed3-3b442be600b6	<i>GNPNAT1</i>		Chr1:156009767-156009727	<i>UBQLN4</i>	Non conclusive
HNSC-US	8c238d30-df8e-4e6b-98fc-21696269a294	<i>KRT17</i>		Chr11:127262292-127262335		Non conclusive
HNSC-US	fafd6f5b-1d76-4537-bd1c-e0bd7b4e2166	<i>KTN1</i>	oncogene	Chr4:95176296-95176410	<i>SMARCAD1</i>	EXPRESSED
HNSC-US	64bb5550-2735-4401-a0db-58ec1020a32d	<i>MAP3K4</i>		Chr4:95176296-60042427		Non conclusive
HNSC-US	8fc1f1be-d2d5-4b3a-9973-f4d964018beb	<i>NCAPH</i>		Chr1:163628659-163628640		Non conclusive
LUSC-US	e6b72c24-1607-43b9-8b8a-7bf83eea5895	<i>ATP6VOC</i>		Chr7:128608087-128608087	<i>TNPO3</i>	Non conclusive

LUSC-US	1f6b2aca-7357-40d1-ba7a-99227d9900a2	<i>B2M</i>	TSG	Chr14:64088805-64088793	<i>WDR89</i>	EXPRESSED
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>C1orf131</i>		Chr2:45310258-45310311		Non conclusive
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>CNIH4</i>		Chr7:158934807-158934829	<i>VIPR2</i>	EXPRESSED
LUSC-US	19f0cb8c-2e57-4310-967f-a9890f1605db	<i>COX5A</i>		Chr3:132099251-132099231		Non conclusive
LUSC-US	0398eae1-7216-4595-80a5-6b117d96e070	<i>CYFIP2</i>		Chr7:88489076-88488997	<i>ZNF804B</i>	Non conclusive
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>DAPL1</i>		ChrX:84343761-84343733	<i>APOOL</i>	EXPRESSED
LUSC-US	6fd72426-f6c8-47ca-a500-d5d3600b9b15	<i>EIF2S1</i>		Chr7:109918576-109918576		Non conclusive
LUSC-US	b913d254-8307-4b8a-8313-d978e32bb38f	<i>EIF5B</i>		Chr8:119911085-119911171		Non conclusive
LUSC-US	422a46b2-a67c-4a7e-923f-9b651ced96f8	<i>FAM210B</i>		Chr12:27139136-27139146	<i>TM7SF3</i>	Non conclusive
LUSC-US	0398eae1-7216-4595-80a5-6b117d96e070	<i>FGGY</i>		Chr22:36471006-36471063		Non conclusive
LUSC-US	e6b72c24-1607-43b9-8b8a-7bf83eea5895	<i>GLRX5</i>		Chr2:96832304-96832242		Non conclusive
LUSC-US	0e2ee54a-51c9-4868-842d-a2a1c1cfb016	<i>KRT5</i>		Chr9:103617685-103617686		Non conclusive
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>MAX</i>	TSG	Chr3:185087525-185087489	<i>MAP3K13</i>	Non conclusive
LUSC-US	b5e2cbda-bbfa-4ef8-a9c4-cb978bef9b23	<i>MED10</i>		Chr8:131232562-131232444	<i>ASAP1</i>	EXPRESSED
LUSC-US	6fd72426-f6c8-47ca-a500-d5d3600b9b15	<i>MYH11</i>	TSG	Chr7:97399222-97399198		Non conclusive
LUSC-US	1ee543d5-b8c0-4f79-8373-6bb6319f2ee2	<i>MYL9</i>		Chr4:2331210-2331241	<i>ZFYVE28</i>	Non conclusive
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>NTS</i>		Chr15:77767421-77767434	<i>HMG20A</i>	EXPRESSED
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>PSMA1</i>		Chr16:1781321-1781328	<i>MAPK8IP3</i>	EXPRESSED
LUSC-US	3666bc65-8e40-409e-9a1f-41583dd6d978	<i>RTF1</i>		Chr7:106287513-106287489		Non conclusive
LUSC-US	9293e197-e38a-4e19-a7d0-1b45d1ad48bd	<i>SPATS2L</i>		Chr2:136784525-136784508		Non conclusive
LUSC-US	b5e2cbda-bbfa-4ef8-a9c4-cb978bef9b23	<i>SSBP1</i>		Chr1:226325706-226325753		Non conclusive

LUSC-US	422a46b2-a67c-4a7e-923f-9b651ced96f8	<i>TFDP2</i>	Chr12:116966949-116966863	Non conclusive
LUSC-US	9af6ed4e-8cdc-4f49-84e9-ba1053b5b3ca	<i>TOP1MT</i>	Chr6:88791065-88791366	Non conclusive
OV-AU	AOCS-159	<i>PFDN2</i>	Chr6:146295348-146295371	Non conclusive
PACA-CA	PCSI_0231	<i>PERP</i>	Chr1:111835475-111835492 <i>CHIA</i>	No RNA-seq DATA
PACA-CA	PCSI_0231	<i>PHAX</i>	Chr4:72593135-72593118	No RNA-seq DATA

Table 9 - Validated somatic processed pseudogenes. For each somatic event, source gene generating the PP, its classification depending on the Cancer Gene Census database, and the genomic coordinates of the insertion site and their corresponding gene name are described.

5.2.4 Evaluation of potential PP-host gene fusion transcripts

Almost half of the identified candidate processed pseudogenes were inserted within annotated genes. Although, the likelihood that a particular PP integrates into a region with transcriptional activity and in the right orientation is low, previous studies have shown a fraction of germline and somatic human PPs to be expressed.

As a first approximation to study the potential functional impact of somatic PPs identified across all tumor types, we explored evidence for expression using RNA-seq data available for 144 samples containing 257 previously identified candidates (51% inserted in intergenic regions, 48% inserted within genes).

From this analysis, we could identify read support (split and paired-end reads) for the expression of 17 PPs, across 14 different samples and 6 different tumor types (BRCA, HNSC, LUSC, OV, SKCM and STAD) (Table 9).

Whereas three of these expressed PPs were located within intergenic regions, the majority (14) were inserted in different parts of genes, generating diverse forms of PP-host gene fusion transcripts with a variety of potential forms of functional interactions. An example is shown in Figure 53. Four of these PPs were inserted outside the coding region of the host gene, but the remaining ten directly affect the coding potential of genes, as we could infer from the RNA-seq data. Seven PPs out of 14 were inserted in the opposite reading direction compared to their host gene.

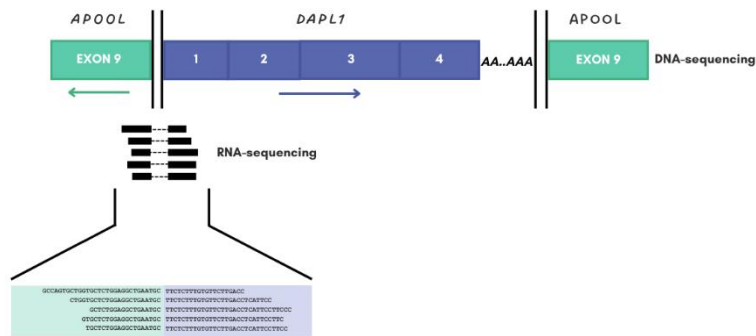


Figure 53. Chimeric DAPL1-APOOL transcript. DAPL1 (source gene) was found as a somatic PP inserted into exon nine of APOOL, on the tumor DNA sequence of a LUSC donor. The expression of the fusion transcript was confirmed with tumor RNA-seq from the same patient. The figure shows the split reads (black) mapping both transcripts together as a chimeric, and its sequences.

Considering supporting reads and performing an *in-silico* translation of the sequence, the reconstruction of the potential PP-host gene fusion transcripts predicts that the major form of PP insertion would generate a premature stop codon within the coding region of the host transcript (Fig 54). This event could be generated either because of the presence of intronic sequences, or because the PP integrated in the opposite direction of the host gene.

Alternatively, we cannot discard that these inverse PP integrations generate antisense transcripts (partial or complete, like for *B2M*) that could interact, in this case, with the transcript of the source gene.

Discussion of chapter 2 starting in section 6 (page 254).

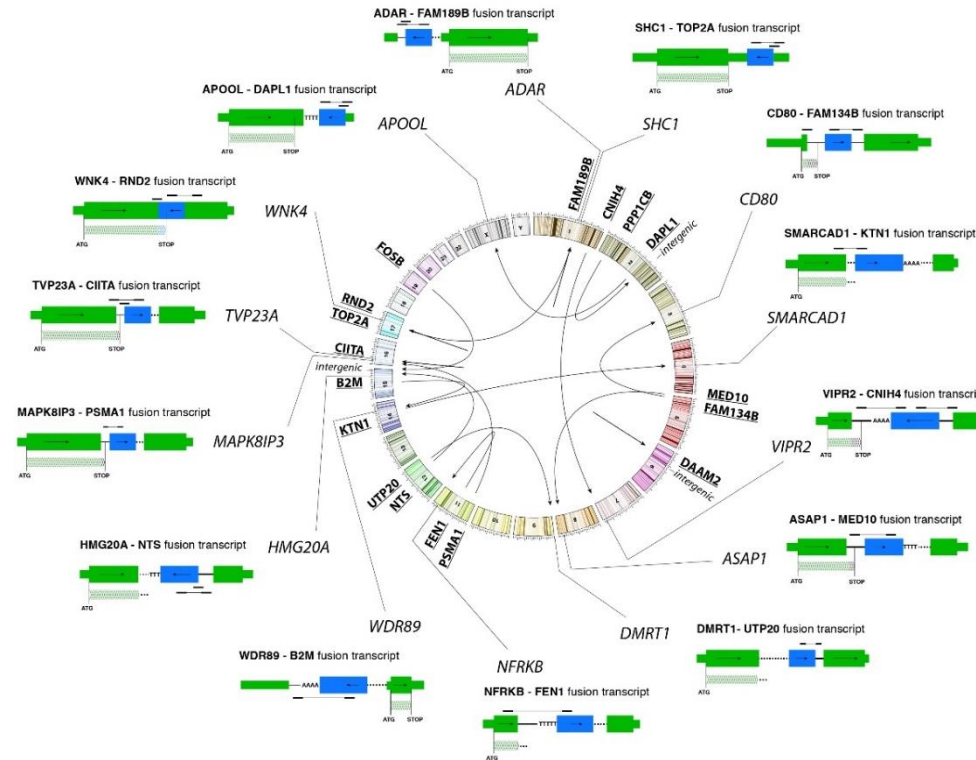


Figure 54. Host gene and processed pseudogene fusion transcripts including 17 somatic events. Circos represent the human reference genome with all chromosomes. Arcs with arrows within this circos correspond to a somatic PP, connecting the source gene (underlined and bold) with the corresponding integration site. All except 3 events are inserted within genes. For these 14 PPs, the predicted fusion transcript structure is shown in the outermost layer of the figure. Coding potential is shown below the fusion transcript representation. Start codon is indicated as ATG and termination as STOP. Dots represent uncertain termination.

5.3. Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data

Chapter 3

Study 1: Catalog of candidate micropeptides for MS/MS searches

5.3.1 Predicting non-reference-based novel transcripts for pancreatic adenocarcinoma samples

The results of these sections are framed within one of our objectives related with the definition of a new catalog of candidate micropeptide sequences for mass-spectrometry searches and interpretations within a Pancreatic cancer study.

Diverse datasets of small open reading frames have been obtained from computational and experimental approaches and are publicly available. Nevertheless, most of these sequences have been identified translating annotated and known human transcripts. Moreover, it is known that transcription and translation are tissue specific.

To end with a more specific cancer-type dataset and to observe novel micropeptides within non-annotated transcripts, we performed *de novo* transcriptome assembly from six independent RNA-seq samples from pancreatic adenocarcinoma human tissue. Our approach to generate these micropeptide-enriched transcriptomes is based on the use of StringTie, a software that generates transcript assemblies for different needs and scenarios. Stringtie predicts transcript's start and stop coordinates based on sudden drops in coverage of the aligned reads. Since different optional parameters can be set up when running StringTie, we first tested the different possibilities for processing the data and selected the one with more indications of micropeptide enrichment.

5.3.1.1 Calibrating StringTie

With the aim of finding the best set of parameters and threshold for StringTie according to our needs, we generated a testing scenario, where we experimented with different settings.

To predict transcripts based on aligned read clusters, StringTie uses a default minimum size of 200 nucleotides length. However, since the aim of this project was to identify short open reading frames of less than 300 coding nt, we decided to test and modify this threshold.

We performed a quick comparison on StringTie results after running the algorithm on one PACA sample. We run StringTie using default values for all parameters as well as modifying the minimum length allowed for the predicted transcripts to 50nt. As expected, we could observe an increase in the number of assembled transcripts of around 21.500 sequences. Although few of them were single exon transcripts covered by a low number of reads, StringTie could also identify short sequences and their splice junctions based on split read detection. At this point of the study, intending to cover most of the transcriptome to later translate it, we decided to continue all the following analysis and datasets creation using 50 nucleotides as the minimum length to predict *de novo* transcripts.

Not only was the minimum transcript length evaluated to adjust the algorithm settings, but also the fraction of multiple-location-mapped reads allowed to be present in a locus (-m). Usually, high multi-mapping reads occur in RNA-seq samples due to transcript isoforms, repetitive elements or low complexity sequences such as poly-A tails. To address this issue when identifying novel transcripts based on read coverage, StringTie was tested and used considering two different values (default and 0,1) as the fraction of multi-mapped reads within the predicted transcript.

We compared the results obtained under both assemblies, and we could observe a decrease of around 5.000 predicted transcripts in the same tumor sample. Although the number of transcripts was lower when excluding multi-mapped reads ($-m = 0,1$), not all the obtained transcripts in this analysis were identical to the previously identified with the default allowed fraction. When considering only unique reads, the coverage across the transcriptome was modified and consequently, read clustering and the prediction of start and stop transcript coordinates. For this reason, predicted transcripts were not all identical in both analyses.

When identifying new transcripts from RNA-seq, there are reasons both to exclude and to retain multi-mapped reads. After careful inspection, we found that the results did not significantly favor one value over the other for the multimapping parameter ($-m$). Therefore, we chose to use both values to create two different sets of smORFs: the default (1,0), which allows the maximum fraction of multi-mapped reads, and 0,1, which considers only unique mapped sequences. The specific choice between these two values depends on the subsequent filtering steps employed to define the catalog of smORFs. The implications and reasons behind this choice will be further described and discussed in the following discussion section.

Both sets of predicted transcripts were used to continue with the identification of novel smORFs in pancreatic adenocarcinoma tumor samples. Therefore, as described in the methodology section of this thesis, two datasets of small open-reading frames were obtained under different parameters applied not only when predicting transcripts but in all the steps. Results observed for each dataset are explained below in sections 5.3.3.1 and 5.3.3.2.

5.3.2 Assessment of transcript clustering based on different criteria

After running StringTie across different RNA-seq, we combined the results to create a representative transcriptome for our pancreatic adenocarcinoma samples. To do this, we applied the StringTie transcript merge mode for the dataset version 1, but also established an in-house strategy to obtain a consensus set for dataset version 2. An examination of the results we observed after applying different criteria while generating the in-house strategy is described in this results section.

To merge transcripts and isoforms identified on different samples, we explored the overlap between their genomic coordinates. In the first approach we only merge those isoforms with the exact same start and end (windows size 0). However, any predicted transcript was identified in all six samples under this criterion, only 3 were defined in 5 out of 6 samples and around 99% of the predicted transcript in each sample were considered unique. Because the identification of transcripts when running StringTie was based on drops in coverage of aligned reads, the prediction of exact same start and end coordinates across samples was extremely unlikely. Therefore, applying this criterion did not appear reasonable.

We then tested a range of window sizes to consider start and end coordinates representing the same transcript even it was predicted in diverse samples. As expected, the number of transcripts predicted across samples that clustered among themselves increased as window sizes increased too. Moreover, the number of unique predicted transcripts decreased, observing a significant change when comparing results using a windows size of 250 and 500. In particular, 92% of the predicted transcripts were not clustered if the criteria used was a windows size of 250bp, while for a window size of 500bp, 87% of them were

exclusively identified in one sample. Around 82 and 85% were considered unique across other windows sizes applied (from 750 to 2000bp), therefore this percentage appeared to be stabilized through window sizes higher than 500bp.

StringTie was not only able to detect start and end coordinates but also define exons. Isoforms with different number of exons or different exon coordinates for a specific transcript were also detected even within one sample. Therefore, we decided that the number of exons should also be considered when merging results across samples. When comparing the results obtained after applying different windows sizes, we not only evaluate the number of shared and unique start-end transcript coordinates. We also count the number of clustered transcripts in each merge, that had the exact same start and end coordinates for all their exons. Since exons were predicted based on split-reads, their coordinates were more precise and reproducible across samples. As mentioned, the number of clustered transcripts was constantly increasing when using larger windows sizes. However, the number of clustered transcripts with exactly all same exon coordinates decreased with larger windows sizes revealing more clustered sequences did not probably represent the same transcript. The percentage of clustered transcripts sharing the same exon coordinates among all clustered transcripts was similar (around 39%) across merging results when the windows sizes were from 500 to 2000 (Fig 55).

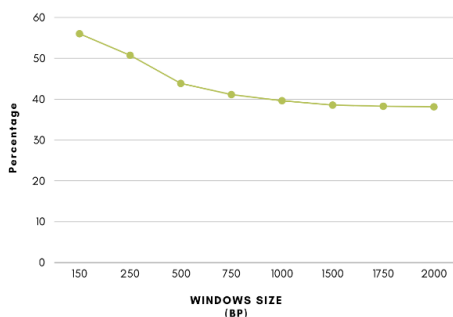


Figure 32. Percentage of clustered transcripts sharing the same exon coordinates among all clustered transcripts. Values calculated for each windows size (bp) used to merged predicted transcripts from different RNA-seq samples.

Considering the percentage of clustered transcripts after applying a range of window sizes, the number of clustered transcripts sharing the exact same exon coordinates across, and how these numbers increased, decreased or maintained similar, we decided to merge transcripts from different samples when their start and end coordinates were shared within 500bp. The clustered transcripts had to share the same number of exons and strand orientation.

5.3.3 Small open-reading frames datasets: insights from two different criteria

As previously mentioned, the aim of this presented study was to obtain a collection of candidate micropeptides, to use as a reference dataset for the mass spectrometry analysis performed on pancreatic adenocarcinoma exosomes. In this stage of the study, we generated two datasets following multiple steps and based on distinct parameters with the aim of capturing a comprehensive range of information while mitigating false positives. While dataset version 1 was slightly more conservative (Table 10), we tried to enlarge the collection of candidates on dataset version 2 (Table 11). However, it was imperative to consider the total number of candidates obtained in each dataset, as mass spectrometry analyses are constrained by dataset size limitations.

Dataset version 1 (DS1)				
Tumor sample	STEPS			
	PREDICT TRANSCRIPTS	COMBINE SAMPLES	GET PEPTIDE SEQUENCES	FILTER
	De-novo transcriptome assembly	Merge and consensus of multiple samples	<i>In-silico</i> translation	Overlap with annotated CDS
PACA1	95.015			551.494 candidate micropeptides
PACA2	106.682			
PACA3	65.222	25.207 merged predicted	838.377 candidate	
PACA4	186.180	transcripts	micropeptides	
PACA5	141.505			
PACA6	196.061			

Table 10. - Predicted transcripts and candidate sequences obtained in each step for database version 1

Dataset version 2 (DS2)							
Tumor sample	STEPS						
	PREDICT TRANSCRIPTS	COMBINE SAMPLES			GET PEPTIDE SEQUENCES	FILTER	
	De-novo transcriptome assembly	Clustering transcripts from multiple samples	Recalculate consensus sequence	Select representative transcripts	<i>In-silico</i> translation	Overlap with annotated CDS	Expression of the transcript
PACA1	90.651						
PACA2	100.963						
PACA3	61.758	589.475 clustered isoforms	589.145	27.849	6.366.662	3.733.227	1.211.051
PACA4	176.703		consensus sequences	merged predicted transcripts	candidate micropeptides	candidate MP	candidate MP
PACA5	136.574						
PACA6	182.098						

Table 11 - Predicted transcripts and candidate sequences obtained in each step for database version 1/2

5.3.3.1 Dataset version 1: a more conservative set of small

ORFs

We decided to start the identification of novel smORFs applying a more restrictive criterion. Following the methodology previously explained and the mentioned StringTie parameters, we first ran the algorithm for six pancreatic adenocarcinoma samples allowing a maximum fraction of multi-mapped reads of 1'0.

5.3.3.1.1 De novo transcript prediction allowing multi-mapped reads

When allowing multi-mapped reads (-m 1'0) a total of 95.015, 106.682, 65.222, 186.180, 141.505, 196.061 transcript isoforms were predicted for samples PACA1, PACA2, PACA3, PACA4, PACA5 and PACA6 respectively. An overview regarding the size of the transcripts, the number of exons, the number of reads covering each sequence, and their distribution across chromosomes is shown in Supplementary figure 3. Around 20.000 predicted transcripts in each of this samples overlapped with an Ensembl annotated transcript just by evaluating their start and end coordinates within a 100bp windows size.

5.3.3.1.2 Consensus set of predicted transcripts using StringTie algorithm

After using StringTie to predict all possible transcripts based on their sequencing coverage, we combine the results of these six samples to obtain a consensus set of sequences corresponding to the pancreatic adenocarcinoma transcriptome. For this dataset version 1, we combined the results of all the samples using StringTie transcript merge mode. After performing the merging step, we get a list of 25.207 predicted transcripts summarizing the pancreatic adenocarcinoma transcriptome of these six samples. Distribution across

chromosomes is shown in Figure 56. Around 10% of them were single-exon transcripts. Genomic coordinates resulting from the use of the StringTie merge mode were recalculated by the algorithm and do not precisely match those identified in each individual sample. Therefore, we were unable to explore the proportion of transcripts originating from each sample neither the degree of overlap between them, as StringTie does not provide this information on its merge output. Considering the recalculated genomic coordinates for these set of combined predicted transcripts, we used StringTie to get the number of reads covering each of them and in all tumor samples separately. Based on these, we observed that supporting reads were not identified for 182 merged predicted transcripts in PACA1, 112 in PACA2, 182 in PACA3, 151 in PACA4, 136 in PACA5 and 108 in PACA6. The median TPM observed for merged predicted transcripts and in each sample range between and 1,9 (PACA1) and 8,7 (PACA2) (Fig 57).

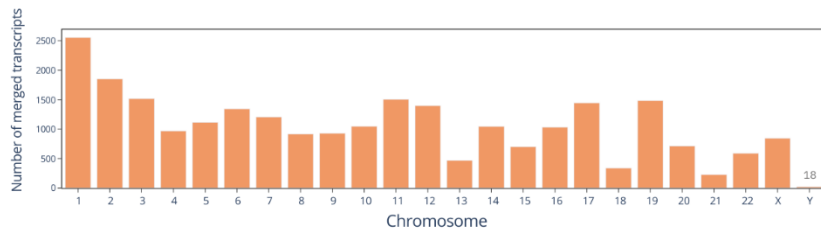


Figure 33. Number of merged transcripts predicted in each human chromosome.

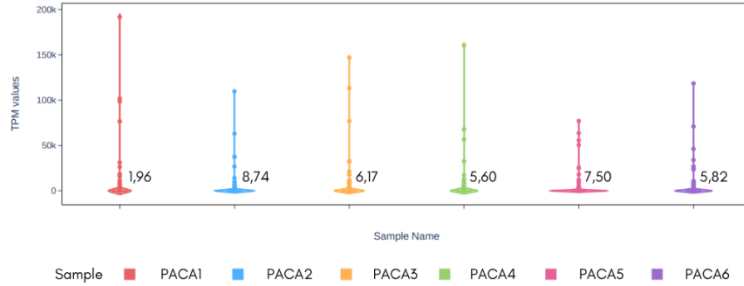


Figure 34. Violin plot showing the distribution of TPM values obtained for each merged transcript and in each sample. Median values are shown within the plot.

We downloaded the DNA sequence of all the transcript isoforms and reconstructed their coding DNA based on the exon coordinates obtained from the StringTie merge mode. Later, to end with a set of amino acid sequences we performed an *in-silico* translation of the 25.207 predicted coding sequences. For this dataset version 1, only the ATG codon was used as a starting point for the translation.

5.3.3.1.3 *In-silico* translation of coding DNA

Through the 3-frames translation of coding DNA sequences we obtained a set of 838.377 candidate small ORFs that range between 7 and 100 amino acids lengths (median size 19 aa). Small ORFs were distributed across all human chromosomes, being 1, 2, 3, 11 and 5 the ones with the highest numbers of short aa sequences (79.508, 71.214, 61.505, 49.530, 45.566 respectively). In contrast, even though chromosome 19 was on the top 5 regarding the number of predicted transcripts as chromosomes 1,2,3 and 11, only 27.401 smORFs were translated within it (Fig 58). The predicted transcript MSTRG.14380.4 identified in chromosome 8 was the cDNA with the highest number of candidates smORFs: 1.537 short amino acid sequences. Note that this predicted transcript was also the longest coding sequence obtained after the merging step, with 161.831 nucleotides.

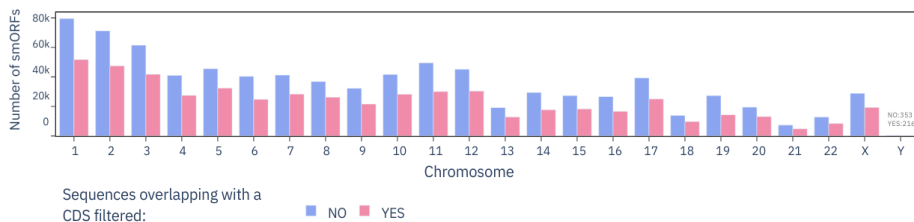


Figure 35. Number of smORFs translated from each human chromosome. Bars in blue show the results after in-silico translation while those in pink correspond to smORFs obtained after filtering them.

5.3.3.1.4 Filtering micropeptides based on their overlap with known CDS

Lastly, to end with a collection of candidates novel micropeptides we filtered out those overlapping with annotated coding sequences. Only the overlap with CDS was considered since we choose to also study smORFs located in 5' or 3' UTRs, introns or long non-coding RNAs. To do so, we performed a Blastn of all the smORF sequences against the human coding sequences. Only those results from Blastn with an e-value lower than 0'001 and an overlap lower than 60% between their sequence and an annotated CDS were kept.

We end with a dataset of 551.494 candidate micropeptides, which represent our dataset version 1. In line with the results obtained in previous steps, chromosome 1, 2 and 3 had the highest number of candidates micropeptides (Fig 58) and their median size was 19 amino acids. MSTRG.14380.4 was the predicted transcript with the highest number of candidates micropeptides (1.342). However, a median of 13 smORFs were identified in each of the predicted transcripts analyzed. An example showing the predicted transcript (MSTRG.12.1) all the smORFs translated and those that were selected after the filtering step is shown in Figure 59.

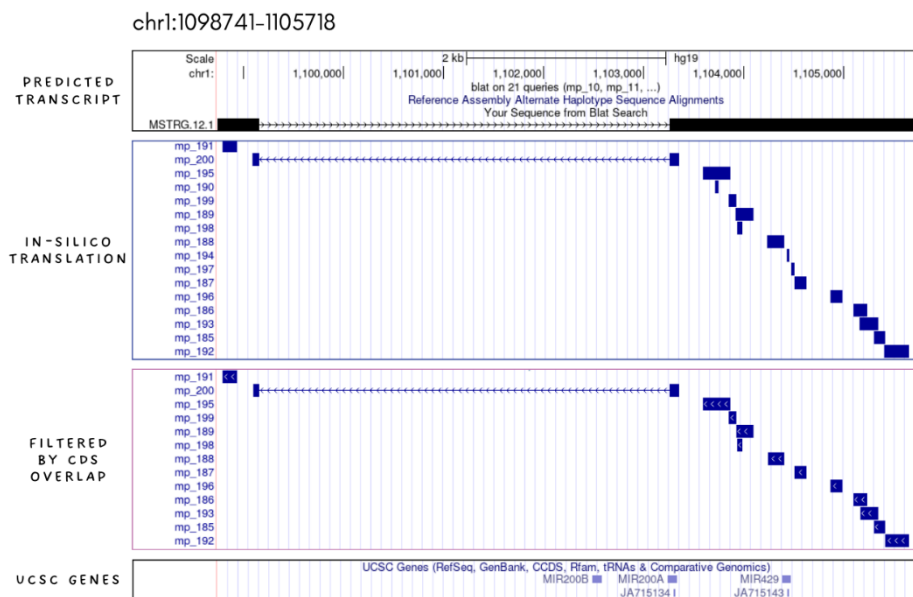


Figure 36. Candidate small ORF identified in one predicted transcript. MSTRG.12.1 (transcript name), located in chr1:1098741-1105718 is shown above the figure in black. smORFs obtained after translation (blue square) and those that do not overlap with known CDS (pink square) are shown below. The transcript was identified in a highly intergenic region.

5.3.3.2 Dataset version 2: inclusion of non-canonical start codons and expression-based filtering for small ORFs

Once we finished the first dataset of small ORFs, we designed a second version with the aim of addressing specific limitations and enhancing our approach based on the insights gained from our initial version. One consideration was the need to adjust certain parameters within the StringTie algorithm. Moreover, we opted for an in-house merging strategy as we encountered uncertainty regarding the criteria used on the merge mode of StringTie. At this point of the study, motivated by the desire to capture a broader range of candidate smORFs, we also decided to expand the set of start codons to include non-canonical ones. However, this expansion was addressed through more stringent filters to effectively reduce the candidate collection. This ensures its

compatibility with mass spectrometry analyses, that results in higher numbers of false positives and less sensitivity when large datasets are used.

5.3.3.2.1 De novo transcript prediction based on unique mapped reads

We started again from the prediction of transcripts based on de-novo assembly and using StringTie algorithm, However, in this case, we did not allow multi-mapped reads (-m 0,1) and only unique mapped reads were considered to calculate the coverage of each transcript. The minimum size length to define a transcript was not changed (-M 50bp). After running StringTie, we get a total of 90.651, 100.963, 61.758, 176.703, 136.574 and 182.098 predicted transcripts for pancreatic adenocarcinoma samples (from PACA1 to PACA6). A summarized description about the size of the transcripts, number of exons, reads covering each sequence, and distribution across chromosomes is shown in Supplementary figure 4.

5.3.3.2.2 Consensus set of predicted transcripts applying an in-house merging strategy

When using StringTie merge mode, the decisions it made to define the combined transcriptome representing all samples were not known nor controlled. For this reason, we defined our merging strategy to end with a consensus set of transcripts including all samples analyzed. Decisions and the criteria considered to define our merging strategy are explained in Methods and Results section. To do so, first we clustered transcript isoforms based on their start and end coordinates, strand and number of exons. Considering the results obtained for the six PACA samples, we obtained 589.475 clustered isoforms, with 82% of them being single-exon transcripts. Once more, chromosomes most represented were 2, 1 and 3 with 53.755, 50.212 and 41.018 clustered isoforms respectively. On this set of predicted transcripts, 492.732 were only predicted in one PACA sample,

and 4.719 in all of them. Since differences on the genomic coordinates were observed across clustered isoforms identified in different samples, we redefine transcript sequences applying different rules to obtain a consensus sequence. After recalculating their genomic coordinates, some isoforms appeared to be replicated and therefore, were deleted. Because of this, the number of consensus isoforms we obtained (589.145), was slightly lower than the number of clustered transcripts. Finally, isoforms were filtered regarding their prior identification in RNA-seq samples to end with a representative set of transcripts. The filter applied to single-exon transcripts was stricter than for the rest, given the small minimum prediction size (50bp) used in StringTie, which does not require split reads but only a cluster or reads aligned to a region. The combination of the six tumor samples allowed us to obtain 27.849 transcripts (Fig 60), where 4.691 were predicted in all pancreatic adenocarcinoma samples analyzed and 225 (0,81%) of them were single-exon transcripts. This was our transcriptome for the second dataset version.

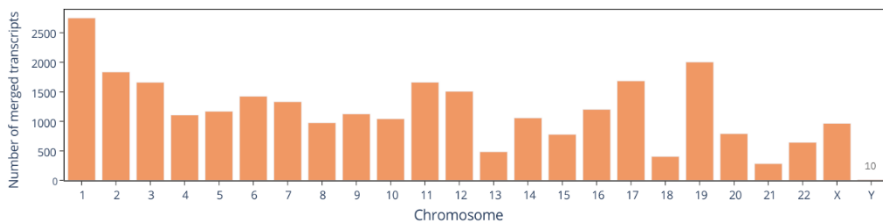


Figure 37. Distribution of merged transcripts (27.849) across all human chromosomes.

For all consensus and filtered transcripts, we analyzed their expression values in all pancreatic adenocarcinoma samples. To do so, we use a specific StringTie function that allows to get the abundance in coverage and TPM values. StringTie could not identify supporting RNA reads for 118 transcripts in PACA1 sample, 17 in PACA2, 110 in PACA3, 53 in PACA4, 54 in PACA5 and 76 in PACA6.

Expression values in TPM had different ranges depending on the tumor sample, where median values range between 1,6 (PACA1) and 10,13 (PACA2). We could also observe differences when compared the expression values observed in those predicted transcripts (4.387) identified in all RNA-seq samples (Fig 61). In particular, sample PACA1 had the lowest expression values for these transcripts (median TPM value 3,09). Not only the overall sample expression was different across them, but also when compared TPMs obtained for each particular predicted transcript. These differences can be seen in Figure 62, representing TPM values for a subset of 3.048 merged transcripts identified in all samples.

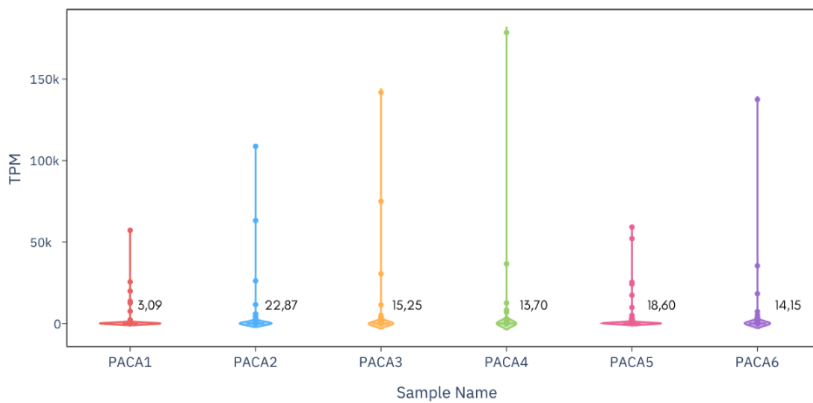


Figure 38. Violin plot showing the distribution of TPM values across six RNA-seq samples. Only the expression values of those transcripts predicted in all samples are represented.

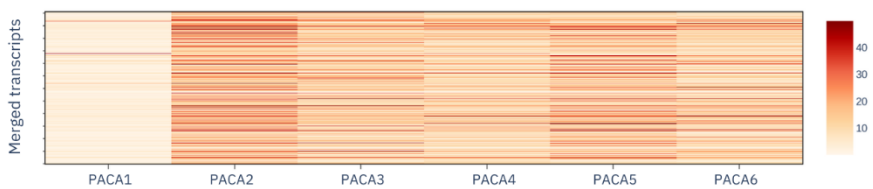


Figure 39. Expression values in TPM observed for a subset (3.048) of merged transcripts predicted in all samples. For each sample (x axis) and transcript (y axis), TPMs are represented in a range of orange colors.

5.3.3.2.3 In-silico translation of coding DNA considering non-canonical start codons

To perform a 3-ORF *in silico* translation, DNA sequences were downloaded, and the coding sequences were reconstructed considering the exon-exon coordinates. From 27.849 transcripts we generated 28.040 coding sequences, since for isoforms with unknown strand their CDS was analyzed in both forward and reverse strands.

Previous studies have shown that ATG is not the unique codon able to initiate translation in humans, but other three-nucleotide combinations too. Although it is known ATG is the most frequently found start codon, for this *in-silico* translation we also consider TTG, CTG, ATT, GTG and ACG. Sequences starting from any of these 6 codons, and with a length between 7 and 100 amino acids until the first were defined as candidate micropeptides. At this stage, we had a collection of 6.366.662 small ORFs, which was over seven times larger than the set of candidate smORFs obtained through *in-silico* translation for version 1. Within this collection of candidates, 82,7% peptides start from a non-canonical start codon. A total of 628.106 candidate micropeptides were identified in chromosome 1, followed by 446.790 and 442.412 in chromosomes 2 and 3, being the top three chromosomes with highest numbers. Contrary, chromosome Y had the lowest number of micropeptides; 2.202 (Fig 63). Transcript isoform MID_6525_1 was the sequence with the highest number of candidate short ORFs, 10.063. It was identified in chromosome 3, from 52.578.244 to 52.719.743 genomic coordinates and formed by 32 exons, resulting in a coding sequence of 9.756 nucleotides. Note, that candidate micropeptides translated from a specific host-transcript could overlap between them. The predicted transcript covered different known protein coding

genes, including *PBRM1*, *GNL3*, *SPCS1* or *NEK4*. This isoform was not observed in the previous version 1.

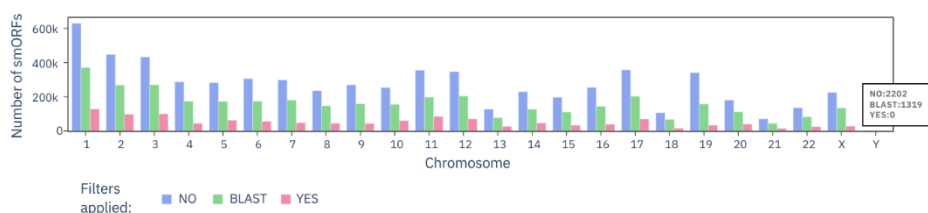


Figure 40. Number of smORFs across human chromosomes after in-silico translation (blue), only when those overlapping with know CDS were excluded (green) or also expression filter was applied (pink).

We then analyzed the type of genomic regions where these candidate micropeptides were located, based on the genomic coordinates of all human annotated genes. From our set of candidate micropeptides, 32,8% were completely located within exons of protein-coding genes, 23,2% in 3' UTR regions, 18,7% in introns and 13% in exons of non-coding genes. Moreover, 4% had part of their sequence overlapping with a coding and around 5% of the candidates did not overlap with any annotated gene suggesting they were in intergenic DNA. According to gene type, 4.381 candidate micropeptides were found overlapping with polymorphic pseudogenes, 2.701 with immunoglobulins, and 1.339 with t-cell receptor genes.

5.3.3.2.4 Filtering micropeptides based on their expression and overlap with known CDS

Considering the size of dataset version 2, we evaluated how to reduce the collection of candidates to get a compatible dataset to perform MS analysis. Compared to dataset version 1, we applied a combination of more stringent filters. Not only did their overlap with known annotated coding sequences but also the expression of the host transcript was taken into account.

Nucleotide sequences of candidate micropeptides were aligned against all annotated ensembl CDS using Blastn. We considered as a good local alignment those results from blastn with an e-value lower than 0'001. The sequence of around 35% candidate micropeptides, overlap more than 90% with a known CDS from Ensembl, whereas 39% overlap more than 50%. Therefore, most of our candidate micropeptides share less than half of their sequence or nothing with a known coding sequence. To end with a smaller dataset containing mostly novel micropeptides in non-annotated coding regions, candidates overlapping more than 30% with any CDS were filtered out. A collection of 3.733.227 candidate micropeptides were obtained, reducing the dataset almost half of its previous size. An example of a predicted transcript (MID_5_1) located in chromosome X and the smORFs selected or excluded after applying this filter is shown in figure 64.

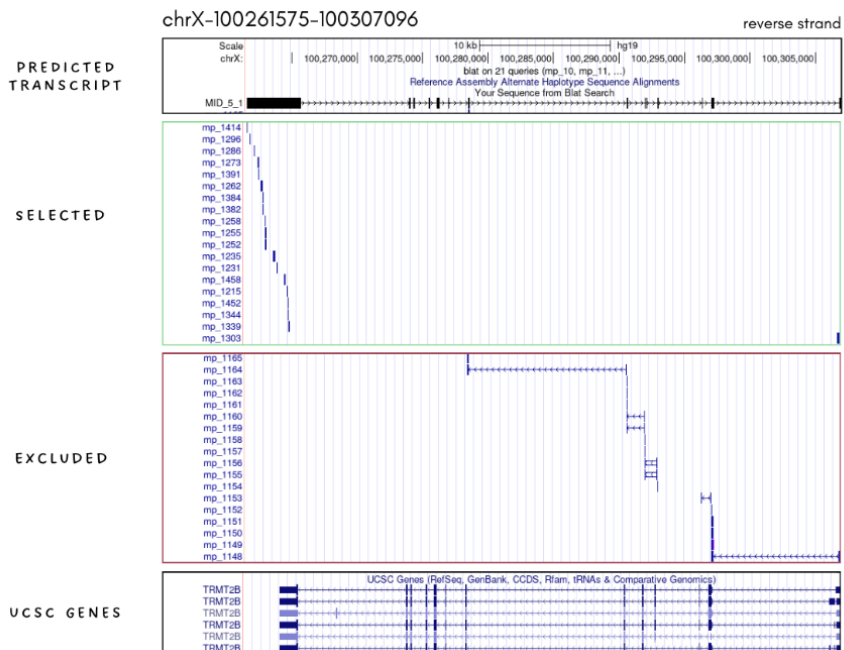


Figure 41. Example of a predicted transcript (MID_5_1) and translated sequences located in chromosome X. Small ORFs selected (green box) after excluding those (red box) that overlap with known CDS are shown. Only few smORFs translated from this transcript are represented.

Finally, we applied a second filter based on expression results that were previously obtained for the host-transcripts and annotated for each candidate micropeptides. We decided to sort out only those short amino acid sequences resulting from a transcript expressed in all six samples. Moreover, its levels of expression should be higher than the median value observed in each pancreatic adenocarcinoma sample. We then end with the final dataset version 2, which enclosed 1.211.051 candidate micropeptides. This dataset was also used as a reference on mass spectrometry analysis.

The highest number of candidates micropeptides in dataset version 2 was identified in chromosome 1 (127.896), followed by chromosome 3 (99.650) and chromosome 2 (97.572) (Fig 63). Candidate micropeptides had a median size of 19 amino acids that matched with the median size previously characterized for smORFs located in intergenic and non-coding regions as well as in UTRs, the three main classes identified in our datasets (Couso & Patraquim, 2017). Moreover, 84% of them started by a non-canonical start codon being CTG the most observed start (268.073 candidates). Almost all candidate micropeptides (97%) were translated from one single-exon of a predicted transcript, whereas the remaining covered between 2 and 4 exons. Note that these exons do not correspond to those already known from protein coding genes but from the transcripts predicted through *de novo* assembly. The predicted transcript MID_6525_1 still was the one with the highest number of candidates micropeptides identified (6.448). Although the majority of the candidates obtained in DS2 were located within protein coding genes (1.103.297), only 2,7% of them had part (less than 30%) of their sequence overlapping with a coding exon. Almost half (43,4%) of the candidates obtained in this dataset were located in 3'UTR regions and in contrast, only 1,2% in 5'UTRs. Finally, we observed a decrease in the percentage of candidate micropeptides located in intergenic regions, that represented 4% of this dataset. Although these candidates were not excluded because of overlapping with known CDS, they had

low expression values and were not represented in all PACA samples. Therefore, after reducing the dataset 1% of the intergenic smORFs previously translated were filtered out.

Together, both datasets are a profitable catalog of candidate smORFs derived from mRNAs expressed in PACA samples and, enriched in non-annotated CDS regions to use for MS/MS analysis.

Discussion of chapter 3 (study 1) starting in section 6 (page 263).

Study 2: Identification of candidate highly conserved micropeptides in intergenic regions

5.3.4 UNICORNs: highly evolutionary constraint intergenic regions

At the beginning of this second study focused on novel micropeptides, most known and published small ORFs were identified in annotated genes. Therefore, we aimed to evaluate less explored DNA sequences such as intergenic regions. We considered a potential approach to identify novel small ORFs was focusing our search on conservation features that could indicate functionality of a sequence. Formerly, the first analysis from The Zoonomia Project was published (5). The Zoonomia Project investigated the genomics of shared and specialized traits in eutherian mammals. By prioritizing making data available, quickly and without restriction, the project supported biological discovery, medical research and the conservation of biodiversity. Among all shared data available on their web page, conservation scores could be downloaded or inspected through specific tracks on the UCSC Genome Browser. These conservation scores were calculated using PyhloP from the Zoonomia whole-genome alignment (v2) of 240 species comprising representatives from more than 80% of mammalian families. The scores were used to identify sites and regions under purifying selection (3,1% in the human genome) including unannotated intergenic regions. UNICORNs were therefore defined as non-coding regions non-annotated in ENCODE3 showing high evolutionary constraints, that could suggest function.

A total of 424.179 UNICORNs were downloaded in GRCh38, distributed across human chromosomes 1 to 22. Highly evolutionary constraint sequences in sexual chromosomes (X and Y) were not provided but excluded on the Zoonomia analysis. Although larger chromosomes tended to have higher numbers of

detected UNICORN, chromosomes 2, 4 and 5 were the top 3 (Fig 65). Therefore, the number of UNICORNS in each chromosome seemed to be correlated with the size of their intergenic DNA. UNICORNS range between 11 and 1.325 nucleotides length, with a mean size of 38 nt (Fig 66).

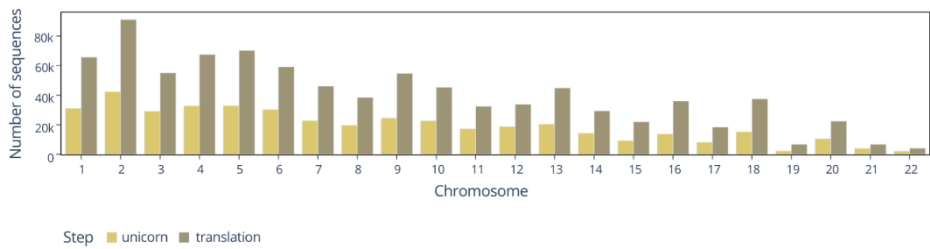


Figure 42. Sequence distribution across chromosomes. Number of UNICORNS (light yellow) and in-silico translated sequences (dark yellow) in each human chromosome.

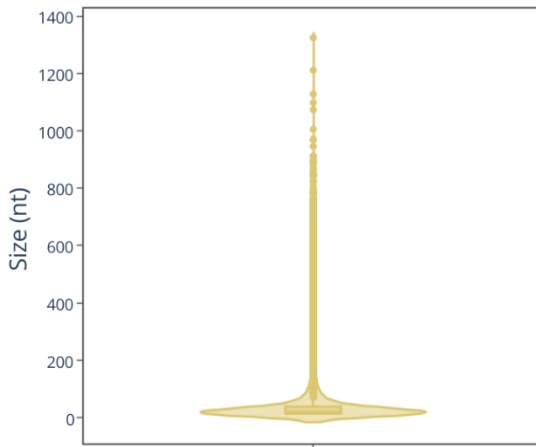


Figure 43. Violin plot showing the range of sizes in nucleotides for all UNICORNS (n=424.189 sequences).

5.3.5 In-silico translated small ORFs located in intergenic regions

In order to identify all possible short amino acid sequences located in highly conserved intergenic regions, we artificially translated all DNA sequences previously defined as UNICORNs by the Zoonomia Project. We did a strongly permissive translation to comprise a wide range of candidates. Nucleotide sequences were translated 6-ORF meaning we read them from the first, second and third nucleotide and in both forward and reverse strands.

In-silico translation of 424.179 UNICORNs resulted in a list of 887.676 amino acid sequences with a length between 10 and 100 codons, considering the established threshold (100aa) used to define micropeptides. Their distribution across chromosomes is shown in figure 65. Almost all (91,39%) of these sequences were shorter than 30 aa and had a mean size of 17,63aa. Only 11 translated sequences reach the maximum size (100 aa). Translation started from the canonical ATG codon in 2,2% of the sequences, being the trinucleotide TTT the most recurrent start codon (5,2%) followed by AAA (4,75%) (Fig 67). Similar percentages were obtained for translated sequences ending with a stop codon (51,3%) or because of the UNICORN termination (48,7%). Also, when we counted sequences translated from forward (444.500) and reverse (443.175) strands or classified depending on the translation starting nucleotide (ORF1 304.291 aa sequences; ORF2 295.873; ORF3 287.511).

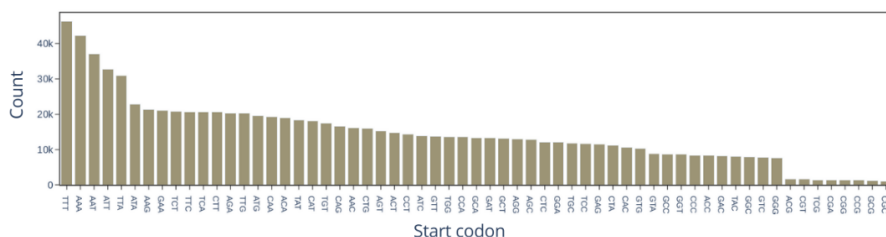


Figure 44. Number of in-silico translated sequences depending on their start codon.

Considering the initial set of UNICORNs, 281.066 of them did not result in any translated sequence because of their size. All these UNICORNs were shorter than 32 nucleotides. Furthermore, shorter UNICORNs and particularly those with an approximate size of 35 nt resulted just into one translated sequence. In contrast, we could identify higher numbers of small peptides from larger conserved intergenic sequences. Although it was not the largest UNICORN, 78 different amino acid sequences were translated from a region located in chromosome 2 between 155.728.521 and 155.729.733 genomic coordinates (1.121 nt length). Summarizing, the number of translated sequences per UNICORN was also correlated with their size in nucleotides.

5.3.6 Candidate ortholog sequences of translated intergenic small ORFs

As these 887.676 candidate small ORFs were short amino acid sequences artificially generated from conserved intergenic regions, more evidence was needed to assume or suggest this micropeptide sequences could have a functional role in humans.

For this reason, we first evaluate the orthology between the in-silico translated human short amino acid sequences and mice genome. Orthology between human and mice sequences could suggest the protein sequence is relevant for the organism and likely functional. We used the Reciprocal Best Hit approach to define pairs of orthologs between both species. Therefore, two complementary tblastn analysis were performed comparing human short amino acid sequences against the reference mice DNA and short peptide sequences identified in mice versus the reference human DNA.

From the first tblastn analysis 887.676 human short aa sequences were compared with mice genome. We obtained 1.104.502 local alignments from

283.608 different human smORFs candidates, 32% of our initial set. However, only 61.585 alignments from 56.350 human intergenic smORFs passed the e-value ($< 1e-05$) and coverage ($>50\%$) filters defined to consider a good local alignment. We did not have translated sequences from 403.970 initial UNICORNs after applying this filter. Around 8% of the amino acid sequences had two or more matches with different regions of the mice genome and many (47,84%) sequences were completely aligned. At this point, sequences were distributed across chromosomes similarly to in previous steps. Human peptide sequences were aligned in all mice chromosomes (Fig 68). A total of 1.516 translated sequences were aligned within sexual chromosomes X and Y, and 37 of them in unplaced scaffolds including GL456233, GL456379, GL456382 and JH584296.

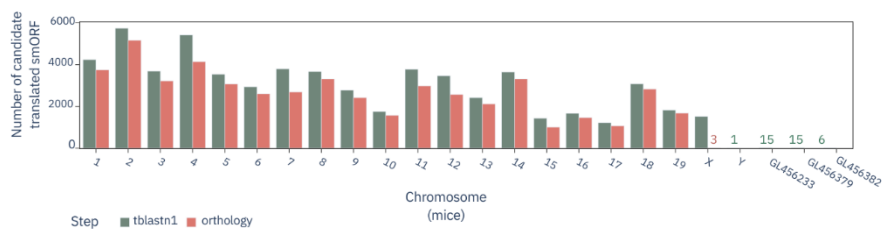


Figure 45. Number of human translated smORFs aligned across mice chromosomes after *tblastn1* (green) and once ortholog sequences were defined (orange).

The amino acid sequences (61.585) selected after the blast filtering step were used for the second *tblastn* analysis. In this case, we compared mice peptide sequences with the human reference genome. After TBLASTN2, 56.304 sequences were aligned in at least one human genome region without considering any filter (58.316.612 alignments). We did not have results for 55 mice peptide sequences.

After this second alignment we define the set of orthologs. Following the definition of RBH, two sequences from different genome species are considered orthologs if align each other as the best hit in the other genome. In this project we only retained 1:1 orthologs, so peptide sequences that aligned in more than

one human or mice genomic region were discarded. Therefore, we inspected the 56.304 sequences aligned in TBLASTN2 and looked for their alignment in TBLASTN1. Combining an automatic search with manual inspection of alignments, we defined a set of 1:1 orthologs counting a list of 50.936 candidate smORFs translated from 18.658 different UNICORNs.

Ortholog sequences were identified and distributed across human chromosomes similarly than in previous steps (Fig 69), and a comparable proportion of candidate smORFs was obtained from forward (25.435) and reverse (25.501) translation. Peptide sequences aligned in mouse Y chromosome, or in unplaced scaffolds did not pass the criteria used to select orthologs (Fig 68). Regarding the size in amino acids of these candidates, we calculated a mean value of 35,26 aa, that was higher compared to those 887.675 aa sequences obtained after performing the in-silico translation (mean=17,63) (Fig 70). These values suggested smaller peptides tended not to fulfill orthology selection criteria, nor had good alignment scores. However, upon examining each sequence's size individually, the majority did not align fully, resulting in a shorter aligned region compared to the initial sequence. Consequently, the orthologous sequences had a smaller size relative to their initial translated sequence since we only evaluated orthology between the aligned regions but not the complete peptides obtained from in-silico translation.

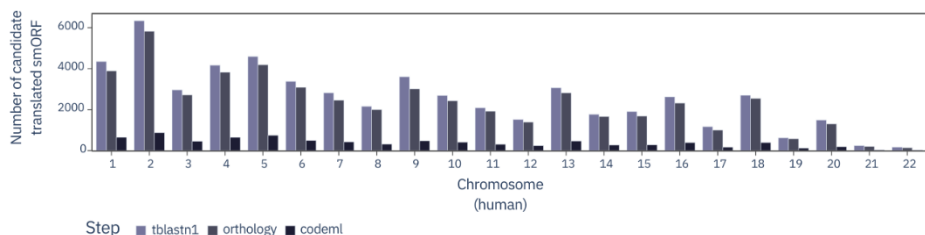


Figure 46. Number of candidate translated smORFs identified in each chromosome. Sequences after the first tblastn (dark blue), once orthology was defined (medium blue) and those selected considering the dn/ds ratio (dark blue) are shown.

Ortholog sequences were then used to calculate the ratio of non-synonymous to synonymous variants between human and mice sequences.

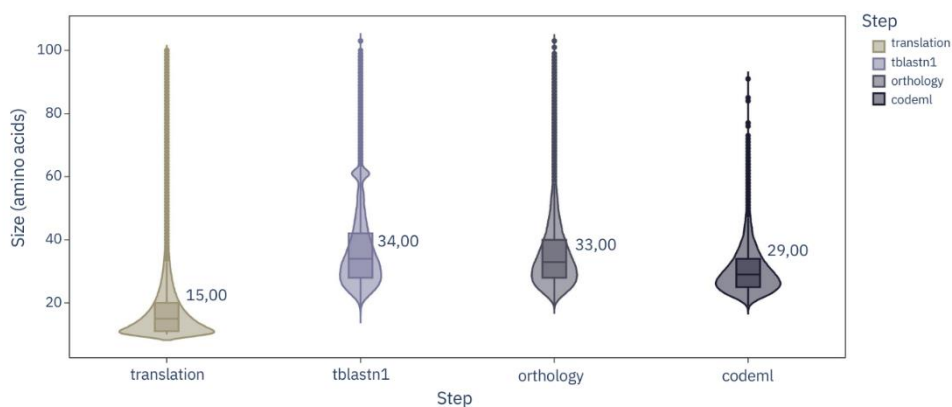


Figure 47. Violin plot corresponding to the range of peptide sizes (aa) of all smORFs selected in each step. Median is shown within the plot.

5.3.7 Calculated dn/ds ratio on known protein-coding genes

In line with the definition of the dn/ds ratio, ortholog pairs with ratios close to 0 are typically under purifying selection and therefore, they are supposed to maintain their role across species. So as to select candidate functional micropeptides from our set of ortholog sequences (50.936) we first established a more precise threshold. Accordingly, we calculated the dn/ds ratio for 300 coding exons shorter than 1000bp to simulate small peptide sequences.

Before doing this calculation, we identified their mouse orthologs pairs following the established methodology and criteria previously used for candidate smORFs. We in-silico translated the nucleotide sequences of 300 coding exons, performed reciprocal alignments comparing peptide and nucleotide sequences from human and mice, and selected 1:1 orthologs. We discarded 11 coding exons from the 300 randomly selected because of their short length (3 bp). After all, we ended with a list of 71 one-to-one ortholog pairs that range from 65 to 781 bp

CDS length. After this selection of orthologs, the number of known coding sequences inspected decreased almost 4 times since 155 regions did not properly align to mouse genome (e-value < 1e-05 and sequences overlap > 50%), and 63 did not satisfy the criteria to define them as 1:1 orthologs. Among these 71 orthologs, 22 corresponded to coding exons from cancer genes. They were distributed across all human chromosomes except 13, 18, 20, 21 and Y, and had a median size of 137 coding bp (min 67, max 781).

Finally, we used the Codeml function of the PAML package to calculate the dn/ds ratio for each ortholog pair. We could not measure this ratio for 12 pairs because of the presence of nucleotide gaps in human or mice sequences. The dn/ds ratios obtained for the remaining 59 coding exons range from 0,001 to 0,8014 (mean=0,12) (Fig 71). Notably, three coding regions exhibited the highest dn/ds values and therefore not closer to 0, had a rate of synonymous substitutions per synonymous site (ds) equal to 0. Accordingly, these human sequences did not have synonymous substitutions when compared to their mouse ortholog. It is likely that we did not detect synonymous variation due to the short length of the analyzed regions, since it is rare and somewhat unlikely in real evolutionary cases of protein coding genes.

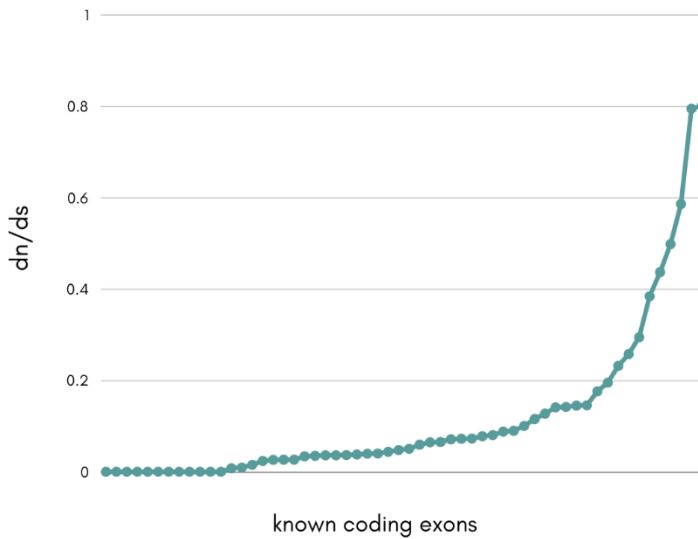


Figure 48. Calculated dn/ds ratio for 59 coding exons and their ortholog pairs in mouse.

Based on the dn/ds ratio (non-normal) distribution of 59 coding exons, we established a threshold to detect outlier values as $Q3 + 1,5 * IQR$ that resulted in 0,3179. Subsequently, we decided to use a dn/ds below 0,32 together with a ds value exceeding 0,1 to consider short peptide sequences as potentially functional micropeptides in nature.

5.3.8 Catalog of candidate intergenic micropeptides from highly conserved regions

The results obtained from known coding exons allowed us to define a dn/ds threshold to later select candidate micropeptides indicating functionality. We applied codeml to calculate the ratio of non-synonymous to synonymous variation in ortholog pairs, including 37.293 previously obtained sequences. Note that we could not calculate the ratio for 13.643 orthologs because of nucleotide gaps in human or mouse aligned sequences. We reject those short peptides that did not pass the dn/ds criteria and end up with a set of 8.289 candidate novel micropeptides located in 6.536 different highly conserved intergenic regions

5.3.9 Preliminary evidence of expressed candidate functional micropeptides

At this point of the study, the analysis previously performed allowed us to identify short peptide sequences under purifying selection and located in intergenic regions highly conserved between species. This information suggested functionality of the peptide, but more proofs were needed to confirm their presence in nature. For this reason, we decided to look for evidence of expression of candidate functional micropeptides in healthy tissues.

We searched for aligned reads covering each candidate micropeptide (8.289) in 135 randomly selected samples comprising 27 different healthy tissues (5 samples per tissue). Next, paired-end reads were filtered out to avoid low-quality and multi-mapped sequences. Also, PE reads where at least one pair overlapped with a known transcripts including non-coding RNAs were excluded. We decided to require at least 5 RNA-seq reads covering the candidate to consider it had signals of expression. Therefore, we could observe 249 candidate micropeptides showing signals of expression in a minimum of one RNA-seq sample, and 13 candidates in 10 samples or more including different tissue types. A total of 29 candidate micropeptides had signals of expression in 60% of the samples (3 out of 5 RNA-seq) for at least one healthy tissue (Table 12), with median coverage values observed across tissues that range from 3,8 to 32. The candidate micropeptide with highest median coverage was located in chr1: 37.099.960-37.100.043 and signals of expression were only detected in healthy muscle samples. Interestingly, we noticed the candidate micropeptide identified in chr5: 93.615.953-93.616.054 had signals of expression in 103 different samples, including all tissue types (26) except muscle (Table 13). This was the candidate with evidence of expression in more samples. This intergenic conserved region is located 1700bp upstream a known lncRNA (*FAM172A*). Even though it

was close to a known gene, we did not detect mate reads of supporting PE within the lncRNA. These preliminary results suggested that the conserved smORF was not part of the known transcript but a different candidate gene.

Candidate micropeptide	Complete tissues	Median coverage
chr14_46837398_46837472	1 Brain	9
chr14_60600471_60600557	1 Pituitary	8,2
chr14_76096277_76096354	3 Brain, Nerve, Pituitary	5,2
chr5_64614555_64614653	8 Adrenal Gland, Blood Vessel, Brain, Esophagus, Pituitary, Prostate, Salivary Gland, Stomach	9,7
chr5_51839816_51839917	1 Ovary	12,6
chr5_93615953_93616054	26 Adipose tissue, Adrenal Gland, Bladder, Blood Vessel, Brain, Breast, Cervix Uteri, Colon, Esophagus, Heart, Kidney, Liver, Lung, Nerve, Ovary, Pancreas, Pituitary, Prostate, Salivary Gland, Skin, Small Intestine, Spleen, Stomach, Thyroid, Uterus, Vagina	18,4
chr13_96087557_96087655	1 Pituitary	5,4
chr8_10862241_10862342	1 Brain	11,4
chr8_10862239_10862343	1 Brain	11,4
chr13_53078789_53078884	1 Salivary Gland	14
chr13_53078850_53078927	1 Salivary Gland	14
chr2_206936805_206936885	2 Pituitary	7
chr6_155829957_155830046	1 Thyroid	9
chr6_155830067_155830168	1 Thyroid	8
chr5_127035507_127035587	1 Small Intestine	3,8
chr10_92682836_92682967	2 Pancreas, Thyroid	7,4
chr11_65808271_65808354	5 Lung, Ovary, Pituitary, Prostate, Thyroid	5
chr1_119017757_119017822	2 Breast, Prostate	4,5
chr1_118875161_118875268	1 Muscle	6,6
chr1_48049010_48049105	1 Kidney	6,2
chr1_37099960_37100043	1 Muscle	32
chr9_116036606_116036713	1 Skin	5
chr4_84527263_84527343	3 Lung, Salivary Gland, Thyroid	31,8
chr7_24173168_24173239	1 Kidney	13,2
chr7_26652489_26652614	1 Spleen	3,8
chr10_123188358_123188456	1 Pituitary	30,8
chr10_123188430_123188537	1 Spleen	4,6
chr4_181739529_181739597	1 Adrenal Gland	20
chr11_115167714_115167806	1 Brain	6,6

Table 12 - Candidate micropeptides (chr_start_end), with signals of expression in at least 60% RNA-seq samples (3 out of 5) of a healthy tissue. Median coverage considering the observed values across all samples where signals were detected is calculated and shown in the last column, colored from red (low coverage values) to blue (high coverage values).

Adipose Tissue	19	Nerve	26,2
Adrenal Gland	14,4	Ovary	16,2
Bladder	14,6	Pancreas	7,8
Blood Vessel	15,2	Pituitary	24
Brain	31,6	Prostate	19,2
Breast	34,2	Salivary Gland	26,2
Cervix Uteri	22,4	Skin	7
Colon	17,2	Small Intestine	24,8
Esophagus	17,4	Spleen	22,8
Heart	5,4	Stomach	16,6
Kidney	17,8	Thyroid	15,4
Liver	5,2	Uterus	23,8
Lung	22,6	Vagina	20,4

Table 13 – Number of reads (mean per tissue) covering candidate micropeptide located in chr5 from 93615953 to 93616054. Colors ranging from lower (red) to higher (blue) values.

5.3.10 Detection of significant clusters of somatic cancer mutations in published smORFs

Intending to explore the role of micropeptides in cancer disease and tumorigenesis, we analyzed clusters of somatic single nucleotide variants from the ICGC cancer genomes in small ORFs. So as to start evaluating the performance of a driver discovery algorithm (OncodriveCLUSTL) in small genes, and while working on the identification of intergenic and novel smORFs, we first analyze a published database (SmProt) of smORFs including 49.065 short peptides.

We run OncodriveCLUSTL for each set of SNVs classified depending on the ICGC project using 6 different parameter combinations (see Methods, Table 4). We then selected the most adjusted combination for each ICGC set of variants based on the KS test and the enrichment in cancer genes, as authors of the algorithm suggested to us (Supplementary Table 4, Supplementary Fig 5). Also, manual inspection of the qq-plot obtained from OncodriveCLUSTL was done.

Independently of the parameter combination, 6 ICGC projects were excluded due to low number of smORFs with clustering signals (< 20) or significant differences and inflation between the expected and observed p-values (BOCA-UK, BTCA-SG, GACA-CN, LUSC-US, OV-AU and UTCA-FR) (Fig 73). Therefore, these ICGC projects were excluded as we could not calculate their KS value.

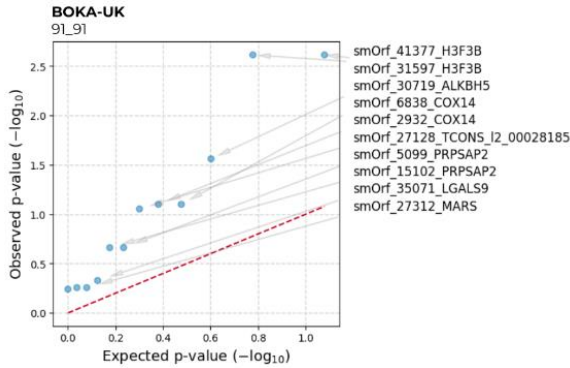


Figure 50. QQ-plot obtained from OncodriveCLUSTL. Example of one ICGC project (BOCA-UK) excluded due to low number of clustering signals (blue dots). The ID name for each small ORFs with detected clusters is shown at the right side.

We could identify significant clusters (q-value < 0,01) in small ORFs for 4 different ICGC projects including ESAD-UK, PACA-AU, PACA-CA and LUSC-KR (Fig 74).

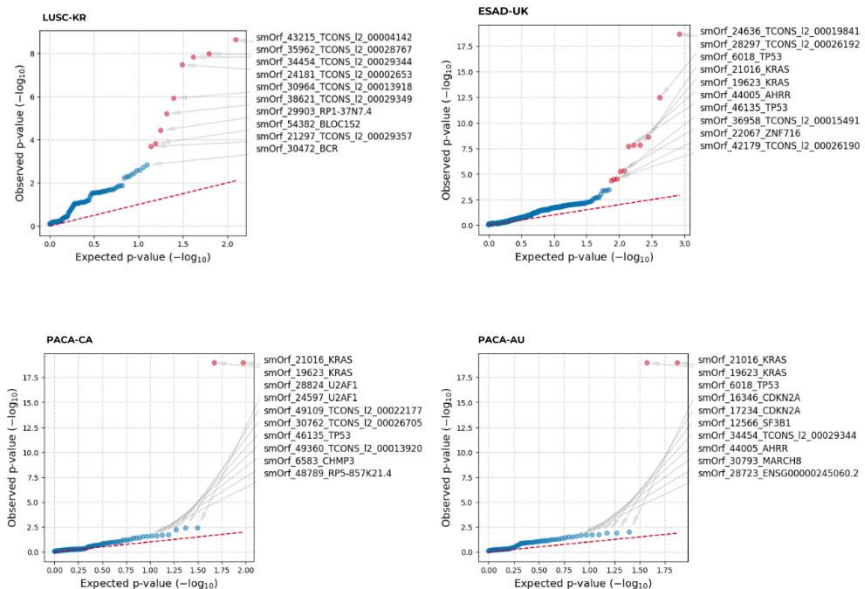


Figure 51. QQ-plots obtained from OncodriveCLUSTL for ICGC projects with significant clusters. Expected and observed p-values are shown for all smORFs with clusters of mutations. SmORFs (dots) colored in red have significant p-values suggesting they are drivers.

In pancreatic adenocarcinoma projects (PACA-AU and PACA-CA) significance was only observed in two smORFs (smOrf_19623_KRAS and smOrf_21016_KRAS)(Table 14). Nevertheless, both smORFs overlapped with KRAS, a known cancer driver gene. Clusters of variants within these smORFs coincide with previously identified significant clusters in the IntOGene study (<https://www.intogen.org/search>). On the other hand, a higher number of significant clusters and therefore, smORFs were detected when we analyzed ESAD-UK (11 significant smORF) and LUSC-KR (9). For these two ICGC projects, significance was not only observed in smORFs overlapping with cancer genes such as KRAS or TP53, but also in lncRNAs annotated by the NONCODE database(204). Based on the Gencode v37 annotation, six smORFs were in intergenic regions. One small ORF (smOrf_24636) located in *LINC00879* (Gencode v37 annotation) had significant clusters of mutations in ESAD tumors. Altered frequency of this lncRNA due to amplification was previously identified in LUSC tumors, suggesting it could be implicated with tumorigenesis (205). Differential expression between tumor and normal genomes were also observed in the GEPIA data portal (<http://gepia.cancer-pku.cn>) for other genes that contained driver-smORFs identified in our analysis. As an example, candidate smOrf_28297 was located within *RP11-274B21.14*, which is highly expressed in Acute Myeloid Leukemia tumors, and smORF_32101 that was located within *ZNF716* a zinc finger protein also highly expressed in Testicular Germ cell tumors.

Together, these results suggested that we could use OncodriveCLUSTL for driver identification of small ORFs. Although we did not have significant results for smaller sets of SNVs, we could characterize 13 different smORFs located in lncRNA or non-annotated regions as driver genes.

ICGC project	smORF	Cancer gene	Chromosome	Start	End	Strand	Length	Mutations	Clustered mutations	Clusters	P-value empirical	Q-value empirical	P-value analytical	Q-value analytical	NONCODE database	Gencode v37
ESAD-UK	smOrf_24636	False	3	94857157	94893419	+	36263	264	195	49	0,001	0,04912	2,22E-19	1,85E-16	TCONS_I2_0001984	LINC00879
	smOrf_28297	False	7	128173707	128231906	+	58200	57	18	8	0,001	0,04912	3,54E-13	1,48E-10	TCONS_I2_0002619	RP11-274B21.14
	smOrf_6018	True	17	7572929	7574008	-	162	6	6	1	0,001	0,04912	2,60E-09	7,25E-07		TP53
	smOrf_19623	True	12	25362729	25398318	-	228	16	16	1	0,001	0,04912	1,55E-08	2,58E-06		KRAS
	smOrf_21016	True	12	25388140	25398318	-	132	16	16	1	0,001	0,04912	1,55E-08	2,58E-06		KRAS
	smOrf_44005	False	5	304314	344021	+	39708	62	23	8	0,001	0,04912	2,05E-08	2,85E-06		AHRR
	smOrf_46135	True	17	7577598	7578425	-	66	35	35	1	0,001	0,04912	4,92E-06	0,0006		TP53
	smOrf_36958	False	2	133066882	133077808	+	10927	25	6	2	0,001	0,04912	5,90E-06	0,0006	TCONS_I2_00015491	
	smOrf_22067	False	7	57530634	57530756	+	123	4	4	1	0,001	0,04912	3,16E-05	0,0028		ZNF716
	smOrf_42179	False	7	128173707	128231906	+	58200	57	18	8	0,001	0,04912	3,37E-05	0,0028	TCONS_I2_0002619	RP11-274B21.14
	smOrf_32101	False	9	38437830	38458403	+	20574	29	5	2	0,001	0,04912	4,63E-05	0,0035	TCONS_I2_00028727	
	PACA-AU	smOrf_19623	True	12	25362729	25398318	-	228	165	165	1	0,001	0,0375	1,11E-03	4,16E-03	
smOrf_21016		True	12	25388140	25398318	-	132	165	165	1	0,001	0,0375	1,11E-03	4,16E-03		KRAS
PACA-CA	smOrf_19623	True	12	25362729	25398318	-	228	176	176	1	0,001	0,0313	1,11E-03	5,22E-03		KRAS
	smOrf_21016	True	12	25388140	25398318	-	132	176	176	1	0,001	0,0313	1,11E-03	5,22E-03		KRAS
LUSC-KR	smOrf_43215	False	10	38945013	38965111	-	20099	64	42	11	0,001	0,0089	2,39E-09	2,99E-07	TCONS_I2_00004142	
	smOrf_35962	False	9	67293593	67332582	+	38990	91	58	17	0,001	0,0089	1,08E-08	6,38E-07	TCONS_I2_00028767	
	smOrf_34454	False	9	68422201	68429113	-	6913	43	33	10	0,001	0,0089	1,53E-08	6,38E-07	TCONS_I2_00029344	
	smOrf_24181	False	1	143719823	143744288	-	24466	89	68	16	0,001	0,0089	3,45E-08	1,08E-06	TCONS_I2_0000265	RP6-206117.1
	smOrf_30964	False	2	114299778	114326239	+	26462	38	18	6	0,001	0,0089	1,24E-06	3,10E-05	TCONS_I2_0001391	PGM5P4-AS1
	smOrf_38621	False	9	68433538	68438627	-	5090	34	31	5	0,001	0,0089	6,37E-06	0,0001	TCONS_I2_00029349	
	smOrf_29903	False	17	18330505	18379019	+	48515	55	36	11	0,001	0,0089	3,77E-05	0,0007		RP1-37N7.4
	smOrf_54382	False	10	102035218	102042693	-	7476	18	15	5	0,001	0,0089	0,0002	0,0024		BLOC1S2
	smOrf_21297	False	9	68430257	68438623	-	8367	45	42	7	0,001	0,0089	0,0002	0,0029	TCONS_I2_0002935	LOC642236

Table 14 - Small ORFs from SmProt database with significant clustering signals (Q-value analytical). Results obtained from OncodriveCLUSTL, for 4 different sets of SNVs (ICGC project). smORF IDs can be translated into the ones used by SmProt database using Supplementary Table 5. Those smORF that overlap with a known cancer gene, are indicated with a "True" in the Cancer gene column. The location of each smORFs considering the NONCODE database and Gencode v37 is shown in the last two columns.

5.3.11 Low number of somatic SNVs acquired in intergenic novel candidate micropeptides

Based on the results found for published smORFs, we decided to search for recurrence of SNVs in our set of novel and intergenic candidate micropeptides (8.289). However, before running OncodriveCLUSTL we annotate all somatic SNVs from the same ICGC projects used in section 5.3.10.

A total of 3.500 candidates were annotated. Disappointingly, we did not find signals of recurrence since the majority of these candidates (2.996) had only one somatic SNV identified acquired in one tumor genome. No more than 3 SNVs were detected in candidates intergenic micropeptides.

Even large numbers of SNVs identified in tumor types such as esophageal adenocarcinoma (ESAD-UK), SNVs were not located within these short peptides. For this reason, OncodriveCLUSTL could not detect any cluster of variants when evaluating 8.289 novel candidate micropeptides.

Discussion of chapter 3 (study 2) starting in section 6 (page 270).

6. Discussion

Analysis of somatic structural variants in CLL and their incorporation into subclonality studies - Chapter 1

Chronic lymphocytic leukemia is the most prevalent leukemia in Western countries, and it is characterized by a highly variable clinical course. Due to its development over several years, it is an interesting cancer model to study subclonality and evolution during cancer progression, response to therapy or relapse (206,207). The present study was focused on Richter transformation (RT), one type of evolution of CLL tumors into a very aggressive large B cell lymphoma (DLBCL) conferring a dismal prognosis. Moreover, prior to this study the mechanisms driving RT were poorly known (196,198,208). As previously mentioned, the present longitudinal study of chronic lymphocytic leukemia was in collaboration with Dr. Ferran Nadeu and Dr. Elias Campo, and involves diverse groups focused on specific goals with the aim of understanding tumor evolution in Richter syndrome patients. Also, Dra. Romina Royo from BSC was strongly involved in this study, coordinating the analysis of variants. My role within the project was focused on structural variation and intratumor heterogeneity. Particularly, one of the challenges we wanted to address was to include large somatic variants to reconstruct the subclonal architecture of each tumor.

To do so, we first evaluated the landscape of somatic variation within CLL tumors, including SNVs, indels and SVs. SNVs were used in the study to identify and characterize subclones and study their particular role in tumor formation and also in the RT. As of that time, there was no available protocol to assign somatic structural variants to subclones to be able to study the role of SVs in tumor evolution and RT. One of the main challenges rely on the calculation of the frequency at which the SVs are present in the sample, and therefore, to which subclone it belongs. While this is relatively easy for SNVs, the limited number of

mapped reads supporting SVs breaks makes the estimation of allele frequency very challenging.

As mentioned, the identification of somatic structural variants was the first step to later continue with the analysis of intratumor heterogeneity and clonal dynamics in CLL. Many algorithms were published at the time we were working on this project to reconstruct tumor subclonality including CloneHD (209), PhyloWGS (105), DPCLust (184) and SciClone (104). These algorithms and consequently most ITH studies are generally based only on SNVs, and occasionally on small insertions and deletions, as their variant allele frequencies can be better estimated from mapped reads. In this context, I participated actively in the definition of the somatic variation landscape and the subclones for the Nadeu et al, 2022 article. As general discussion on this part (please see the discussion of the main results of this study in Annex 10.3 and in the thesis of Dra. Romina Royo, UB 2023), this study provides a transversal reconstruction of the generation and evolution of the tumor and the subclones, as well as evidence of the presence of RT cells already in early stages of the tumor, encouraging the exploration of early detection protocols for the clinic.

SNVs and indels usually occur more frequently than structural variants within a tumor, so their abundance makes easier the subclonal reconstruction. Moreover, these variants were detected more accurately by variant callers, are less complex and affect only a few nucleotides so calculating their frequency was less challenging. Finally, subclonal reconstruction was even more affordable to obtain just by using coding mutations encompassing important driver events that could be clearly identified from whole-exome sequencing with high depth, which was a rapid, cheaper and comprehensive technology (207).

Although CLL is not known to be a tumor type with high numbers of somatic structural variants, the availability of longitudinal samples was key to study

intratumor heterogeneity and consider their inclusion. Nevertheless, our approach was to reconstruct ITH using single nucleotide variants and indels, since they are more abundant, to later include SVs into those subclones evolving similarly. Considering that, we worked on a strategy to calculate structural variant allele frequencies and translate them into cancer cell fraction. These values will allow us to infer them into previously defined subclones.

In the context of subclonality and SVs, variant allele frequencies are calculated based on the number of reads covering the mutation including mutated and non-mutated fragments. Since structural variants disrupt a large region of the DNA, often containing additional variants close to the breakpoints, their representation within the sequenced sample is not properly translated into the final reference-based alignment that we do for variant calling (i.e. the BAM files), as SVs supporting reads are difficult to align. This, results in a drop of the number of reads (i.e. coverage) around SVs breakpoints and a difficulty in calculating the VAF, as done with SNVs. Therefore, we needed to design and implement a different protocol for the calculation of the VAF for SVs. After considering several possibilities based on the calculation and normalization with genome-wide coverages of the samples, we explored the use of the mutated reads of the SV region, that is, those supporting the variant to infer its variant allele frequency (see results). In this context, it is important to highlight that no proper benchmarking set was available for subclonal SVs and to fully assess the reliability of our approach.

We could apply the defined strategy and calculate variant allele frequencies and their corresponding cancer cell fraction for all structural variants identified in case 63. Longitudinal samples were crucial to track genetic alterations and explore the evolution of cell populations over time. Summarizing the results, we could calculate clonal CCF values for those inversions involving the *ATM* gene present in all collected samples at any time point, whereas we saw an increase in the

frequency of other SVs after the first chemoimmunotherapy was given to the patient.

In any case, we initially considered to publish a methodology for assigning SVs to subclones, but then SVClone (210) was published together with the compendium of articles from The Pan-Cancer Analysis of Whole Genomes. This computational method was developed in fact, for inferring the cancer cell fraction of SV breakpoints from whole-genome sequencing data following a strategy similar to the proposed here. This is why we made the strategic decision of devoting the energy and time to other emerging projects described below. Because of the challenges we encountered for the determination of the VAF for SVs, the publication of a methodology and our focus on other research opportunities, the rest of the CLL study was progressing at a different speed, and we were not ready to include these results when the paper was submitted.

Identification of somatic processed pseudogenes in cancer and evaluation of their functional impact – Chapter 2

Cancer is a complex genetic disease, where the transformation of normal cells to malignant cells is generally driven by a combination of mutations acquired on the DNA sequence. Because of that, the study of these genomic events occurring somatically is essential to understand the basis behind tumor formation and progression. Somatic variation might also provide new clinical markers for a better diagnostic or to select precise treatments. Furthermore, identifying and characterizing these genomic events allow us reclassifying tumors depending on the genetic profile instead of the primary site.

In addition to investigating the landscape of cancer somatic mutations, we extended our focus to include the exploration of somatic retrotransposition

events, particularly somatic processed pseudogenes. We considered the study of such retrotransposition events might harbor implications for human cancer health. Furthermore, our decision was taken based on previous research (94,110,113,119) in this area. Observations from other research groups that began to identify these somatic events in specific cancer samples prompted us to direct attention toward the potential significance of somatic retrotransposition events in a wider context.

Among other large-scale projects intending to identify common patterns of mutations in cancer, the Pan-Cancer Analysis of Whole Genomes was a worldwide initiative collecting 2.600 genomes. PCAWG was coordinated by a series of working groups comprising more than 700 scientists. Our role specifically involved active participation in working group six, primary focus in the analysis of somatic structural variation. Participating in the PCAWG was a significant step in our research journey.

The chance to extensively explore data provided by PCAWG was also a significant motivation for our research. With a comprehensive collection of tumor whole genomes surpassing those previously studied in the context of somatic retrotransposition, we initiated a search for somatic processed pseudogenes across all 34 tumor types provided by this international cancer initiative.

Since standard protocols for the identification of somatic PPs were not published at the time we started our research, by exploring different bioinformatic strategies. Our analysis was based on the combination of automatic searchers for somatic structural variants that could support PPs integration, with manual inspection and validation using tumor and normal sequencing reads from the same tumor genome.

Through the evaluation of the testing set of 48 LUSC tumor genomes by applying diverse criteria combinations, we came up with a conservative protocol

to identify candidate processed pseudogenes. Among the different filters included within our protocol, some were determinant to distinguish from potential false positives, as somatic PPs can be easily confused with the source gene at exon level. For example, the most accurate automatic detection occurred within the most restrictive dataset (4), when structural variants representing both insertion sites and one splice junction event were required. Within this dataset, only one out of 26 candidates could not be manually validated. We perceived these criteria too stringent to define candidates. Furthermore, if a splice junction event was additionally requested, we conducted our search with consideration of a conservative definition of PP, where fragments of one single exon were assumed not to be retrotransposed. The exclusion of single exon insertions as candidate pseudogenes was due to the challenging and potentially unfeasible distinction of their origin.

Regarding the insertion site, we established a minimum nucleotide distance of 100Kb between the insertion site and the source gene responsible for producing the inserted cDNA. This measure aimed to prevent the inclusion of recombination events occurring within a chromosome. Likely, processed pseudogenes are not inserted close to their source gene location.

Second, we did not expect genomic deletions within the insertion region but just a break where the cDNA was interpolated. For this reason, we defined a distance flanking insertion coordinates of less than 350bp. This threshold was determined based on an in-depth consideration of the insert size of our sequencing reads, and the inherent limitations of variant callers when identifying large structural variants using short paired-end reads. The value was carefully selected to account for the precise coordinate error typically associated with variant calling in such scenarios. The nature of these widely studied genomic elements together with the observations done through bioinformatic searches in

this set of tumor genomes were considered to continue developing the automatic searching protocol.

Third, the inserted sequence must include at least 50bp from an exon of the same source gene. This criterion differed from the guidelines applied previous identification protocols. Cooke et al. considered the presence of putative pseudogenes if tumor DNA contained at least three exons from a single gene, with a minimum of two observed canonical splice junctions (110). When extremely short nucleotide regions are inserted into the genome, determining their origin and whether they are newly created through DNA replication or repair, or the consequence of transposition events becomes challenging and requires manual inspection. On the other hand, we anticipated that nucleotide sequences of sufficient length, uniquely aligning to a specific coding sequence would likely arise from the deletion of a coding fragment and its later insertion in the genome (“cut and paste”), or from a retrotransposition event (“copy and paste”). In either case, the result would be mistaken with processed pseudogenes. Additionally, no clear deletions within the source genes were identified across all somatic SVs detected on each tumor. Therefore, single-exon insertions were also considered candidate processed pseudogenes.

Applying the final criteria to all the PCAWG tumor genomes (2589) we could identify 433 candidate processed pseudogenes. Compared to previous studies done in the context of somatic processed pseudogenes acquired in tumor genomes, we could identify the largest number of candidate events mainly because a less stringent criteria were used for the automatic search, sing-exon candidates were included, a higher number of tumor genomes were analyzed and for all of them whole-genome sequencing data was evaluate. In this line, we were able to identify candidates inserted within intergenic regions too.

Notably, experimental validation of somatic processed pseudogenes was not possible in this project due to the lack of fresh material from the tumor samples analyzed.

Even pancreatic tumor type was also included in the collection of samples analyzed by Cooke et al. they could not detect any processed pseudogene across 11 genomes. Contrary to our candidate's selection, genomes corresponding to pancreatic tumors (240 genomes) had the higher number of candidate somatic insertions. When candidates' counts were normalized by the number of tumor genomes analyzed for each tumor type, we could observe a higher prevalence of candidate insertions in head and neck (20,45%) and esophageal adenocarcinomas (8,05 %), which were not included in Cooke et al. study. Lung squamous cell carcinoma and gastric adenocarcinoma had 29,16% % and 7,8% of candidates with at least one acquired PP, being consistent with findings from previous studies where they observed prevalences of 19% and 9% respectively (110).

Although we increased the number of tumor genomes analyzed compared to previous studies (110), validated processed pseudogenes were mainly identified in the same tumor types including lung squamous cell carcinoma and colorectal adenocarcinoma. Processed pseudogenes were newly observed in head and neck squamous cell carcinoma, esophageal adenocarcinoma, ovarian, breast and pancreatic cancer, even some of these tumor types were included previously.

Undoubtedly, in our hands lung squamous cell carcinoma was the tumor type that acquires the highest number of processed pseudogenes (29,16 of donors), as previously reported (110,113). Interestingly, results do not show acquired processed pseudogenes on lung adenocarcinoma (LUAD), a tumor type included on non-small cell lung carcinoma (NSCLC), as LUSC is. As the number of tumor genomes analyzed was similar for both types (40 for LUAD and 48 for

LUSC), we can discard the idea of not identifying PPs because of few genomes. We suggested that the formation of PPs might be specific to certain tumor types, probably depending more on the type of cell affected, than on the organ in which it grows.

The fact that somatic PPs appear only on specific tumor types across a total of 34, suggests a specific mechanism behind the formation of this genomic event, which could explain why some tumors acquire PPs whereas others do not.

Our observations suggest a correlation between the acquisition of somatic processed pseudogenes and somatic retrotransposition events, which primarily include solo-L1 insertions. As shown in the same PCAWG publication where these results were published (211), the highest frequencies of somatic PPs were identified in those tumor types (ESAD, HNSC, LUSC, COAD) that also exhibited significantly enrichment in somatic retrotransposition events. Compared to other tumor types and different classes of structural variants, these tumors had a higher fraction of mobile element insertions. This trend is consistent with the established association between the activity of L1 machinery and the formation of processed pseudogenes (212). Additionally, 71% of the insertion sites defined for the 45 validated processed pseudogenes were within repeat elements, and half of them specifically within L1.

Across all 433 candidate processed pseudogenes, we could count 393 different source genes, where 31 of them appeared retrotransposed in more than 2 tumor genomes. The protein coding gene *TRMT10C* was the most recurrent source gene detected in six different tumor genomes from LUSC, HNSC, GACA and LINC. Copy number variation, and in particular gains on this gene have been reported in LUSC (59% of the TCGA patients) and HNSC (41%) being the highest frequencies across 26 tumor types. However, we cannot directly link *TRMT10C* gains with the formation of this processed pseudogenes since the event is not

identified in the same tumor genomes. Any of these six candidates were manually validated.

Among the candidate source genes, 32 were defined as cancer genes including tumor suppressor genes and oncogenes by the COSMIC database (87). Therefore, less than 9% of the candidate PPs arise from cancer genes. This proportion was also seen considering only validated PPs. Six cancer source genes (*B2M*, *DEK*, *MYH11*, *MYH9*, *PML AND SRGAP3*) were found in more than one candidate processed pseudogene counting 12 different events. Three candidates out of these 12 (*B2M* in one ESAD and one LUSC genome, and *MYH11* acquired in other LUSC genome) could be validated through manual inspection (insertion site and splice junctions confirmed).

The gene ontology enrichment analysis (<https://geneontology.org/>) performed on the 393 different candidate source genes and for cellular component GO terms showed an enrichment of 18,05 (FDR 8.70E-03) in the eukaryotic translation initiation factor 4F complex. This group of proteins found within cells work collaboratively in the initial stages of translation. Overexpression of eIF4F complex components has been observed in several cancers, contributing to increased translation of specific oncogenes. Again, we could not directly correlate this overexpression with PPs formation, and the fact these particular source genes are retrotransposed on tumors since we did not perform differential expression analysis on tumors to prove it. Moreover, we did not prove if enrichment analysis of random sets of genes also points out that particular GO term or if it was specific for this somatic event. Still, it has been shown that overall, genes acting as template for somatic PPs are among the top quartile of expressed genes for each specific tumor type (110).

Focusing on the insertion site of the 433 candidates somatic PPs, half of them appeared to be in intergenic regions while the other half were found in 202

different known genes. Low insertion site recurrence was observed across the 433 events. Only 6 genes appear as the insertion loci for more than one candidate PPs. Recurrence across the insertion sites (3%) where candidate PPs was lower than the observed regarding the source genes (7%), suggesting a relatively random pattern of insertion locations but likely not in the retrotransposition of the source genes. Nevertheless, integration of PPs tends to occur on open-active chromatin regions, as many events appeared inside other expressed genes, since there are fewer genes compared to intergenic regions in the human genome.

More analyses using RNA-seq expression or epigenetic data of all tumor genomes acquiring somatic processed pseudogenes are needed to clearly understand the causes of retrotransposing specific genes into determined genome locations.

To decipher the potentially functional impact of somatic PPs, we evaluated RNA expression of 257 PP events. For the majority of them we were not able to determine supporting RNA reads and therefore, we had inconclusive results. However, we could confirm the expression of 17 fusion PP-host gene or locus. Contrary to the results shown in Cooke et al. study, we confirmed the expression of three processed pseudogenes landing in intergenic regions. Moreover, and even the challenging RNA-seq alignment performed when repeat sequences are included, 4 out of 17 expressed fusion PP-host genes had L1 or other repeat elements within the insertion site. Expressed processed pseudogenes also include four events resulting from the retrocopy of cancer genes, such as *CIITA*, *FEN1*, *KTN1* and *B2M*, which could point to a potential functional interaction and an impact in the biology of the tumor.

Evidence of aberrant fusion transcripts encouraged us to predict the chimeric peptide sequence resulting from them. Sequencing RNA reads joining intron sequences of the host gene together with the source gene were found.

Accordingly, we assumed their translation even in their wild-type form does not codify for proteins. The majority of the aberrant fusion transcripts generate premature stop codons within the coding region of the host transcript, and particularly within their intron sequences. Translation of the processed pseudogene was only predicted for the *WNK4-RND2* fusion transcript, since cDNA *WNKK4* was inserted within an exon of the host gene.

Finally, high variation across tumor genes of both source genes and insertion locus, showed a distinctive nature of these somatic events, with no recurrence across patients in terms of the affected genes or regions. This diversity limits their potential for practical and clinical applications such as identification of targets or biomarkers. Moreover, using somatic processed pseudogenes in precision medicine is highly improbable, as they lack the necessary uniformity. Although they might not be directly applicable in medicine, their study remains significant as they could either act as passenger mutations but also potentially confer functional advantages to the tumor cells.

Together with an extensive identification of structural variants promoted by LINE-1 retrotransposition on PCAWG data, and following PCAWG rules for publishing, our work was published as one section of a broader study of retrotransposition in cancer (211). All the observations illustrate the relevant role of L1 in remodeling the landscape variation of cancer genomes and their potential implications for the formation and development of tumors.

Identification and characterization of novel candidate micropeptides using publicly available genomic and transcriptomic cancer data – Chapter 3

Study 1: Catalog of candidate micropeptides for MS/MS searches

Algorithms developed to find open-reading frames have generally discarded small ORFs as coding genes, mainly because of their short length and their level of uncertainty. But these small ORFs can be translated into micropeptides and have important functional roles. Although great efforts have been made to identify these new coding genes, they are still poorly studied compared to known annotated protein coding genes. The identification itself is already highly challenging, as some of the parameters that are characteristic of micropeptides come close to thresholds that are defined to eliminate noise within studies. For example, the short length, or the potential absence of introns within coding micropeptide DNA regions generate fewer number of supporting mapped reads for micropeptides. Another issue is the differentiation of micropeptides from real exons of longer known genes, as most of the micropeptides so far have been defined in annotated genes. In this frame we worked in collaboration with researchers from VHIO and CNIO to understand the potential role of micropeptides in metastatic processes in pancreatic adenocarcinoma (PACA). The general goal of this part of the study was to identify and characterize micropeptide sequences in exosomes secreted by Pancreatic tumor cells, using a combination of experimental, mass spectrometry and bioinformatic approaches. In proteomic, mass spectrometry peptides are commonly identified by matching MS/MS observed spectra against a theoretical spectrum of all candidate peptides represented in a reference protein sequence dataset. The characteristics of this

dataset of candidate peptides are crucial to ensure a proper balance between having a high enrichment in micropeptides, without losing good candidates and without incorporating false positives. At the BSC we focused on the design and generation of this catalog. Our goal was to improve, adapt and change the standard and default datasets, i.e with known annotated proteins, that are typically used in proteomic MS/MS studies for one more specific towards micropeptide identification. This involved several challenges, mostly related to the identification and inclusion of unknown potentially functional expressed peptides, with expression patterns often close to transcriptional noise. This is actually a new approach known as proteogenomics, where novel peptides are identified by searching MS/MS spectra against a customized protein sequence dataset generated from genomic and transcriptomic data (213). There are different strategies to generate customized protein sequence datasets, and the optimal choice really depends on the goals of the experiment and type of novel peptides the study seeks to identify. Taking this into account, we based our strategy on performing *de novo* transcriptome assembly of RNA-seq samples which will predict known and novel transcripts.

As our work was focused on the metastatic processes in pancreatic adenocarcinoma and transcription and translation are tissue specific, transcriptome assembly was performed from 6 randomly selected ICGC samples corresponding to the same tumor type (PACA).

For the generation of this catalog there was no standard methodology or unified strategy to perform *de novo* transcriptome assembly from RNA-seq samples. Instead, there were several software tools available that could be used with different and specific criteria closer to our needs. In this study StringTie algorithm was used for both *de novo* assembly and quantification of the predicted transcripts. It uses a genome-guided transcriptome assembly approach along with concepts from *de novo* assembly. Based on published studies, this algorithm

exhibits good accuracy in reconstructing transcript structures, is compatible with paired-end RNA-seq libraries and it is more sensitive to genes with low expression levels than other algorithms (214,215). In terms of processing time, it only takes around 20 minutes per sample when run on the MareNostrum 4 supercomputing, and subsequently it was not highly time consuming when launching many tests.

Importantly, StringTie allowed us to fine-tune parameters that we considered optimal for the search of short ORFs. First, the minimum size length was significantly decreased, and transcripts were predicted starting at 50 nucleotides length. Second, we could evaluate the inclusion or exclusion of multi-mapped reads that are usually present in RNA-seq samples. Although including multi-mapped reads (-m 1,0) resulted in the prediction of around 6.000 more transcripts per sample, we did not have sufficient information to assess how many false positives we were also including. This is why we generated DS1 and DS2, with and without considering multimapping reads, respectively. Most of the efforts in this part went to tuning the different parameters for each one of the steps involved in the generation of the catalog, like for example finding the right expression thresholds, the right merging strategy to build the final candidate transcripts, which required specific modifications of the StringTie protocol, or to evaluate the gain of considering non-canonical start codons. Again, our goal was to enrich this dataset in micropeptides without taking much noise as false positive expression signals.

Expression values vary across samples even when looking at the same predicted transcript, but this could be due to differences in cellular composition of samples, technical factors during sample processing or sequencing, time or developmental stages, and external environmental factors.

Generally, median TPM values in our datasets were not higher than 9, indicating, as expected, a prevalence of low expressed observed in all sample showed low levels of expression for most of the predicted transcripts. Note that as an example, expression levels (TPM) for housekeeping genes such as *ACTB*, *GAPDH*, *UBC* or *ADA* observed in the experiment E-MTAB-2706 provided by the Expression Atlas (216) across pancreatic adenocarcinoma cell lines were around 4997, 3540, 1403,75 and 31,47 respectively. Moreover, the average TPM expression for the first two housekeeping genes (*ACTB* and *GAPDH*) was 2162,95 and 1362,81 in our PACA samples. As micropeptides are not really annotated across the genome, our approach needs to consider potential low-expressed and unknown genomic regions, as potential micropeptide genes. Other signals of the potential presence of false positives within our set is the comparison of canonical and non-canonical start codons found in other datasets. Despite this can be due to fragmentation of the candidate transcripts, and general estimates are also biased towards canonical ATG starting genes, it is also likely that a fraction of our transcripts, and final peptides is derived from transcriptional noise.

Although necessary, we did not have the chance and time to perform a proper comparative evaluation of the level of overlap between the different approaches, or whether known expressed transcripts, eventually also known micropeptides, in pancreatic adenocarcinoma were actually enriched within our datasets. Performing these analyses is necessary to provide further support and to evaluate *de novo* and merging strategies applied here to define the PACA transcriptome.

During the generation of this PACA transcriptome we also encountered computational limitations, when relaxing the thresholds and the dataset increased. For example, the size of DS2, where non-canonical start codons were also considered, was computationally and algorithmically not compatible with the MS/MS analysis at the CNIO. Measures, like increasing the stringency in

expression helped us reduce the size of the dataset, although also enriched the catalog towards highly expressed regions, which is not optimal for micropeptide studies.

An evaluation of the position of our candidate small ORFs across the types of annotated regions in the genome, we observed interesting results. A classification of the candidate sequences based on their location compared to annotated genes showed that the majority of candidate ORFs (1.103.297) were located within protein coding genes. Nevertheless, as a result of the Blast filtering step, only 2,7% of them had part (less than 30% in DS2) of their sequence overlapping with a coding exon whereas the remaining ones were within UTRs, or introns. Interestingly, the highest frequency (43,4%) of candidates was observed in 3' UTR regions. While previous studies aiming to classify smORFs (124,150) depending on their location, clearly describe smORFs in 5' UTRs, also called uORFs, downstream smORFs (dORFs) located in 3'UTRs have been less explored. Small ORFs have also been identified in 3' UTRs by ribosome profiling and proteomics, but their frequency tends to be lower than the observed for uORFs. As an example, in a study performed in 2021 analyzing ORFs from OpenProt and sORFs.org, researchers identified 14,4% of novel ORFs overlapping 5' UTRs and only 2,8% in 3'UTRs (149). A similar tendency was observed also in zebrafish embryos. Upstream small ORFs have been systematically characterized and their functions are well known. They act as cis-regulators of the translation of downstream canonical ORFs, and often repress their translation. Upstream ORFs are considered the main class of regulatory small ORFs, and it has been observed that regulation through them is conserved across vertebrates for dozens of genes. Moreover, in many cases their translation starts from non-AUG start codons. Contrary dORFs, those located in 3'UTRs and that appeared to be enriched in our datasets, have not been as much characterized. However, a study published in 2020 reported dORFs enhance translation of their canonical ORFs in both human

cells and zebrafish embryos, indicating a novel and strong post-transcriptional regulatory mechanism (217). Also in dataset 2, around 22% and 21% of the candidates were located in introns or exons of non-coding genes, two smORFs classes that have also been well described in previous studies. In summary, although it is generally known that a significant fraction of the translated ORFs maps to untranslated regions and sequences previously considered noncoding we did not expect an enrichment in sequences located in 3'UTRs rather than in 5' UTRs, introns or non-coding genes. We suggest this enrichment could be due to 3' UTRs are longer than 5' UTR regions in protein coding genes so it is more likely to identify higher numbers of ORFs.

A curated and systematic characterization of all the candidate micropeptides identified in both datasets, based on their classification, could shed light to better understand the results obtained. Expression analysis of each of these groups of candidates could also demonstrate if for example, UTR regions had generally higher expression levels compared to intronic regions or lncRNAs in our defined transcriptome. If this was the case, identifying more candidate sequences in 3'UTRs could not only be because of their larger size (218) but also because their inclusion after all the filtering steps applied. Finally, we also considered that the fact we were stringent with single-exon predicted transcripts and expression abundance resulted in a lower detection of intergenic smORFs.

As we foresee, our sets of *in-silico* translated sequences likely include high numbers of false positive micropeptides even though we ended up reducing both sets. Because of that, at any point of this study all micropeptides were just considered candidates. It is also important to consider the technical debate on whether low levels of expression used to predict transcripts, or the small number of samples evaluated had been sufficient to represent the pancreatic adenocarcinoma transcriptome. To better define and determine these transcripts, a larger cohort of patients is clearly needed.

The generation of this catalog was performed in collaboration with the group at the VHIO, where we regularly discussed with them the results of applying different types of filters, for example. Finally, both datasets of candidate micropeptides were used in MS/MS experiments of pancreatic tumor samples performed at the CNIO, yielding a total of 439 candidate micropeptides.

Interestingly, 167 of these micropeptides were defined in our datasets. Many of these short peptides have been detected in previously annotated non-protein coding regions of the human genome, including UTRs, lncRNAs or pseudogenes. Only 23 out of 439 micropeptides were selected based on their enriched expression in pancreatic adenocarcinoma compared to healthy pancreas, the consistency of their detection across databases and evidences of micropeptide functionality. Despite the PI moved to another location, which inevitably affected the normal progress of this activity, there are still plans to continue with this work. Marion Martínez and Dra. Maria Abad are still working on their functional characterization at VHIO. Finally, we are planning a more extensive description of all 439 identified micropeptides using other computational approaches such as PhyloCSF (162), to evaluate their conservation, IUPred3 (219) to identify disordered protein regions within them and the ELM prediction tool (220) to detect short linear motifs that can be protein interaction sites.

Study 2: Identification of candidate highly conserved micropeptides in intergenic regions

Initiating our exploration of micropeptides in the first study, in which we provided a catalog of novel candidate micropeptides from transcripts expressed in PACA tumors, we realized the underexplored nature of micropeptides located in intergenic human genomic regions. In addition, identifying and characterizing smORFs within gene regions adds the challenge of demonstrating its independent role beyond the role of the surrounding genes. It is known that intergenic ORFs are the most numerous (96% of the smORFs) in human DNA, with a median size of 22 codons. However, many seem not to undergo transcription and to be randomly generated by our genomes rather than have a functional role (124,175).

Identification and annotation of small ORFs is per se, challenging due to their short length compared to known genes and, because of prediction algorithms limitations (171). Furthermore, due to the high numbers of intergenic ORFs, and to avoid inflating the estimates of functional smORFs, these short intergenic peptides are less considered. As expected, micropeptide studies are usually focused on those more likely to be functional. Despite this, we attempted to evaluate these less explored DNA sequences considering they can be a promising source of potentially functional intergenic micropeptides.

Due to the availability and quality of data, as a main strategy, we decided to start the search of micropeptide from the genome, by searching all small intergenic ORFs with different levels of functional evidence, instead of directly exploring the transcriptome as we did in the previous study. These functional evidence were explored using evolutionary and comparative genomic approaches and tools. We also explored other types of functional support, like the known differences in nucleotide composition between functional (coding) and non-functional sequences, previously used in multiple studies for the prediction of

genes in newly sequenced genomes (221). Unfortunately, the signals that we explored with Francisco Cámara (Roderic Guigó's group at the CRG) were too noisy.

Aiming to describe micropeptides with a role in tumorigenesis, we expected their function to be essential for controlling basic cell functions required to survive. It is known that cancer genes such as oncogenes and tumor suppressors are widely conserved through evolution (222). We then expect that functional micropeptides with important cell functions (e.g. cancer-related) will also be conserved across different species and taxonomical clades. This is why comparative analyses of genomes and transcriptomes from multiple species at varying evolutionary distances have been powerful to identify functional coding and non-coding sequences.

So as to restrict our genome-wide search of intergenic micropeptides, we targeted unannotated intergenic constrained regions (UNICORNs) previously identified through a 240 species alignment and published by The Zoonomia Project (5).

UNICORN nucleotide sequences were in-silico translated assuming they were intronless and all their sequence was coding. In fact, it has been shown that new annotated coding genes located in regions previously defined as non-coding have significantly a smaller number of exons, and around 88% of them are single-exon genes (123). Assuming intronless sequences was not so far from what we could somehow expect.

On top of sequence conservation, we included other evolutionary measures, in this case, the ratio of synonymous versus non-synonymous substitution rates (dN/dS) that informs about the selective pressure linked to a coding region and suggests functionality. The reliability of this ratio, which is also used for the identification of positive selection in evolutionary processes, depends on the

number of substitutions identified between two species, which at the same time depends on the evolutionary distance of the two sequences and their length. This means that more recent and shorter micropeptides with low substitution counts might not generate reliable dN/dS ratios and could be classified as non-functional, or neutrally evolving. To find a balance between reliability of the coding alignments between two ortholog candidate micropeptides, needed for the dN/dS calculations, and enough evolutionary distance to ensure a minimum number of synonymous and non-synonymous substitutions (163,223), we used mice for this analysis. A preliminary evaluation of this strategy applied to known functional micropeptides including *AGD3*, Myoregulin, *NBDY*, *SPAAR*, Minion, Phospholamban and Sarcolipin, showed ratios associated with functionality, giving support to the potential benefits in using this tool.

At the level of strategy, we encountered the typical challenges associated with short sequences, which affected all our steps, from the reliability of the alignments to demonstrating expression and the micropeptide using RNAseq data. Because we prioritize stringency and reliability of the sequences found (i.e. absence of false positives), rather than sensitivity, we applied strict thresholds to some of the steps that might have filtered out good micropeptides. For example, as mentioned, very recent micropeptides, despite providing alignments with good quality, often do not have enough substitutions when compared with other species to calculate dN/dS. Older micropeptides might, on the other hand, present problems defining reliable orthologs and aligning their sequences. In addition, the use of sequence conservation also implies that our study might identify the most conserved fragments of longer micropeptides, but not the entire micropeptide.

We know from previously published studies (129,130,171), that micropeptides seem to be less conserved than protein-coding genes, so we expected many of them were not covered on our search. Moreover, it is also

important to bear in mind that the set of small functional proteins not only includes conserved regions but sequences that have recently emerged *de novo* from previously noncoding sequences. Due to their origin from randomly occurring ORFs, *de novo* proteins are also remarkably short. *De novo* originated proteins are known to be species specific, may not be present in the species gene annotations and show little or no signatures of purifying selection, which limits our search. In fact, recently emerged *de novo* genes show high evolutionary rates when compared with more conserved genes (150,224). For all these reasons, our search based on evolutionary conservation and purifying selection limited the identification of functional micropeptides.

Finally, we also used expression as another measure that could indicate functionality. As micropeptides are short and appear to show low levels of expression (129), we could then also miss real micropeptides because the limited availability of good quality raw RNA-seq data and the limited sequencing coverage, which determines the number of final supporting reads. As an example of these limitations, paired-end reads supporting a few candidates showed us that our set of highly conserved smORFs included part of larger candidate transcripts, and therefore we could confirm that in some cases we were not defining the entire smORF but their conserved nucleotides.

For most of the candidate micropeptides signals of expression were not strongly supported but just detected in one sample each. However, we determine evidence of expression in 103 samples including all tissues except muscle for one interesting candidate located in chromosome 5 between 93.615.953 and 93.616.054 bp. The conserved region was around 1700 bp upstream a known lncRNA (*FAM172A*). Even it was close to a known gene, mate reads of the ones aligned across the candidate did not cover *FAM172A* in the samples analyzed. This data suggested that the conserved smORF was not part of the known transcript but a different candidate gene. As it has been observed for other regulatory

smORFs (150), the function of this candidate could be hypothetically linked with the regulation of the lncRNA. Moreover, signals of expression in *FAM172A* can be observed in the same tissue types in the GTEX portal, with the lowest expression values in muscle tissue samples and correlating with the observed signals of expression seen for our candidate smORF.

It should be noted that smORFs shorter than the library size used when sequencing the sample will not be detected if reads are selected depending on their size before the alignment step. RNA-seq samples and other experimental analysis such as microarrays defined precisely to identify expression of short peptides, are needed to confirm these candidates are transcribed in cells. Evidence of expression not only confirms their presence in nature but also provides information about, for example, their tissue specificity or functionality. Furthermore, proteomic data from MS/MS studies or Ribo-seq is needed to validate all these candidate micropeptides accurately. Absolutely, this catalog of novel candidate micropeptides in intergenic regions could be used to analyze raw publicly available MS/MS experiments.

Lastly intending to explore the role of micropeptides in cancer disease and tumorigenesis, we analyzed the recurrence of somatic single nucleotide variants in smORFs. We started with a set of published micropeptides to test driver-discovery algorithms that had not been directly developed for small ORFs. We aimed to identify driver smORFs based on the presence of significant clusters of mutations detected in the ICGC tumor genomes and using OncodriveCLUSTL, a driver-discovery algorithm (7). The number of SNVs was not sufficient for 6 ICGC projects, where OncodriveCLUSTL could not provide good results nor clusters of variants. For these tumor types, a potential approach to obtain better results in future analysis could be increasing the number of samples analyzed, and consequently of variants. Due to differences in mutation rates between tumors, to do so we should be concerned about the specific type and subtype when

searching for more data. At the end, our candidate intergenic smORFs regions did not show enough somatic SNVs from the ICGC tumors analyzed. Because of the low number of variants in these candidates, OncodriveCLUSTL was not able to identify clusters, and even less significant and recurrent variants within the candidates. Based on recurrent and clusters of mutations, we could not suggest a role in cancer for our candidate intergenic smORFs.

We propose other experimental and computational approaches can be applied to evaluate their potential role in cancer, considering that more precise measurements are needed to identify functionality even low signal intensities. In terms of genomic data, gain- and loss-of-function mutations, or copy number alterations within these candidates could also shed light to understand their implication with cancer disease. Differential expression levels comparing tumor and normal tissues will also provide more insights into their function, as well as defined tissue specificity. Proteomic data including MS/MS and Ribosome profiling experiments will not only confirm their presence in humans but also allow us to understand in which cell types are translated, and their abundance.

It is important to know that smORFs can have regulatory effects on neighboring genes (150), and not all are translated into micropeptides neither have their own function. Intergenic conserved regions, can also be regulatory elements regulating gene expression and be structural DNA features such as transcription factor binding sites, contribute to chromatin structure organization of maintain genomic stability (225,226). Thus, these small, conserved intergenic regions will not be transcribed and translated and therefore not considered micropeptides.

In summary, the catalog of candidate micropeptide sequences we provided, a total of 8.289 sequences, is a valuable source of information to perform more analysis including experimental validation. Available algorithms developed to

detect smORFs such as MiPepid (164), sORF Finder (160) or PhyloCSF (162) can be applied on UNICORN regions or candidate smORFs identified to support our findings. Moreover, lastly, an increasing number of intergenic smORFs have been annotated and previously published databases such as sorfs.org (175) or nORFs.org (149) have been updated. Then, we considered it is also important to evaluate the overlap between these recently published smORFs and the candidates identified in our conservation study. As described in this discussion, many questions are still open in the present study, becoming an interesting research line to continue exploring even if it has been usually missed.

7. General overview

The presented thesis is a compilation of genomic studies done with the aim of understanding tumor genomes and the biology behind them. Using diverse classical and novel bioinformatic tools, together with manual inspection of the data to understand and question automatic searches, the focus has primarily been on evaluating two genomic elements: somatic processed pseudogenes and micropeptides.

The landscape of cancer and biomedical research have been notably driven by the potential of next-generation sequencing, which has also changed the way we undertake biological questions leading to important scientific discoveries. Unlike classical genetic studies, a comprehensive evaluation of the human genome now serves as the foundation for more specific research questions. Genomic data, along with other omics data such as transcriptomics, proteomics and epigenomics, have unravel the complexity of tumors, deciphering germline and somatic alterations, gene expression alterations, and environmental changes affecting the way genes normally work. All this information, is essential to understand cancer disease.

Large-scale initiatives and publicly available data, such as The Cancer Genome Atlas or The International Cancer Genome Consortium, have played a crucial role. These initiatives, collect extensive datasets fostering collaborative efforts and accelerating discoveries (88,227). Indeed, the present thesis could not have been done without all these cancer genomic and transcriptomic data available.

Even this vast amount of publicly available data useful for cancer research, there are still many challenges to deal with. Variability across data has been one of these limitations. In one hand, technical information regarding how data was collected, quality, or a clear description of the methodology used to produce the data is occasionally or partially missing. Therefore, it is not allways easy to select

which samples can be included in our specific study, or to understand why we are facing with particular results. Moreover, due to the wide range of available algorithms, to integrate and compare data, data harmonization is needed to ensure compatibility (54,228). If it is not already done, data should be reanalyzed considering the requirements of the research.

Ensuring broad data sharing is nowadays essential. However, even the great initiatives and efforts done in this field, data is not always easily accessible. Moreover, although data agreements are clearly needed, they can be time consuming and a bottleneck in the research. In this line, another challenge we usually encounter is the lack of publicly available clinical data, which is fundamental for translating genomic research findings into actionable insights. The combination of genomic discoveries together with clinical data promotes precision medicine, investing this knowledge in the hands of clinicians and health care systems to really benefit our society. Yet, the challenges persist also in this line, due to ethical management of sensitive patient data and integration of multi-omic information into routine clinical practice. Nevertheless, efforts are being done to accelerate both, the use of clinical data in research, and the application and integration of biological individualized findings into precision medicine. As an example, Genomics England alongside UK National Healthcare System, analyzed WGS data from almost 14000 tumors, integrating genomic data with real-world treatment and outcome data within a secure research environment (229).

While challenges and limitations must be considered in both research and its clinical applications, there is undeniable recognition of the important role that NGS and omics data play in advancing cancer studies and healthcare. Consequently, there is a need to address and overcome these limitations. Efforts should be directed towards comprehensive strategies that enable the extensive utilization of cancer data, facilitating the understanding of the information embedded within tumor cells. By working on these challenges, we can unlock the

potential of genomic insights, paving the way for transformative developments in both cancer research and precision medicine.

8. Conclusions

Chapter 1:

1) Current short read-based sequencing analysis of somatic structural variants in cancer analysis, and structural variant heterogeneity limits the estimation of variant allele frequency to later classify them in tumor subclones.

2) Given this limitation, manual inspection of coverage variability around somatic breakpoints, and the accurate identification of aligned supporting reads is required to reliably estimate and infer their cancer cell fraction to fully characterize intratumor heterogeneity.

Chapter 2:

3) We have been able to define a strategy combining automatic searches with manual inspection to identify somatic processed pseudogenes in cancer. The application of this strategy to the PCAWG cohort allowed us to identify somatic PPs across different tumor types.

4) The distribution of somatic PPs within tumor genomes appeared to be enriched in protein coding genes, and particularly in intronic regions.

5) We observed a heterogeneous distribution of somatic PPs across tumor types, which seems to be correlated with the level and activity of tumor somatic retrotransposition driven by LINE-1 elements.

6) We could identify expressed somatic processed pseudogenes and reconstruct the resulting fusion transcripts. These PP-host gene fusions suggested that somatic PPs can have a functional impact on cell's transcriptional activity.

Chapter 3:

Study 1:

7) The generation of a micropeptide customized catalog based on pancreatic adenocarcinoma expression of non-annotated coding regions, allowed us to identify micropeptides in exosomes secreted by the same tumor type through mass spectrometry.

Study 2:

8) The combination of evolutionary approaches, including nucleotide sequence identity and coding substitution ratio (dn/ds) across species, allowed us to define a set of initial candidates small ORFs located in human intergenic regions.

9) An initial inspection of healthy transcriptomic samples enabled us to observe signals of expression within a few candidate intergenic smORFs. Compared to known protein coding genes, candidate smORFs exhibited low expression levels, presenting a challenge in their assessment using RNA-seq samples.

9. References

1. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* [Internet]. 2020 Feb 5; Available from: <http://www.nature.com/articles/s41588-019-0562-0>
2. Nadeu F, Royo R, Maura F, Dawson KJ, Dueso-Barroso A, Aymerich M, et al. Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis. *Leukemia* [Internet]. 2020 [cited 2023 Feb 8];34:1929–33. Available from: <https://doi.org/10.1038/s41375-020-0730-3>
3. Nadeu F, Royo R, Massoni-Badosa R, Playa-Albinyana H, Garcia-Torre B, Duran-Ferrer M, et al. Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. *Nat Med*. 2022 Aug 1;28(8):1662–71.
4. Cmero M, Ong CS, Yuan K, Schröder J, Mo K, Group PE and HW, et al. SVclone: inferring structural variant cancer cell fraction. *bioRxiv* [Internet]. 2017;172486. Available from: <https://www.biorxiv.org/content/early/2017/08/04/172486>
5. Genereux DP, Serres A, Armstrong J, Johnson J, Marinescu VD, Murén E, et al. A comparative genomics multitool for scientific discovery and conservation. *Nature*. 2020 Nov 12;587(7833):240–5.
6. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Vol. 45, *Nature Genetics*. 2013. p. 580–5.
7. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers. *Bioinformatics*. 2019;35(22):4788–90.
8. Deoxyribonucleic Acid (DNA) [Internet]. [cited 2022 Jul 5]. Available from: <https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid>
9. Definition of gene - NCI Dictionary of Genetics Terms - NCI [Internet]. [cited 2022 Jul 5]. Available from: <https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/gene>
10. Genetics [Internet]. [cited 2022 Jul 5]. Available from: <https://www.nigms.nih.gov/education/fact-sheets/Pages/genetics.aspx>
11. Crick F. Central dogma of molecular biology. *Nature*. 1970;227(5258):561–3.
12. Cseke LJ, Wu W, Kaufman PB. RNA sequencing and analysis. *Handbook of Molecular and Cellular Methods in Biology and Medicine, Second Edition*. 2003;2015(11):237–70.
13. McClean P. A History of Genetics and Genomics. North Dakota State University: PLSC 411/611 - Genomics [Internet]. 2011;(September). Available from: <https://www.ndsu.edu/pubweb/~mcclean/plsc411/History-of-Genetics-and-Genomics-narrative-and-overheads.pdf>
14. Discovery of DNA Double Helix: Watson and Crick | Learn Science at Scitable [Internet]. [cited 2022 Jul 5]. Available from: <https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/>
15. Timeline: History of genomics – YourGenome [Internet]. [cited 2022 Jul 5]. Available from: <https://www.yourgenome.org/facts/timeline-history-of-genomics/>

16. Evolution: Library: The Discovery of DNA's Structure [Internet]. [cited 2022 Jul 5]. Available from: https://www.pbs.org/wgbh/evolution/library/06/3/l_063_01.html
17. Claussnitzer M, Cho JH, Collins R, Cox NJ, Emmanouil T, Hurler ME, et al. A brief history of human disease genetics. 2020;577(7789):179–89.
18. 1966: Genetic Code Cracked [Internet]. [cited 2022 Jul 5]. Available from: <https://www.genome.gov/25520300/online-education-kit-1966-genetic-code-cracked>
19. Regis ED. The forgotten code cracker. Vol. 297, Scientific American. Scientific American Inc.; 2007. p. 50–1.
20. Deciphering the Genetic Code - National Historic Chemical Landmark - American Chemical Society [Internet]. [cited 2022 Jul 5]. Available from: <https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/geneticcode.html>
21. Avila J, Mayor F, Ruiz-Desviat L, Margarita Salas (1938–2019). Nature. 2019 Dec 12;576(7786):208–208.
22. Passarge E. Origins of human genetics. A personal perspective. European Journal of Human Genetics [Internet]. 2021;29(7):1038–44. Available from: <http://dx.doi.org/10.1038/s41431-020-00785-7>
23. Ohta T. Neutral Theory. In: Brenner's Encyclopedia of Genetics: Second Edition. Elsevier Inc.; 2013. p. 67–8.
24. Restriction Enzymes | Learn Science at Scitable [Internet]. [cited 2022 Jul 5]. Available from: <https://www.nature.com/scitable/topicpage/restriction-enzymes-545/>
25. 1968: First Restriction Enzymes Described [Internet]. [cited 2022 Jul 5]. Available from: <https://www.genome.gov/25520301/online-education-kit-1968-first-restriction-enzymes-described>
26. Kresge N, Simoni RD, Hill RL. The characterization of restriction endonucleases: the work of Hamilton Smith. J Biol Chem. 2010 Jan 15;285(3):e2.
27. Coffin JM, Fan H. The Discovery of Reverse Transcriptase. Vol. 3, Annual Review of Virology. Annual Reviews Inc.; 2016. p. 29–51.
28. GNN - Genetics and Genomics Timeline [Internet]. [cited 2022 Jul 5]. Available from: http://www.genomenewsnetwork.org/resources/timeline/1972_Berg.php
29. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
30. da Cunha M de LR de S. Molecular Biology in Microbiological Analysis. In: Reference Module in Food Science. Elsevier; 2019.
31. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.
32. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. 2001.

33. The Human Genome Project [Internet]. [cited 2022 Jul 5]. Available from: <https://www.genome.gov/human-genome-project>
34. 2003: Human Genome Project Completed [Internet]. [cited 2022 Jul 5]. Available from: <https://www.genome.gov/25520492/online-education-kit-2003-human-genome-project-completed>
35. UNESCO research shows women career scientists still face gender bias [Internet]. [cited 2022 Jul 22]. Available from: <https://en.unesco.org/news/unesco-research-shows-women-career-scientists-still-face-gender-bias>
36. Richmond ML. Opportunities for women in early genetics. *Nat Rev Genet.* 2007;8(11):897–902.
37. Women in genetics [Internet]. [cited 2022 Jul 22]. Available from: <https://frontlinegenomics.com/women-in-genetics/>
38. Rich Sobel. 22 Women Geneticists Who Should be Famous! | by Rich Sobel | An Injustice! [Internet]. 2020 [cited 2022 Oct 21]. Available from: <https://aninjusticemag.com/22-women-geneticists-who-should-be-famous-bb046977c5ae>
39. Richmond ML. Women in the early history of genetics. William Bateson and the Newnham College Mendelians, 1900-1910. *Isis.* 2019;92(1):55–90.
40. Women in genetics [Internet]. 2020 [cited 2022 Oct 21]. Available from: <https://frontlinegenomics.com/women-in-genetics/>
41. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Vol. 31, *Nucleic Acids Research.* Oxford University Press; 2003. p. 51–4.
42. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The Ensembl genome database project. *Nucleic Acids Res.* 2002 Jan 1;30(1):38–41.
43. Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch’Ang LY, et al. The international HapMap project. *Nature.* 2003 Dec 18;426(6968):789–96.
44. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Vol. 526, *Nature.* Nature Publishing Group; 2015. p. 68–74.
45. Koepfli KP, Paten B, O’Brien SJ, Antunes A, Belov K, Bustamante C, et al. The genome 10K project: A way forward. *Annu Rev Anim Biosci.* 2015 Feb 1;3:57–111.
46. The Cancer Genome Atlas Program - NCI [Internet]. [cited 2022 Oct 20]. Available from: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>
47. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* [Internet]. 2020;18:9–19. Available from: <https://doi.org/10.1016/j.csbj.2019.11.002>
48. Gibbs RA. The Human Genome Project changed everything. Vol. 21, *Nature Reviews Genetics.* Nature Research; 2020. p. 575–6.

49. Manzoni C, Kia DA, Vandrovцова J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Brief Bioinform.* 2018;19(2):286–302.
50. Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. *Genome Med* [Internet]. 2013 Sep 13 [cited 2022 Aug 5];5(9):79. Available from: <http://genomemedicine.biomedcentral.com/articles/10.1186/gm483>
51. Sequencing Technology | Sequencing by synthesis [Internet]. [cited 2022 Sep 16]. Available from: <https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html>
52. Metzker ML. Sequencing technologies the next generation. Vol. 11, *Nature Reviews Genetics*. 2010. p. 31–46.
53. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost.* 2019;45(7):661–73.
54. Dotolo S, Esposito Abate R, Roma C, Guido D, Preziosi A, Tropea B, et al. Bioinformatics: From NGS Data to Biological Complexity in Variant Detection and Oncological Clinical Practice. *Biomedicines.* 2022 Aug 24;10(9):2074.
55. El-Metwally S, Ouda OM, Helmy M. Next Generation Sequencing Technologies and Challenges in Sequence Assembly [Internet]. *SPRINGER BRIEFS IN SYSTEMS BIOLOGY*. 2014. Available from: <http://www.springer.com/series/10426>
56. Bacher U, Shumilov E, Flach J, Porret N, Joncourt R, Wiedemann G, et al. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. Vol. 8, *Blood Cancer Journal*. Nature Publishing Group; 2018.
57. McCombie WR, McPherson JD, Mardis ER. Next-generation sequencing technologies. *Cold Spring Harb Perspect Med.* 2019 Nov 1;9(11).
58. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature.* 2022;604(7906):437–46.
59. Sequencing Coverage for NGS Experiments [Internet]. [cited 2022 Oct 21]. Available from: <https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html>
60. Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: Where are we now? *Proc Natl Acad Sci U S A* [Internet]. 2021 Dec 28 [cited 2022 Oct 14];118(52):e2109019118. Available from: <https://pnas.org/doi/full/10.1073/pnas.2109019118>
61. GenBank Overview [Internet]. [cited 2022 Oct 14]. Available from: <https://www.ncbi.nlm.nih.gov/genbank/>
62. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. Vol. 17, *Nature Reviews Genetics*. Nature Publishing Group; 2016. p. 257–71.
63. Complex Disease [Internet]. 2022 [cited 2022 Oct 21]. Available from: <https://www.genome.gov/genetics-glossary/Complex-Disease>

64. Morganti S, Tarantino P, Ferraro E, D'Amico P, Viale G, Trapani D, et al. Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life. Vol. 133, *Critical Reviews in Oncology/Hematology*. Elsevier Ireland Ltd; 2019. p. 171–82.
65. Qin D. Next-generation sequencing and its clinical application. *Cancer Biol Med*. 2019;16(1):4–10.
66. Zhang XD. Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. *J Pharmacogenomics Pharmacoproteomics*. 2015;06(02).
67. What Is Cancer? - NCI [Internet]. [cited 2023 Feb 24]. Available from: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
68. Faguet GB. A brief history of cancer: Age-old milestones underlying our current knowledge database. Vol. 136, *International Journal of Cancer*. Wiley-Liss Inc.; 2015. p. 2022–36.
69. Blackadar CB. Historical review of the causes of cancer. Vol. 7, *World Journal of Clinical Oncology*. Baishideng Publishing Group Co., Limited; 2016. p. 54–86.
70. Hansford S, Huntsman DG. Boveri at 100: Theodor Boveri and genetic predisposition to cancer. Vol. 234, *Journal of Pathology*. John Wiley and Sons Ltd; 2014. p. 142–5.
71. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;
72. Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. Vol. 33, *Nature Genetics*. 2003. p. 238–44.
73. Hanahan D, Weinberg R a, Francisco S. The Hallmarks of Cancer Review University of California at San Francisco. 2000;100:57–70.
74. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Vol. 144, *Cell*. 2011. p. 646–74.
75. Hanahan D. Hallmarks of Cancer: New Dimensions. Vol. 12, *Cancer Discovery*. American Association for Cancer Research Inc.; 2022. p. 31–46.
76. Definition of somatic mutation - NCI Dictionary of Cancer Terms - NCI [Internet]. [cited 2023 Aug 11]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/somatic-mutation>
77. Luzzatto L. Somatic mutations in cancer development. In: *Environmental Health: A Global Access Science Source*. 2011.
78. Chin L, Hahn WC, Getz G, Chin L, Hahn WC, Getz G, et al. Making sense of cancer genomic data Making sense of cancer genomic data. 2012;534–55.
79. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Vol. 109, *Cancer Science*. Blackwell Publishing Ltd; 2018. p. 513–22.
80. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol* [Internet]. 2016 Jun 6 [cited 2023 Aug 11];17(1):1–14. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4>

81. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* [Internet]. 2010 Jul 3 [cited 2023 Aug 11];38(16):e164. Available from: [/pmc/articles/PMC2938201/](https://pubmed.ncbi.nlm.nih.gov/20332167/)
82. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* [Internet]. 2001 Jan 1 [cited 2023 Aug 11];29(1):308. Available from: [/pmc/articles/PMC29783/](https://pubmed.ncbi.nlm.nih.gov/11814644/)
83. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020 581:7809 [Internet]. 2020 May 27 [cited 2023 Aug 11];581(7809):434–43. Available from: <https://www.nature.com/articles/s41586-020-2308-7>
84. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* [Internet]. 2018 Jan 1 [cited 2023 Aug 11];46(D1):D1062–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/29165669/>
85. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L Haines . [et al]* [Internet]. 2013 [cited 2023 Aug 11];07(SUPPL.76):Unit7.20. Available from: [/pmc/articles/PMC4480630/](https://pubmed.ncbi.nlm.nih.gov/24480630/)
86. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* [Internet]. 2003 Jul 7 [cited 2023 Aug 11];31(13):3812. Available from: [/pmc/articles/PMC168916/](https://pubmed.ncbi.nlm.nih.gov/12575454/)
87. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* [Internet]. 2019 Jan 8 [cited 2023 Aug 11];47(D1):D941–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/30371878/>
88. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* [Internet]. 2013 Oct 1 [cited 2023 Aug 11];45(10):1113. Available from: [/pmc/articles/PMC3919969/](https://pubmed.ncbi.nlm.nih.gov/24046052/)
89. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods* 2013 10:11 [Internet]. 2013 Sep 15 [cited 2023 Aug 11];10(11):1081–2. Available from: <https://www.nature.com/articles/nmeth.2642>
90. Olafsson S, Anderson CA. Somatic mutations provide important and unique insights into the biology of complex diseases. Vol. 37, *Trends in Genetics*. Elsevier Ltd; 2021. p. 872–81.
91. Reddy EP, Reynolds RK, Santos E, Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* 1982 300:5888 [Internet]. 1982 [cited 2023 Aug 11];300(5888):149–52. Available from: <https://www.nature.com/articles/300149a0>
92. Hussen BM, Abdullah ST, Salihi A, Sabir DK, Sidiq KR, Rasul MF, et al. The emerging roles of NGS in clinical oncology and personalized medicine. Vol. 230, *Pathology Research and Practice*. Elsevier GmbH; 2022.
93. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell* [Internet]. 2013;153(1):17–37. Available from: <http://dx.doi.org/10.1016/j.cell.2013.03.002>

94. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of somatic retrotransposition in human cancers. *Science* (1979). 2012 Aug 24;337(6097):967–71.
95. Ojha J, Ayres J, Secreto C, Tschumper R, Rabe K, Dyke D Van, et al. Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia Key Points. 2015; Available from: www.bloodjournal.org
96. Espiritu SMG, Liu LY, Rubanova Y, Bhandari V, Holgersen EM, Szyca LM, et al. The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* [Internet]. 2018;173(4):1003-1013.e15. Available from: <https://doi.org/10.1016/j.cell.2018.03.029>
97. Niida A, Nagayama S, Miyano S, Mimori K. Understanding intratumor heterogeneity by combining genome analysis and mathematical modeling. *Cancer Sci*. 2018;109(4):884–92.
98. Marusyk A, Polyak K. Tumor heterogeneity: causes and consequences. [cited 2018 Jul 4]; Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2814927/pdf/nihms-160602.pdf>
99. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. *Cell*. 2013 Apr 25;153(3):666–77.
100. Landau DA, Sun C, Rosebrock D, Herman SEM, Fein J, Sivina M, et al. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat Commun* [Internet]. 2017;8(1). Available from: <http://dx.doi.org/10.1038/s41467-017-02329-y>
101. Dentre SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*. 2021 Apr 15;184(8):2239-2254.e39.
102. Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, et al. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLoS Med*. 2015;12(2).
103. Braggio E, Kay NE, Vanwier S, Tschumper RC, Smoley S, Eckel-Passow JE, et al. Longitudinal genome-wide analysis of patients with chronic lymphocytic leukemia reveals complex evolution of clonal architecture at disease progression and at the time of relapse. *Leukemia*. 2012;26(7):1698–701.
104. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput Biol*. 2014;10(8).
105. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol*. 2015;16(1):1–20.
106. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993–8.
107. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113–20.

108. Kazazian HH. Mobile elements: drivers of genome evolution. *Science* [Internet]. 2004 Mar 12 [cited 2023 Nov 15];303(5664):1626–32. Available from: <https://pubmed.ncbi.nlm.nih.gov/15016989/>
109. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* 2009 10:10 [Internet]. 2009 Oct [cited 2023 Nov 15];10(10):691–703. Available from: <https://www.nature.com/articles/nrg2640>
110. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JMC, et al. Processed pseudogenes acquired somatically during cancer development. *Nat Commun*. 2014 Apr 9;5.
111. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Vol. 345, *Science*. American Association for the Advancement of Science; 2014.
112. Ding W, Lin L, Chen B, Dai J. L1 elements, processed pseudogenes and retrogenes in mammalian genomes. *IUBMB Life*. 2006;58(12):677–85.
113. Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol*. 2013 Mar 13;14(3).
114. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*. 2004 May 20;429(6989):268–74.
115. Vanin EF. PROCESSED PSEUDOGENES: CHARACTERISTICS AND EVOLUTION [Internet]. Vol. 19, *Ann. Rev. Genet*. 1985. Available from: www.annualreviews.org
116. Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human pseudogenes. *Genome Res*. 2003;13(12):2559–67.
117. Miller TLA, Orpinelli F, Leonel J, Buzzo L, Galante PAF. sideRETRO: a pipeline for identifying somatic and poly-morphic insertions of processed pseudogenes or retrocopies. *Bioinformatics* [Internet]. 2021;37(3):419–21. Available from: <https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa689/5876827>
118. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res*. 2005;33(8):2374–83.
119. Goodier JL. Retrotransposition in tumors and brains [Internet]. 2014. Available from: <http://www.mobilednajournal.com/content/5/1/11>
120. de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, et al. Primary Immunodeficiency Caused by an Exonized Retroposed Gene Copy Inserted in the CYBB Gene. *Hum Mutat*. 2014;35(4):486–96.
121. Kazazian HH. Processed pseudogene insertions in somatic cells. Vol. 5, *Mobile DNA*. BioMed Central Ltd.; 2014.
122. Y Miki, I Nishisho, A Horii, Y Miyoshi, J Utsunomiya, KW Kinzler, et al. *Cancer Research*. 1992 [cited 2023 Jan 9]. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Available from: <https://pubmed.ncbi.nlm.nih.gov/1310068/>

123. Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, et al. A hidden human proteome encoded by “non-coding” genes. *Nucleic Acids Res.* 2019 Sep 5;47(15):8111–25.
124. Couso JP, Patraquim P. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol* [Internet]. 2017 Sep 12;18(9):575–89. Available from: <http://www.nature.com/articles/nrm.2017.58>
125. Makarewicz CA, Olson EN. Mining for Micropeptides. *Trends Cell Biol* [Internet]. 2017 Sep 1;27(9):685–96. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0962892417300648>
126. Choi SW, Kim HW, Nam JW. The small peptide world in long noncoding RNAs. *Brief Bioinform.* 2018 Jun 29;
127. Zhe Ji, Ruisheng Song, Aviv Regev, Kevin Struhl. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* [Internet]. 2015 [cited 2022 Sep 9]; Available from: <https://elifesciences.org/articles/08890>
128. Yeasmin F, Yada T, Akimitsu N. Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics. Vol. 9, *Frontiers in Genetics*. Frontiers Media S.A.; 2018.
129. Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO Journal.* 2014;33(9):981–93.
130. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods.* 2016;13(2):165–70.
131. Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, et al. The translation of non-canonical open reading frames controls mucosal immunity. *Nature.* 2018 Dec 20;564(7736):434–8.
132. Cao X, Slavoff SA. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. *Exp Cell Res* [Internet]. 2020;(March):111973. Available from: <https://doi.org/10.1016/j.yexcr.2020.111973>
133. Galindo MI, Pueyo JJ, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* 2007 May;5(5):1052–62.
134. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol.* 2008 Dec;70(6):1487–501.
135. Zhang Q, Vashisht AA, O'Rourke J, Corbel SY, Moran R, Romero A, et al. The microprotein Minion controls cell fusion and muscle formation. *Nat Commun.* 2017 Jun 1;8.
136. Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, et al. MTORC1 and muscle regeneration are regulated by the LINCO0961-encoded SPAR polypeptide. *Nature.* 2017 Jan 12;541(7636):228–32.
137. Huang JZ, Chen M, Chen D, Gao XC, Zhu S, Huang H, et al. A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol Cell.* 2017 Oct 5;68(1):171-184.e6.

138. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol.* 2007 Jun;9(6):660–5.
139. Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, et al. Toddler: An embryonic signal that promotes cell movement via apelin receptors. *Science* (1979). 2014;343(6172).
140. Kikuchi K, Fukuda M, Ito T, Inoue M, Yokoi T, Chiku S, et al. Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation. *Nucleic Acids Res.* 2009;37(15):4987–5000.
141. Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell.* 2015 Feb 12;160(4):595–606.
142. Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, et al. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* (1979). 2013;341(6150):1116–20.
143. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, et al. Muscle physiology: A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* (1979). 2016 Jan 15;351(6270):271–5.
144. Bi P, Ramirez-Martinez A, Li H, Cannavino J, McAnally JR, Shelton JM, et al. Control of muscle formation by the fusogenic micropeptide myomixer. *Science* (1979). 2017 Apr 21;356(6335):323–7.
145. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A. A human short open reading frame (sORF)-Encoded polypeptide that stimulates DNA end joining. *Journal of Biological Chemistry.* 2014 Apr 18;289(16):10950–7.
146. D’Lima NG, Ma J, Winkler L, Chu Q, Loh KH, Corpuz EO, et al. A human microprotein that interacts with the mRNA decapping complex. *Nat Chem Biol.* 2017 Feb 1;13(2):174–80.
147. Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, et al. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer’s disease genes and A. 2001 [cited 2023 Feb 10]; Available from: www.pnas.org/cgi/doi/10.1073/pnas.101133498
148. Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan J, et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab.* 2015 Mar 3;21(3):443–54.
149. Matthew DCN, Kohze R, Erady C, Meena N, Hayden M, Cooper DN, et al. A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Res.* 2021;31(2):327–36.
150. Ruiz-Orera J, Albà MM. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. Vol. 35, *Trends in Genetics.* Elsevier Ltd; 2019. p. 186–98.
151. Choteau SA, Wagner A, Pierre P, Spinelli L, Brun C. MetamORF: A repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database.* 2021;2021.

152. Ruiz-Orera J, Albà MM. Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures. *NAR Genom Bioinform.* 2019 Apr 1;1(1).
153. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* [Internet]. 2015 Dec 14;16(1):179. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0742-x>
154. Khitun A, Slavoff SA. Proteomic Detection and Validation of Translated Small Open Reading Frames. *Curr Protoc Chem Biol.* 2019 Dec 1;11(4):e77.
155. 7.13C: Homologs, Orthologs, and Paralogs - Biology LibreTexts [Internet]. [cited 2023 Jan 24]. Available from: [https://bio.libretexts.org/Bookshelves/Microbiology/Microbiology_\(Boundless\)/07%3A_Microbial_Genetics/7.13%3A_Bioinformatics/7.13C%3A_Homologs_Orthologs_and_Paralogs#:~:text=A%20homologous%20gene%20\(or%20homolog,sequence%20are%20not%20necessarily%20homologous.](https://bio.libretexts.org/Bookshelves/Microbiology/Microbiology_(Boundless)/07%3A_Microbial_Genetics/7.13%3A_Bioinformatics/7.13C%3A_Homologs_Orthologs_and_Paralogs#:~:text=A%20homologous%20gene%20(or%20homolog,sequence%20are%20not%20necessarily%20homologous.)
156. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. *Nat Biotechnol* [Internet]. 2022 Jul 13;40(7):994–9. Available from: <https://www.nature.com/articles/s41587-022-01369-0>
157. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS Genet.* 2008 Dec;4(12).
158. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution [33]. Vol. 267, *Nature*. Nature Publishing Group; 1977. p. 275–6.
159. Jeffares DC, Tomiczek B, Sojo V, dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods in Molecular Biology.* 2015;1201:65–90.
160. Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu SH. sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics.* 2009 Dec 14;26(3):399–400.
161. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005 Aug;15(8):1034–50.
162. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* [Internet]. 2011 Jul 1 [cited 2023 Feb 7];27(13):i275–82. Available from: <https://academic.oup.com/bioinformatics/article/27/13/i275/178183>
163. Straub D, Wenkel S. Cross-species genome-wide identification of evolutionary conserved microproteins. *Genome Biol Evol.* 2017 Mar 1;9(3):777–89.
164. Zhu M, Gribskov M. MiPepid: MicroPeptide identification tool using machine learning. *BMC Bioinformatics* [Internet]. 2019 Nov 8 [cited 2023 Feb 7];20(1):559. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3033-9>

165. Ji X, Cui C, Cui Q. smORFunction: a tool for predicting functions of small open reading frames and microproteins. *BMC Bioinformatics*. 2020 Dec 1;21(1).
166. Skarshewski A, Stanton-Cook M, Huber T, Al Mansoori S, Smith R, Beatson SA, et al. uPEPPERoni: An online tool for upstream open reading frame location and analysis of transcript conservation [Internet]. 2014. Available from: <http://u pep-scmb>.
167. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc*. 2012;7(8):1534–50.
168. Glish GL, Vachet RW. The basics of mass spectrometry in the twenty-first century. Vol. 2, *Nature Reviews Drug Discovery*. 2003. p. 140–50.
169. What is Mass Spectrometry? | Broad Institute [Internet]. [cited 2023 Jan 31]. Available from: <https://www.broadinstitute.org/technology-areas/what-mass-spectrometry>
170. Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, et al. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal Chem*. 2016 Apr 5;88(7):3967–75.
171. Leong AZX, Lee PY, Mohtar MA, Syafruddin SE, Pung YF, Low TY. Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. Vol. 29, *Journal of Biomedical Science*. BioMed Central Ltd; 2022.
172. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*. 2013 Jan;9(1):59–64.
173. Vanderperre B, Lucier JF, Roucou X. HALtORF: A database of predicted out-of-frame alternative open reading frames in human. *Database*. 2012;2012.
174. Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G. SORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res*. 2016;44(D1):D324–9.
175. Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res*. 2018 Jan 1;46(D1):D497–502.
176. Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief Bioinform* [Internet]. 2017 Jan 29;19(4):bbx005. Available from: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx005>
177. Brunet MA, Brunelle M, Lucier JF, Delcourt V, Levesque M, Grenier F, et al. OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res*. 2019 Jan 8;47(D1):D403–10.
178. Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, Leutz A. UORFdb - A comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res*. 2014 Jan 1;42(D1).
179. Wan J, Qian SB. TISdb: A database for alternative translation initiation in mammalian cells. *Nucleic Acids Res*. 2014 Jan 1;42(D1):D845.

180. Saha S, Chatzimichali EA, Matthews DA, Bessant C. PITDB: A database of translated genomic elements. *Nucleic Acids Res.* 2018 Jan 1;46(D1):D1223–8.
181. Liu W, Xiang L, Zheng T, Jin J, Zhang G. TranslatomeDB: A comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res.* 2018 Jan 1;46(D1):D206–12.
182. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: An updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D230–4.
183. Moncunill V, Gonzalez S, Beà S, Andrieux LO, Salaverria I, Royo C, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotechnol* [Internet]. 2014 Nov 26 [cited 2018 Jul 10];32(11):1106–12. Available from: <http://www.nature.com/articles/nbt.3027>
184. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. *Cell.* 2012 May 25;149(5):994–1007.
185. Wala JA, Bandopadhyay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* [Internet]. 2018 Mar 13 [cited 2018 Jul 10];28(4):581–91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29535149>
186. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012 Sep;28(18).
187. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015 Oct 22;526(7574):519–24.
188. Dentre SC, Wedge DC, Loo P Van. Principles of reconstructing the subclonal architecture of cancers.
189. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res* [Internet]. 2002 Apr 1 [cited 2023 Aug 11];12(4):656–64. Available from: <https://genome.cshlp.org/content/12/4/656.full>
190. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* [Internet]. 1990 [cited 2023 Aug 11];215(3):403–10. Available from: <https://pubmed.ncbi.nlm.nih.gov/2231712/>
191. Li H, Joh YS, Kim H, Paek E, Lee SW, Hwang KB. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics.* 2016 Dec 22;17.
192. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012 Sep 6;489(7414):57–74.
193. Nadeu F, Clot G, Delgado J, Martín-García D, Baumann T, Salaverria I, et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia.* 2018;32(3):645–53.
194. Araf S, Wang J, Korfi K, Pangault C, Kotsiou E, Rio-Machin A, et al. Genomic profiling reveals spatial intra-tumor heterogeneity in follicular lymphoma. *Leukemia* 2018

- 32:5 [Internet]. 2018 Feb 8 [cited 2023 Dec 27];32(5):1261–5. Available from: <https://www-nature-com.sire.ub.edu/articles/s41375-018-0043-y>
195. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* [Internet]. 2012 Mar 8 [cited 2023 Dec 27];366(10):883–92. Available from: <https://www-nejm-org.sire.ub.edu/doi/full/10.1056/NEJMoa1113205>
 196. Jain P. Richter’s Transformation in Chronic Lymphocytic Leukemia. *Oncology Journal*. 2012.
 197. Klintman J, Appleby N, Stamatopoulos B, Ridout K, Eyre TA, Robbe P, et al. Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood*. 2021 May 20;137(20):2800–16.
 198. Rossi D, Gaidano G. Richter syndrome: Molecular insights and clinical perspectives. Vol. 27, *Hematological Oncology*. 2009. p. 1–10.
 199. Fabbri G, Khiabani H, Holmes AB, Wang J, Messina M, Mullighan CG, et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *Journal of Experimental Medicine*. 2013 Oct 21;210(11):2273–88.
 200. Rossi D, Spina V, Deambrogi C, Rasi S, Laurenti L, Stamatopoulos K, et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood*. 2011 Mar 24;117(12):3391–401.
 201. Scandurra M, Rossi D, Deambrogi C, Rancoita PM, Chigrinova E, Mian M, et al. Genomic profiling of Richter’s syndrome: recurrent lesions and differences with *de novo* diffuse large B-cell lymphomas. *Hematol Oncol*. 2010 Jun 13;28(2):62–7.
 202. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. *Nature* 2020 578:7793 [Internet]. 2020 Feb 5 [cited 2023 Nov 25];578(7793):112–21. Available from: <https://www-nature-com/articles/s41586-019-1913-9>
 203. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. *Nature* 2020 578:7793 [Internet]. 2020 Feb 5 [cited 2023 Nov 25];578(7793):82–93. Available from: <https://www-nature-com/articles/s41586-020-1969-6>
 204. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, et al. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res* [Internet]. 2005 Jan 1 [cited 2024 Jan 3];33(Database issue). Available from: <https://pubmed.ncbi.nlm.nih.gov/15608158/>
 205. Liu B, Chen Y, Yang J. LncRNAs are altered in lung squamous cell carcinoma and lung adenocarcinoma. *Oncotarget* [Internet]. 2017 Apr 4 [cited 2024 Jan 7];8(15):24275. Available from: [/pmc/articles/PMC5421846/](https://pubmed.ncbi.nlm.nih.gov/28421846/)
 206. González-Rincón J, Gómez S, Martínez N, Troulé K, Perales-Patón J, Derdak S, et al. Clonal dynamics monitoring during clinical evolution in chronic lymphocytic leukaemia. *Sci Rep*. 2019 Dec 1;9(1).
 207. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013 Feb 14;152(4):714–26.

208. Jain P, O'Brien S. Richter's Transformation in Chronic Lymphocytic Leukemia | Cancer Network. Oncology (Williston Park) [Internet]. 2012;26(12):1146–52. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23413591> <http://www.cancernetwork.com/oncology-journal/richters-transformation-chronic-lymphocytic-leukemia>
209. Fischer A, Vázquez-García I, Illingworth CJR, Mustonen V. High-definition reconstruction of clonal composition in cancer. *Cell Rep*. 2014 Jun 12;7(5):1740–52.
210. Cmero M, Yuan K, Ong CS, Schröder J, Adams DJ, Anur P, et al. Inferring structural variant cancer cell fraction. *Nat Commun*. 2020;11(1).
211. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature Genetics* 2020 52:3 [Internet]. 2020 Feb 5 [cited 2023 Nov 15];52(3):306–19. Available from: <https://www.nature.com/articles/s41588-019-0562-0>
212. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* [Internet]. 2000 Apr [cited 2023 Nov 15];24(4):363–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/10742098/>
213. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. Vol. 11, *Nature Methods*. Nature Publishing Group; 2014. p. 1114–25.
214. Perteua M, Perteua GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5.
215. Liu X, Zhao J, Xue L, Zhao T, Ding W, Han Y, et al. A comparison of transcriptome analysis methods with reference genome. *BMC Genomics*. 2022 Dec 1;23(1).
216. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* [Internet]. 2022 Jan 7 [cited 2023 Dec 7];50(D1):D129–40. Available from: <https://dx.doi.org/10.1093/nar/gkab1030>
217. Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N, Bazzini AA. Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J* [Internet]. 2020 Sep 9 [cited 2023 Dec 8];39(17). Available from: </pmc/articles/PMC7459409/>
218. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. *Genome Biol* [Internet]. 2002 Feb 28 [cited 2023 Dec 7];3(3):1–10. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-3-reviews0004>
219. Abor Erd G', Os ", Atyásaty'atyás Pajkos M', Dosztányi Z, Dosztányi D. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res* [Internet]. 2021 Jul 2 [cited 2023 Dec 9];49(W1):W297–303. Available from: <https://dx.doi.org/10.1093/nar/gkab408>
220. Kumar M, Michael S, Alvarado-Valverde J, McRossed D, Sign@száros B, Sámano-Sánchez H, Zeke A, et al. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res* [Internet]. 2022 Jan 7 [cited 2023 Dec 9];50(D1):D497–508. Available from: <https://dx.doi.org/10.1093/nar/gkab975>

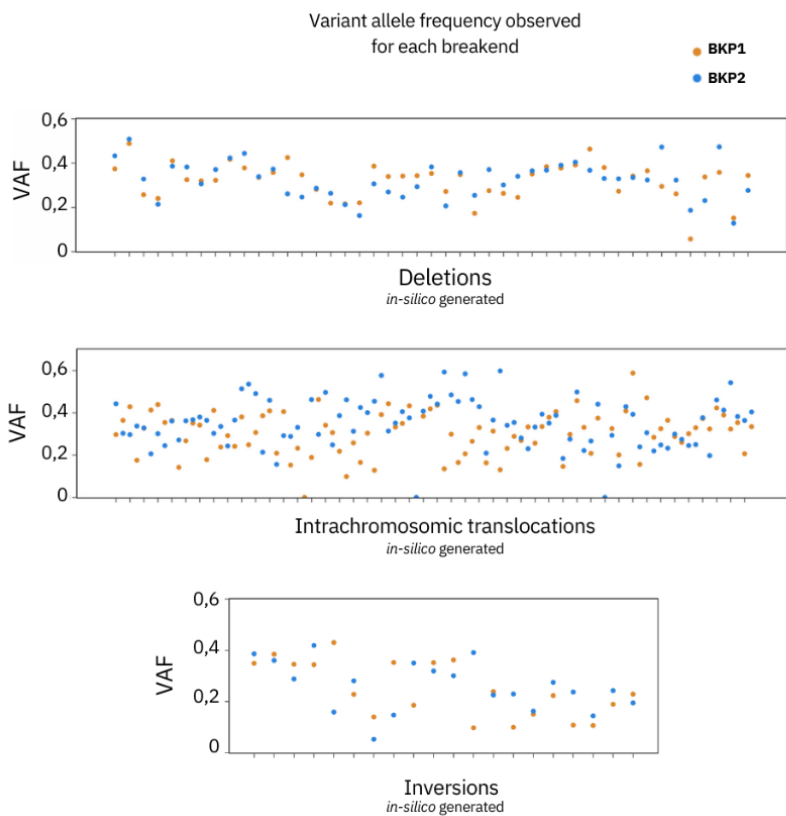
221. Alioto T, Blanco E, Parra G, Guigó R. Using geneid to Identify Genes. *Curr Protoc Bioinformatics*. 2018;64(1):1–26.
222. Zingde SM. Cancer genes. *Western Journal of Medicine* [Internet]. 1993 Sep 10 [cited 2024 Jan 7];158(3):273. Available from: [/pmc/articles/PMC1311753/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/1311753/)
223. Kawashima T. Comparative and evolutionary genomics. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. 2018 Jan 1;1–3:257–67.
224. Montañés JC, Huertas M, Messeguer X, Albà MM. Evolutionary Trajectories of New Duplicated and Putative De Novo Genes. *Mol Biol Evol*. 2023 May 1;40(5).
225. Khaitovich P, Kelso J, Franz H, Visagie J, Giger T. Functionality of intergenic transcription: An evolutionary comparison. *PLoS Genet* [Internet]. 2006 [cited 2024 Jan 1];2(10):171. Available from: www.plosgenetics.org
226. Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, Park BK, et al. Regulatory Roles of Conserved Intergenic Domains in Vertebrate Dlx Bigene Clusters. *Genome Res* [Internet]. 2003 Apr 4 [cited 2023 Dec 30];13(4):533. Available from: [/pmc/articles/PMC430168/](https://pubmed.ncbi.nlm.nih.gov/12430168/)
227. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database (Oxford)* [Internet]. 2011 [cited 2024 Jan 8];2011. Available from: [/pmc/articles/PMC3263593/](https://pubmed.ncbi.nlm.nih.gov/2163593/)
228. Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat Biotechnol*. 2021 Sep 1;39(9):1141–50.
229. Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, Jones L, et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med* [Internet]. 2024 Jan 11; Available from: <https://www.nature.com/articles/s41591-023-02682-0>

10. Annex

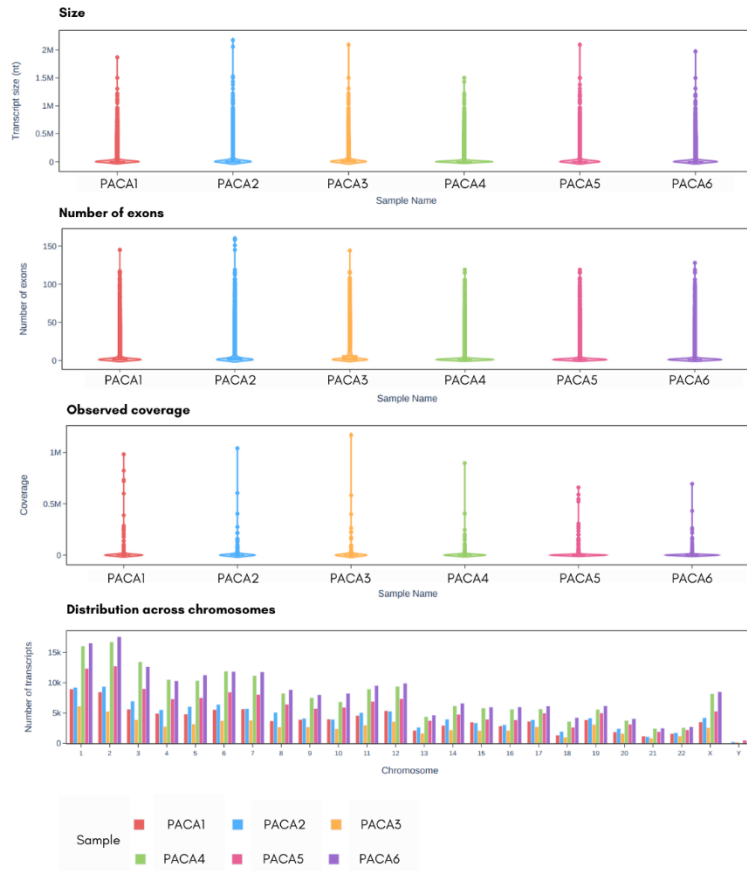
10.1 Supplementary figures



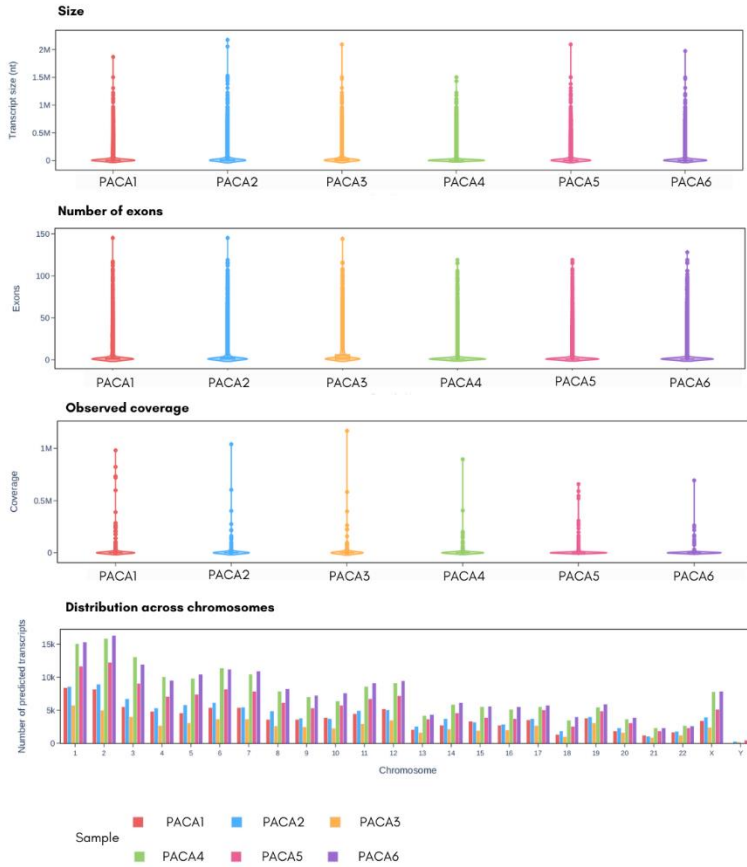
Supplementary Figure 1. Coverage distribution across four genomic regions where SVs have not been identified. Each line correspond to a genomic sample from case 29. Blue represents normal genome, whereas pink and orange two different tumor samples.



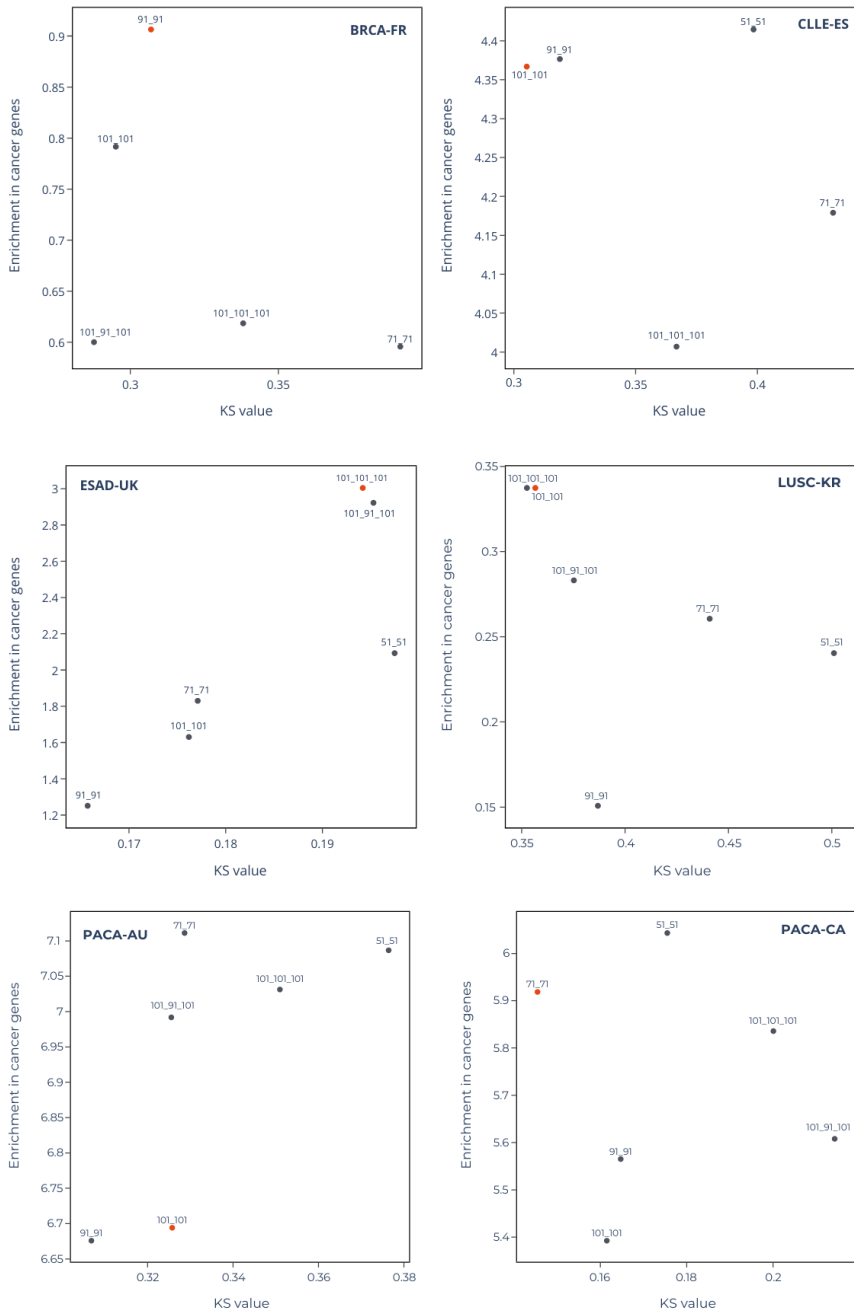
Supplementary Figure 2. Variant allele frequencies calculated for each breakend (blue and orange). Results are shown for deletions, intrachromosomal translocations and inversions (x axis) from the *in-silico* genome.



Supplementary Figure 3. Summary including the size (nucleotides), number of exons, observed coverage and distribution across chromosomes of the transcripts predicted in each PACA sample. Database version 1 strategy.



Supplementary Figure 4. Summary including the size (nucleotides), number of exons, observed coverage and distribution across chromosomes of the transcripts predicted in each PACA sample. Database version 2 strategy.



Supplementary figure 5. KS value compared the enrichment in cancer genes, obtained for 6 different ICGC cancer projects. Each dot represents one OncodriveCLUSTL test using an specific combination of parameters. Dots in red indicate the selected combination.

10.2 Supplementary tables

TIME POINT	SV ID	CHR	POS	STRAND	CHR	POS	STRAND	SV TYPE	LOCATION (bkp1)	LOCATION (bkp2)	CLL DRIVERS	DLBCL DRIVERS
T1-PB	SV_6	11	106417594	+	11	110207731	+	INV	intergenic	intergenic	ATM	ATM
T1-PB	SV_7	11	108122764	-	11	116153393	-	INV	ATM	intergenic	ATM	ATM
T1-PB	SV_11	X	147727409	+	X	147738814	+	INV	AFF2	AFF2		
T1-PB	SV_12	X	147731093	-	X	147738794	-	INV	AFF2	AFF2		
T1-LN	SV_9	11	106417594	+	11	110207731	+	INV	intergenic	intergenic	ATM	ATM
T1-LN	SV_10	11	108122764	-	11	116153393	-	INV	ATM	intergenic	ATM	ATM
T1-LN	SV_18	X	147727409	+	X	147738814	+	INV	AFF2	AFF2		
T1-LN	SV_19	X	147731093	-	X	147738787	-	INV	AFF2	AFF2		
T2	SV_21	11	106417594	+	11	110207731	+	INV	intergenic	intergenic	ATM	ATM
T2	SV_22	11	108122764	-	11	116153393	-	INV	ATM	intergenic	ATM	ATM
T2	SV_24676	13	26926979	-	13	47816848	-	INV	CDK8	intergenic		FOXO1
T2	SV_26	13	47816186	-	13	47816851	-	INV	intergenic	intergenic		
T2	SV_27	13	47816221	+	13	57245235	-	DEL	intergenic	intergenic		
T2	SV_46	X	147727409	+	X	147738814	+	INV	AFF2	AFF2		
T2	SV_47	X	147731093	-	X	147738787	-	INV	AFF2	AFF2		

TIME POINT	SV ID	CHR	POS	STRAND	CHR	POS	STRAND	SV TYPE	LOCATION (bkp1)	LOCATION (bkp2)	CLL DRIVERS	DLBCL DRIVERS
T3	SV_16	2	137757587	+	2	137757897	-	DEL	THSD7B	THSD7B		
T3	SV_29	4	136358882	-	4	140818871	-	INV	intergenic	MAML3		
T3	SV_30	4	140808906	-	4	140856382	-	INV	MAML3	MAML3		
T3	SV_56	7	52263031	-	7	52263480	-	INV	intergenic	intergenic		
T3	SV_66	9	8495226	+	9	31606757	-	DEL	PTPRD	intergenic	CDKN2A	PTPRD,CDKN2A,CDKN2B
T3	SV_308	9	131108130	+	11	118477716	-	TRA	SLC27A4	PHLDB1		
T3	SV_89	9	131209465	+	11	118814376	+	TRA	MIR1268A	intergenic		
T3	SV_309	9	131222219	-	11	118861651	+	TRA	MIR1268A	intergenic		
T3	SV_88	9	131231071	+	11	118477448	+	TRA	MIR1268A	PHLDB1		
T3	SV_72	10	85468180	+	10	85468343	-	DEL	intergenic	intergenic		
T3	SV_75	11	7897130	-	11	7898281	-	INV	LOC283299	LOC283299		
T3	SV_184	11	33731887	+	11	33738914	-	DEL	CD59	CD59		
T3	SV_79	11	63968132	-	11	63973300	+	DUP	STIP1	intergenic		
T3	SV_80	11	67352245	+	11	67353857	-	DEL	GSTP1	GSTP1		RAD9A
T3	SV_82	11	106417594	+	11	110207731	+	INV	intergenic	intergenic	ATM	ATM
T3	SV_83	11	108122764	-	11	116153393	-	INV	ATM	intergenic	ATM	ATM
T3	SV_84	11	118576119	+	11	118867843	+	INV	intergenic	intergenic		
T3	SV_85	11	118802233	-	11	118888669	+	DUP	intergenic	RPS25		
T3	SV_86	11	118841362	-	11	118888405	-	INV	intergenic	RPS25		

TIME POINT	SV ID	CHR	POS	STRAND	CHR	POS	STRAND	SV TYPE	LOCATION (bkp1)	LOCATION (bkp2)	CLL DRIVERS	DLBCL DRIVERS
T3	SV_187	11	118867609	-	11	118867870	-	INV	intergenic	intergenic		
T3	SV_102	13	26926978	+	13	47816848	+	INV	CDK8	intergenic		FOXO1
T3	SV_103	13	27116592	-	13	47841044	+	DUP	intergenic	intergenic		FOXO1
T3	SV_106	13	47816186	-	13	47816851	-	INV	intergenic	intergenic		
T3	SV_107	13	47816221	+	13	57245235	-	DEL	intergenic	intergenic		
T3	SV_108	13	48933458	+	13	48990742	-	DEL	RB1	RB1		
T3	SV_160	X	147727409	+	X	147738814	+	INV	AFF2	AFF2		
T3	SV_161	X	147731093	-	X	147738787	-	INV	AFF2	AFF2		

Supplementary Table 1. Somatic structural variants identified by variant callers and manually validated in CLL case 63. Known CLL and DLBCL driver genes are annotated if involved within the structural variant.

TIME POINT	SV ID	LENGTH	BREAKPOINT 1				BREAKPOINT 2				AVERAGE		DIFFERENCE (BKP)	
			REPEAT1	VAF1	CNA1	CCF1	REPEAT2	VAF2	CNA2	CCF2	VAF	CCF	VAF	CCF
T1-PB	SV_6	3790138	LINE/L1	0,3120	2	0,5940	SINE/Alu	0,4961	2	1,0224	0,4041	0,8082	0,1842	0,4285
T1-PB	SV_7	8030630		0,3584	2	0,7386		0,3238	2	0,6673	0,3411	0,7029	0,0346	0,0714
T1-PB	SV_11	11406		0,2727	2	0,5583		0,3077	2	0,6299	0,2902	0,5941	0,0350	0,0716
T1-PB	SV_12	7702		0,2500	2	0,5118	SINE/MIR	0,3077	2	0,6299	0,2788	0,5708	0,0577	0,1181
T1-LN	SV_9	3790138	LINE/L1	0,3796	1,86	0,7354	SINE/Alu	0,6216	2	1,3036	0,5006	1,0195	0,2420	0,5682
T1-LN	SV_10	8030630		0,4000	2	0,8389		0,4340	1,4133	0,6389	0,4170	0,7389	0,0340	0,2000
T1-LN	SV_18	11406		0,1806	2	0,3762		0,2284	2	0,4757	0,2045	0,4757	0,0478	0,0996
T1-LN	SV_19	7695		0,1500	2	0,3125	SINE/MIR	0,2941	2	0,6127	0,2221	0,4626	0,1441	0,3002
T2	SV_21	3790138	LINE/L1	0,4424	2	0,9183	SINE/Alu	0,8884	2	1,8440	0,6654	1,3811	0,4460	0,9257
T2	SV_22	8030630		0,3539	2	0,7346		0,3936	1,4133	0,5735	0,3738	0,6541	0,0397	0,1610
T2	SV_24676	20889870	LINE/L1	0,0732	2	0,1519	SINE/Alu	0,0706	2	0,1466	0,0719	0,1492	0,0025	0,0052
T2	SV_26	666		na	na	na		na	na	na	na	na	na	na
T2	SV_27	9429015		0,1396	2	0,2897	LTR/ERVL	0,1709	2	0,3547	0,1552	0,3222	0,0313	0,0650
T2	SV_46	11406		0,7293	2	1,5036		0,5983	2	1,2336	0,6638	1,3686	0,1310	0,2700
T2	SV_47	7695		0,7295	2	1,5042	SINE/MIR	0,6187	2	1,2758	0,6741	1,3900	0,1108	0,2284
T3	SV_16	311		na	na	na		na	na	na	na	na	na	na
T3	SV_29	4459990	LINE/L1	0,1197	3	0,3796		0,0972	3	0,3085	0,1085	0,3441	0,0224	0,0711
T3	SV_30	47477		0,0889	3	0,2819		0,2804	3	0,8895	0,1846	0,5857	0,1915	0,6076
T3	SV_56	450		na	na	na		na	na	na	na	na	na	na

TIME POINT	SV ID	LENGTH	BREAKPOINT 1				BREAKPOINT 2				AVERAGE		DIFFERENCE (BKP)	
			REPEAT1	VAF1	CNA1	CCF1	REPEAT2	VAF2	CNA2	CCF2	VAF	CCF	VAF	CCF
T3	SV_66	23111532	LINE/L2	0,3093	2	0,6314		0,2985	2	0,6768	0,0215	0,6541	0,0107	0,0454
T3	SV_308		SINE/Alu	0,1667	2	0,3525		0,1667	3	0,5287	0,1667	0,4406	0,0000	0,1762
T3	SV_89		DNA/hAT-Charlie	0,3158	2	0,6678	SINE/Alu	0,4667	1	0,4935	0,3912	0,5680	0,1509	0,1744
T3	SV_309		Simple repeat	0,1000	1	0,1057	Simple repeat	0,4000	3	1,2689	0,2500	0,6873	0,3000	1,1632
T3	SV_88			0,4522	1	0,4781		0,3507	3	1,1124	0,4014	0,7953	0,1015	0,6342
T3	SV_72	164		na	na	na		na	na	na	na	na	na	na
T3	SV_75	1152	LINE/L1	0,0592	3	0,1877	LINE/L1	0,0617	3	0,1958	0,0605	0,1918	0,0025	0,0081
T3	SV_184	7028		0,1304	3	0,4137		0,1693	3	0,5370	0,1498	0,4753	0,0389	0,1233
T3	SV_79	5169		0,1667	3	0,5287		0,0588	3	0,1866	0,1127	0,3577	0,1078	0,3421
T3	SV_80	1613		0,2543	3	0,8066		0,2295	3	0,7281	0,2419	0,7673	0,0247	0,0785
T3	SV_82	3790138	LINE/L1	0,3855	2,88	1,1662	SINE/Alu	0,7948	2	1,6809	0,5901	1,4235	0,4093	0,5147
T3	SV_83	8030630		0,4128	2	0,8731		0,4203	2,42	1,0715	0,4166	0,9723	0,0075	0,1984
T3	SV_84	291725	LTR/Gypsy	0,3210	2	0,6789		0,3202	3	1,0158	0,3206	0,8473	0,0008	0,3369
T3	SV_85	86437		0,4409	1	0,4662		0,1758	3	0,5577	0,3084	0,5120	0,2651	0,0915
T3	SV_86	47044	LINE/L1	0,4444	1	0,4700		0,2830	3	0,8978	0,3637	0,6839	0,1614	0,4278
T3	SV_187	262		na	na	na		na	na	na	na	na	na	na
T3	SV_102	20889871		0,3636	2	0,7690	LINE/L1	0,3721	2	0,7869	0,3679	0,7780	0,0085	0,0179
T3	SV_103	20724453		0,3056	2	0,6462	LTR/ERV1	0,2813	2	0,5948	0,2934	0,6205	0,0243	0,0514
T3	SV_106	666		na	na	na		na	na	na	na	na	na	na

TIME POINT	SV ID	LENGTH	BREAKPOINT 1				BREAKPOINT 2				AVERAGE		DIFFERENCE (BKP)	
			REPEAT1	VAF1	CNA1	CCF1	REPEAT2	VAF2	CNA2	CCF2	VAF	CCF	VAF	CCF
T3	SV_107	9429015		0,3527	2	0,7460		0,2859	2	0,1000	0,3193	0,6753	0,0669	0,6460
T3	SV_108	57285		0,5349	1	0,5656	LINE/L1	0,5395	1	0,2000	0,5372	0,5680	0,0046	0,3656
T3	SV_160	11406		0,6769	2	1,4316		0,6818	2	0,3100	0,6794	1,4368	0,0049	1,1216
T3	SV_161	7695		0,5966	2	1,2617	SINE/MIR	0,6158	2	0,3000	0,6062	1,2820	0,0192	0,9617

Supplementary Table 2. Variant allele frequency and cancer cell fraction calculated for each breakpoint and structural variant (average) detected in CLL tumors from case 63. Difference between VAF and CCF calculated for each BKP of a SV are also shown in the last two columns. The frequency of those variants marked in light grey was not calculated because they were shorter than the threshold used.

T1-PB	T1-LN	T2	T3	Structural Variant	Samples
0,81	1,02	1,38	1,42	11:106417594_11:110207731	All
0,70	0,74	0,65	0,97	11:108122764_11:116153393	All
0,59	0,48	1,37	1,44	X:147727409_X:147738814	All
0,57	0,46	1,39	1,28	X:147731093_X:147738794	All
0,00	0,00	0,15	0,78	13:26926979_13:47816848	Chemoimmunotherapy
0,00	0,00	0,32	0,68	13:47816221_13:57245235	Chemoimmunotherapy
0,00	0,00	0,00	0,34	4:136358882_4:140818871	Richter
0,00	0,00	0,00	0,59	4:140808906_4:140856382	Richter
0,00	0,00	0,00	0,65	9:8495226_9:31606757	Richter
0,00	0,00	0,00	0,44	9:131108130_11:118477716	Richter
0,00	0,00	0,00	0,57	9:131209465_11:118814376	Richter
0,00	0,00	0,00	0,69	9:131222219_11:118861651	Richter
0,00	0,00	0,00	0,80	9:131231071_11:118477448	Richter
0,00	0,00	0,00	0,19	11:7897130_11:7898281	Richter
0,00	0,00	0,00	0,48	11:33731887_11:33738914	Richter
0,00	0,00	0,00	0,36	11:63968132_11:63973300	Richter
0,00	0,00	0,00	0,77	11:67352245_11:67353857	Richter
0,00	0,00	0,00	0,85	11:118576119_11:118867843	Richter
0,00	0,00	0,00	0,51	11:118802233_11:118888669	Richter
0,00	0,00	0,00	0,68	11:118841362_11:118888405	Richter
0,00	0,00	0,00	0,62	13:27116592_13:47841044	Richter
0,00	0,00	0,00	0,57	13:48933458_13:48990742	Richter

Supplementary Table 3 – Cancer cell fraction calculated for each somatic SV across all longitudinal and spatial CLL samples (case 63). Last column indicates if the SV was detected in all samples (All), in samples after chemoimmunotherapy (T2 and T3) (Chemoimmunotherapy) or only once the tumor transformed into DLBCL; T3 sample (Richter).

ICGC PROJECT	SNVs	Parameters	Clusters	Elements	P-values > 0,01	KS statistic	KS p-value	CG enrichment
BRCA-FR	354.963	51-51	21	19	19	0,4428	6,52E-04	0,577
		71-71	23	21	21	0,3912	2,09E-03	0,596
		91-91	27	25	25	0,307	1,37E-02	0,906
		101-101	28	26	26	0,2951	1,68E-02	0,792
		101-91-101	28	26	24	0,2877	2,98E-02	0,6
		101-101-101	28	26	26	0,3381	3,73E-03	0,618
CLLE-ES	384.029	51-51	25	23	21	0,3984	1,61E-03	4,41
		71-71	26	25	23	0,4311	2,12E-04	4,18
		91-91	26	26	24	0,3189	1,14E-02	4,38
		101-101	27	27	25	0,3053	1,45E-02	4,37
		101-91-101	27	27	20	0,277	7,55E-02	4,05
		101-101-101	27	27	24	0,3668	2,11E-03	4,01
ESAD-UK	8.694.577	51-51	1.046	583	565	0,1975	8,80E-20	2,093
		71-71	1.217	690	670	0,1771	7,45E-19	1,83
		91-91	1.346	776	755	0,1657	1,37E-18	1,251
		101-101	1.418	835	816	0,1762	1,26E-22	1,63
		101-91-101	1.429	835	796	0,1953	4,54E-27	2,922
		101-101-101	1.418	835	792	0,1942	1,24E-26	3,004
PACA-AU	966.241	51-51	67	64	59	0,3764	5,10E-08	7,087
		71-71	78	71	66	0,3287	7,50E-07	7,111
		91-91	82	75	72	0,3068	1,63E-06	6,676
		101-101	82	75	73	0,3258	2,13E-07	6,694
		101-91-101	82	75	64	0,3255	1,54E-06	6,992
		101-101-101	82	75	68	0,351	5,40E-08	7,031
PACA-CA	1.234.194	51-51	97	84	80	0,1755	1,26E-02	6,043
		71-71	110	94	89	0,1454	4,17E-02	5,919
		91-91	127	104	100	0,1647	7,72E-03	5,565
		101-101	132	108	104	0,1615	7,75E-03	5,392
		101-91-101	134	108	97	0,2143	2,18E-04	5,608
		101-101-101	132	108	100	0,2001	5,52E-04	5,836
LUSC-KR	1.144.634	51-51	231	112	89	0,501	3,45E-21	0,24
		71-71	237	115	94	0,4409	3,69E-17	0,261
		91-91	245	123	103	0,3868	2,20E-14	0,151
		101-101	249	125	107	0,3566	1,10E-12	0,337
		101-91-101	247	125	88	0,3381	1,77E-09	0,222
		101-101-101	240	125	91	0,3525	1,28E-10	0,337

Supplementary Table 4. OncodriveCLUSTL tests done for each ICGC project based on different combinations of parameters. For each test, the number of clusters identified, the KS statistic and the enrichment in cancer genes are shown. Marked in green, the selected parameter combination for each ICGC project. Only those sets of variants with more than 20 clusters of variants were evaluated and are represented here.

smORF	Ribo-seq	Literature Mining	MS	Databases
smOrf_24636	.	SPROHSA006339	.	.
smOrf_28297	.	SPROHSA009881	.	.
		SPROHSA175114,		
		SPROHSA176941,		
		SPROHSA178278,		
smOrf_6018		SPROHSA180088,		
		SPROHSA180602,		
		SPROHSA183712,		
		SPROHSA018303	.	.
smOrf_19623	.	SPROHSA013707	.	.
smOrf_21016	.	SPROHSA013708	.	.
smOrf_44005	.	SPROHSA028972	.	.
smOrf_46135	.	SPROHSA018304	.	.
smOrf_36958	.	SPROHSA005007	.	.
smOrf_22067	.	SPROHSA009188	.	.
smOrf_42179	.	SPROHSA009880	.	.
smOrf_32101	.	SPROHSA011377	.	.
smOrf_19623	.	SPROHSA013707	.	.
smOrf_21016	.	SPROHSA013708	.	.
smOrf_19623	.	SPROHSA013707	.	.
smOrf_21016	.	SPROHSA013708	.	.
smOrf_43215	.	SPROHSA012045	.	.
smOrf_35962	.	SPROHSA011384	.	.
smOrf_34454	.	SPROHSA011390	.	.
smOrf_24181	.	SPROHSA003433	.	.
smOrf_30964	.	SPROHSA005004	.	.
smOrf_38621	.	SPROHSA011391	.	.
smOrf_29903	.	SPROHSA019697	.	.
smOrf_54382	.	.	.	SPROHSA141874
smOrf_21297	.	SPROHSA011394	.	.

Supplementary Table 5. Small ORF ID (first column) and their corresponding ID in the SmProt database, indicating also if they were included in the database because of being detected by Ribo-seq experiments, or MS, as well as if they were previously published (literature mining) or appeared in other databases.

10.3 Publications

Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition.

OPEN

Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition

Bernardo Rodriguez-Martin ^{1,2,3}, Eva G. Alvarez^{1,2,3,920}, Adrian Baez-Ortega^{4,920}, Jorge Zamora^{1,3,5,6,920}, Fran Supek ^{7,8,920}, Jonas Demeulemeester ^{9,10}, Martin Santamarina^{1,2,11}, Young Seok Ju ^{12,13}, Javier Temes ^{1,3}, Daniel Garcia-Souto ¹¹, Harald Detering^{2,14,15}, Yilong Li⁶, Jorge Rodriguez-Castro¹¹, Ana Dueso-Barroso^{16,17}, Alicia L. Bruzos^{1,2,11}, Stefan C. Dentre^{9,18,19}, Miguel G. Blanco ^{20,21}, Gianmarco Contino²², Daniel Ardeljan ²³, Marta Tojo^{5,14}, Nicola D. Roberts ⁶, Sonia Zumalave ¹¹¹, Paul A. Edwards ^{24,25}, Joachim Weischenfeldt ^{26,27,28}, Montserrat Puiggròs¹⁷, Zechen Chong^{29,30}, Ken Chen ³¹, Eunjung Alice Lee³², Jeremiah A. Wala ^{33,34,35,36}, Keiran M. Raine ¹³, Adam Butler¹³, Sebastian M. Waszak ²⁶, Fabio C. P. Navarro ^{37,38,39}, Steven E. Schumacher^{33,34,35}, Jean Monlong ⁴⁰, Francesco Maura^{13,41,42}, Niccolò Bolli^{41,42}, Guillaume Bourque ⁴³, Mark Gerstein^{37,38,39}, Peter J. Park ^{44,45}, David C. Wedge^{13,18,46}, Rameen Beroukhi^{33,35,36}, David Torrents^{8,17}, Jan O. Korbel ^{26,47}, Iñigo Martincorena⁶, Rebecca C. Fitzgerald²², Peter Van Loo ^{9,10}, Haig H. Kazazian²³, Kathleen H. Burns ^{48,49}, PCAWG Structural Variation Working Group⁵⁰, Peter J. Campbell ^{6,51,921} ⁵², Jose M. C. Tubio ^{1,2,3,13,921} ⁵³ and PCAWG Consortium⁵²

About half of all cancers have somatic integrations of retrotransposons. Here, to characterize their role in oncogenesis, we analyzed the patterns and mechanisms of somatic retrotransposition in 2,954 cancer genomes from 38 histological cancer subtypes within the framework of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. We identified 19,166 somatically acquired retrotransposition events, which affected 35% of samples and spanned a range of event types. Long interspersed nuclear element (LINE-1; L1 hereafter) insertions emerged as the first most frequent type of somatic structural variation in esophageal adenocarcinoma, and the second most frequent in head-and-neck and colorectal cancers. Aberrant L1 integrations can delete megabase-scale regions of a chromosome, which sometimes leads to the removal of tumor-suppressor genes, and can induce complex translocations and large-scale duplications. Somatic retrotranspositions can also initiate breakage-fusion-bridge cycles, leading to high-level amplification of oncogenes. These observations illuminate a relevant role of L1 retrotransposition in remodeling the cancer genome, with potential implications for the development of human tumors.

LINE-1 retrotransposons are widespread repetitive elements in the human genome, representing 17% of the entire DNA content^{1,2}. Using a combination of cellular enzymes and self-encoded proteins with endonuclease and reverse transcriptase activity, L1 elements copy and insert themselves at new genomic sites, in a process called retrotransposition. Most of the approximately 500,000 L1 copies in the human reference genome are truncated, inactive elements that are unable to retrotranspose. A small subset of them, around 100–150 L1 loci, remain active in the average human genome, acting as source elements, a small number of which consists of highly active copies termed hot-L1s^{3–5}. These L1 source elements are usually transcriptionally repressed, but epigenetic changes that occur in tumors may promote their expression and allow them to retrotranspose^{6,7}. Somatic L1 retrotransposition usually introduces a new copy of the 3' end of the L1 sequence, and can also mobilize unique DNA sequences located immediately downstream of the source element, in a process called 3' transduction^{8–9}. L1 retrotransposons

can also promote the somatic trans-mobilization of Alu elements, SINE-VNTR-Alu (SVA) elements and processed pseudogenes, which are copies of mRNAs that have been reverse transcribed into DNA and inserted into the genome with the machinery of active L1 elements^{10–12}.

Approximately 50% of human tumors contain somatic retrotranspositions of L1 elements^{13–15}. Previous analyses indicate that although a fraction of somatically acquired L1 insertions in cancer may influence gene function, the majority of retrotransposon integrations in a single tumor represent passenger mutations with little or no effect on cancer development^{7,16}. Nonetheless, L1 elements are capable of promoting other types of genomic structural alterations in the germline and somatically, in addition to canonical L1 insertion events^{16–18}; the effect of these alterations remains largely unexplored in the context of human cancer¹⁹.

To further understand the roles of retrotransposons in cancer, we developed strategies to analyze the patterns and mechanisms of

A full list of authors and affiliations appears at the end of the paper.

somatic retrotransposition in 2,954 cancer genomes from 38 histological cancer subtypes within the framework of the PCAWG project¹, many of which had not been evaluated for retrotransposition. On the basis of the robustness of the retrotransposition calls, we retained 296 tumors that were preliminarily excluded by the PCAWG Consortium²¹ (see Methods). Our analyses identify patterns and mutational mechanisms of structural variation in human cancers that are mediated by L1 retrotransposition. We found that the aberrant integration of L1 retrotransposons has a relevant role in remodeling the architecture of the cancer genome in some human tumors, mainly by promoting megabase-scale deletions that, occasionally, generate genomic consequences that may promote cancer development through the removal of tumor-suppressor genes, such as *CDKN2A*, or trigger the amplification of oncogenes, such as *CCND1*.

Results

The landscape of somatic retrotransposition in a large cancer whole-genome dataset. We ran our bioinformatic pipelines (Methods and Supplementary Note) to explore somatic retrotransposition on whole-genome sequencing data from 2,954 tumors and their matched normal pairs, across 38 cancer types (Supplementary Fig. 1 and Supplementary Table 1). The analysis retrieved a total of 19,166 somatically acquired retrotranspositions that were classified into six categories (Fig. 1a and Supplementary Table 2). Comprising 98% (18,739 out of 19,166) of the events, L1 integrations (14,967 solo-L1, 3,669 L1-transductions, and 103 L1-mediated rearrangements, which mainly comprised deletions) overwhelmingly dominate the landscape of somatic retrotransposition in the PCAWG dataset (Fig. 1a,b). By contrast, elements of the lineages Alu (Supplementary Fig. 2) and SVA (comprising 130 and 23 somatic copies, respectively) and processed pseudogenes, with 274 events, represent minor categories.

The core pipeline, TraFiC-mem (Supplementary Fig. 3)—which was used to explore somatic retrotransposition in PCAWG—was validated by single-molecule whole-genome sequencing data analysis of one cancer cell line with high retrotransposition rate and its matched normal sample, confirming the somatic acquisition of 295 out of 308 retrotransposition events (false discovery rate <5%, Supplementary Fig. 4a,b). To further evaluate TraFiC-mem, we reanalyzed a mock cancer genome into which we had previously seeded somatic retrotransposition events at different levels of tumor clonality, and then simulated sequencing reads to the average level of coverage of the PCAWG dataset. The results confirmed a high precision (>99%) of TraFiC-mem, and a recall ranging from 90 to 94% for tumor clonalities from 25 to 100%, respectively (Supplementary Fig. 4c–e).

We observed marked variation in the retrotransposition rate across PCAWG tumor types (Fig. 1c and Supplementary Table 3). Overall, 35% (1,046 out of 2,954) of all cancer genomes have at least one retrotransposition event. However, esophageal adenocarcinoma, head-and-neck squamous carcinoma, lung squamous carcinoma and colorectal adenocarcinoma are significantly enriched in somatic retrotranspositions (Mann–Whitney *U*-test, $P < 0.05$; Fig. 1c,d and Supplementary Fig. 5). These four tumor types alone account for 70% (13,373 out of 19,166) of all somatic events in the PCAWG dataset, although they represent just 9% (266 out of 2,954) of the samples. This is particularly noticeable in esophageal adenocarcinoma, in which 27% (27 out of 99) of the samples show more than 100 separate somatic retrotranspositions (Fig. 1c), making L1 insertions the most frequent type of structural variation in esophageal adenocarcinoma (Fig. 1c). Furthermore, retrotranspositions are the second-most frequent type of structural variants in head-and-neck squamous and colorectal adenocarcinomas (Fig. 1c). To gain insights into the genetic causes that make some cancers more prone to retrotransposition than others, we looked for associations

between retrotransposition and driver mutations in cancer-related genes. This analysis revealed an increased L1 retrotransposition rate in tumors with *TP53* mutations (Mann–Whitney *U*-test, $P < 0.05$; Supplementary Fig. 6), and supports previous analyses that have suggested that *TP53* functions to restrain mobile elements^{22,23}. We also observe a widespread correlation between L1 retrotransposition and other types of structural variation (Spearman's $\rho = 0.44$, $P < 0.01$; Supplementary Fig. 7), a finding that is most likely a consequence of a confounding effect of *TP53*-mutated genotypes (Supplementary Fig. 6).

We identified 43% (7,979 out of 18,636) somatic retrotranspositions of L1 inserted within gene regions including promoters, of which 66 events hit cancer-associated genes. The analysis of expression levels in samples with available transcriptome data, revealed four genes—including the *ABL* oncogene—with L1 retrotranspositions in the proximity of promoter regions that showed significant overexpression compared with the expression in the remaining samples of the same tumor type (Student's *t*-test, $q < 0.10$; Supplementary Fig. 8a–c). The structural analysis of RNA-sequencing data identified instances in which portions of a somatic retrotransposition within a gene exonize, a process that sometimes involves cancer-associated genes (Supplementary Fig. 8d). In addition, we found evidence of aberrant fusion transcripts arising from the inclusion of processed pseudogenes in the target host gene and expression of processed pseudogenes landing in intergenic regions (Supplementary Fig. 8e).

Dissecting the genomic features that influence the landscape of L1 retrotranspositions in cancer. The genome-wide analysis of the distribution of somatic L1 insertions across the cancer genome revealed considerable variation in the rate of L1 retrotransposition (Fig. 2a and Supplementary Table 4). To understand the reasons behind such variation, we studied the association of L1 event rates with various genomic features. We first investigated whether the distribution of somatic L1s across the cancer genome could be determined by the occurrence of L1-endonuclease target-site motifs. We used a statistical approach based on negative binomial regression to deconvolute the influence of multiple overlapping genomic variables²⁴; this analysis showed that close matches to the motif have a 244-fold increased L1 rate, compared with non-matched motifs (Fig. 2b and Supplementary Fig. 9a). Adjusting for this effect, we found a strong association with DNA replication time; the latest-replicating quarter of the genome was 8.9-fold enriched in L1 events (95% confidence interval, 8.25–9.71) compared with the earliest-replicating quarter (Fig. 2b,c and Supplementary Fig. 9b). Recent work²⁵ has shown that L1 retrotransposition has a strong cell-cycle bias, and preferentially occurs during S phase. Our results are in agreement with these findings and suggest that L1 retrotransposition peaks in the later stages of nuclear DNA synthesis.

Next, we examined L1 rates in open chromatin measured using DNase hypersensitivity and, conversely, in closed heterochromatic regions by analyzing K9-trimethylated histone H3 (H3K9me3)²⁶. When adjusting for the confounding effects of L1 motif content and replication time²⁴, we found that somatic L1 events are enriched in open chromatin (1.27-fold in the top DNase hypersensitivity bin; 95% confidence interval, 1.14–1.41; Fig. 2b) and depleted in heterochromatin (1.72-fold, 95% confidence interval, 1.57–1.99; Fig. 2b). This finding differs from previous analyses, which have suggested that L1 insertions favored heterochromatin²⁷—a discrepancy that we believe to be due to the confounding effect between heterochromatin and late-replicating DNA regions, which was not addressed in previous analyses. We also found a negative association of L1 rate with features of active transcription of chromatin, characterized by fewer L1 events at active promoters (1.63-fold; Supplementary Fig. 9c), a slight but significant reduction in L1 rates in highly expressed genes (1.25-fold lower; 95% confidence interval, 1.16–1.34; Fig. 2b) and a further depletion at H3K36me3 (1.90-fold reduction in the highest

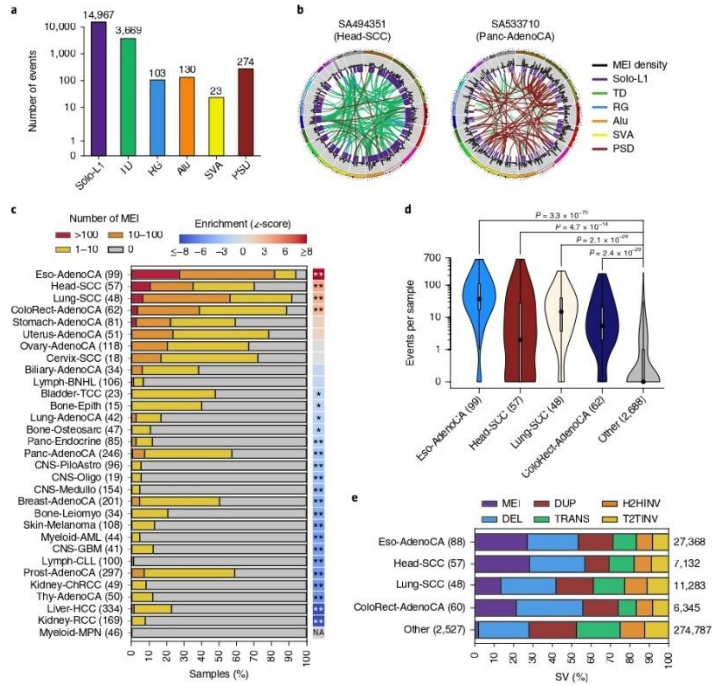


Fig. 1 | Landscape of somatic retrotransposition across human cancers. a Number of somatic retrotransposition events identified in 2,954 cancer genomes across six categories: solo-L1, L1-mediated transductions (TD), L1-mediated rearrangements (RG), Alu, SVA and pseudogenes (PSD). **b** Left, circos plot showing a head-and-neck tumor (Head-SCC) with high retrotransposition rate (638 somatic events). Right, a single pancreatic adenocarcinoma sample harboring around 26% (70 out of 274) of all processed pseudogenes identified in the PCAWG cohort. Chromosome ideograms are shown around the outer ring with individual rearrangements represented as arcs; colors match the type of rearrangement. **c** For 31 PCAWG cancer types with sample size of $n \geq 15$, data show the proportion of tumor samples with >100 (red), 10–100 (orange), 1–10 (yellow) and 0 (gray) somatic retrotranspositions. The number of samples analyzed for each tumor type is shown in parentheses. Retrotransposition enrichment or depletion for each tumor type together with the level of significance (zero-inflated negative binomial regression) is shown. * $P < 0.05$, ** $P < 0.01$, NA, not applicable. **d** Distribution of retrotransposition events per sample across the four tumor types significantly enriched in somatic retrotranspositions; the remaining tumors are grouped into ‘Other’. The number of samples from each group is shown in parentheses; point, median; box, 25th to 75th percentiles (interquartile range); whiskers, data within 1.5x the interquartile range. P values indicate significance from a two-tailed Mann–Whitney U -test. The y axis is shown on a logarithmic scale. **e** For the same four tumor types in **d**, the fraction of structural variants (SV) belonging to six classes is shown: mobile element insertions (MEI), deletions (DEL), duplications (DUP), translocations (TRANS), head-to-head inversions (H2HINV) and tail-to-tail inversions (T2TINV). The total number of structural variants per cancer type is indicated on the right side of the panel.

tertile; 95% confidence interval, 1.59–2.29; Fig. 2b), a mark of actively transcribed regions deposited in the body and at the 3' end of active genes³⁰. Further details on these associations are shown in Supplementary Fig. 9c–e and described in the Supplementary Note.

The contribution of L1 source elements to the pan-cancer retrotransposition burden. We used somatically mobilized L1 3' transduction events to trace L1 activity to specific source elements³¹. This strategy revealed 124 germline L1 loci in the human genome

that are responsible for most of the genomic variation generated by retrotransposition in the PCAWG dataset²¹ (Supplementary Table 5). To our knowledge, 52 of these loci represent previously unreported source elements in human cancer³¹. We analyzed the relative contribution of individual source elements to retrotransposition burden across cancer types, and found that retrotransposition is generally dominated by five hot-L1 source elements that alone give rise to half of all somatic transductions (Fig. 3a). This analysis revealed a dichotomous pattern of hot-L1 activity, with source

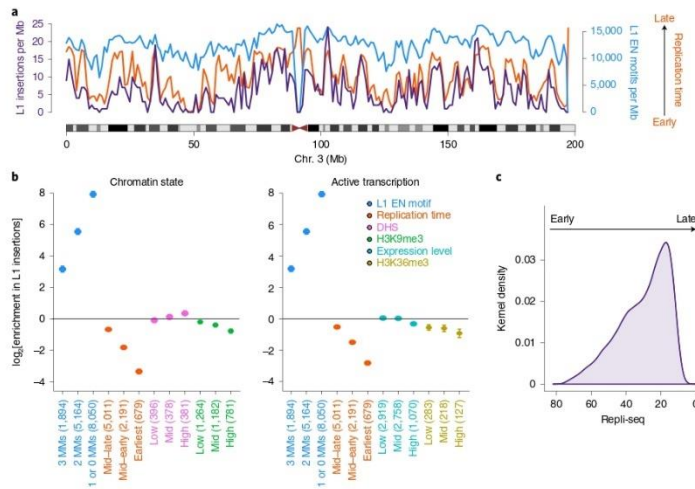


Fig. 2 | Distribution of L1 somatic insertions across the cancer genome and its association with genome organization features. Genome-wide analysis of the distribution of 15,906 somatic L1 insertions, which include solo-L1 and L1 transductions with a 3'-poly(A) breakpoint characterized to base-pair resolution. **a**, The L1 insertion rate (purple) is shown together with the L1 endonuclease (EN) motif density (blue) and replication timing (orange). The data are represented per 1-Mb window. For illustrative purposes, only chromosome 3 is shown. **b**, Association between L1 insertion rate and multiple predictor variables at single-nucleotide resolution. Enrichment scores (thick dots) are adjusted for multiple covariates and compare the L1 insertion rate in bins 1–3 for a particular genomic feature (L1 endonuclease motif, replication timing, open chromatin, histone marks and expression level) versus bin 0 of the same feature, which therefore always has log-transformed enrichment = 0 by definition and is not shown. The error bars represent 95% confidence intervals. The number of observations per bin is provided in parentheses. MMs, the number of mismatches with respect to the consensus L1 endonuclease motif (see Supplementary Note). Heterochromatic regions and transcription elongation are defined based on H3K9me3 and H3K36me3 histone marks. Accessible chromatin is measured through DNase hypersensitivity. **c**, L1 insertion density, using kernel density estimate (KDE), along the replication timing spectrum. DNA replication timing is expressed on a scale from 80 (early) to 0 (late).

elements that we have termed Strombolian and Plinian, given their similarity to these two types of volcanoes (Fig. 3b). Strombolian source elements are relatively indolent and produce small numbers of retrotranspositions in individual tumor samples, although they are often active and contribute substantially to overall retrotransposition in the PCAWG dataset. By contrast, Plinian elements are rarely active across tumors, but in these isolated cases, their activity is fulminant, causing large numbers of retrotranspositions.

At the individual tumor level, although we observed that the number of active source elements in a single cancer genome varied from 1 to 22, typically only 1 to 3 loci were operative (Fig. 3c). There is a correlation between somatic retrotranspositions and the number of active germline L1 source elements among PCAWG samples (Fig. 3d); this is likely one of the factors that explains why esophageal adenocarcinoma, lung and head-and-neck squamous carcinoma account for higher retrotransposition rates—in these three tumor types we also observed higher numbers of active germline L1 loci (Fig. 3c). Occasionally, somatic L1 integrations that retain their full length may also act as a source for subsequent somatic retrotransposition events²⁷, and may reach high activity rates, leading them to dominate retrotransposition in a given tumor. For example, in a remarkable head-and-neck tumor sample, SA197656, we identified one somatic L1 integration at 4p16.1 that then triggered 18 transductions from its new site, with the next most active element

being a germline L1 locus at 22q12.1, which accounted for 15 transductions (Supplementary Table 5).

Genomic deletions mediated by somatic L1 retrotransposition.

In cancer genomes with high somatic L1 activity rates, we observed that some L1 retrotransposition events followed a distinctive pattern that consisted of a single cluster of reads, associated with copy-number loss, for which the mates unequivocally identified one extreme of a somatic L1 integration with, apparently, no local, reciprocal cluster that supported the other extreme of the L1 insertion (Fig. 4a). Analysis of the associated copy-number changes identified the missing L1 reciprocal cluster at the far end of the copy-number loss, indicating that this pattern represents a deletion that occurred in conjunction with the integration of an L1 retrotransposon (Fig. 4b; see the Supplementary Note for additional information on how to interpret the paired-end mapping data from this and other figures). These rearrangements—called L1-mediated deletions—have been observed to occur somatically with engineered L1s in cultured human cells^{14,15} and naturally in the brain¹⁶, and are most likely the consequence of an aberrant mechanism of L1 integration.

We developed specific algorithms to systematically identify L1-mediated deletions, and applied these methods across all PCAWG tumors. We identified 90 somatic events that matched the patterns described above, causing deletions of different size, which ranged

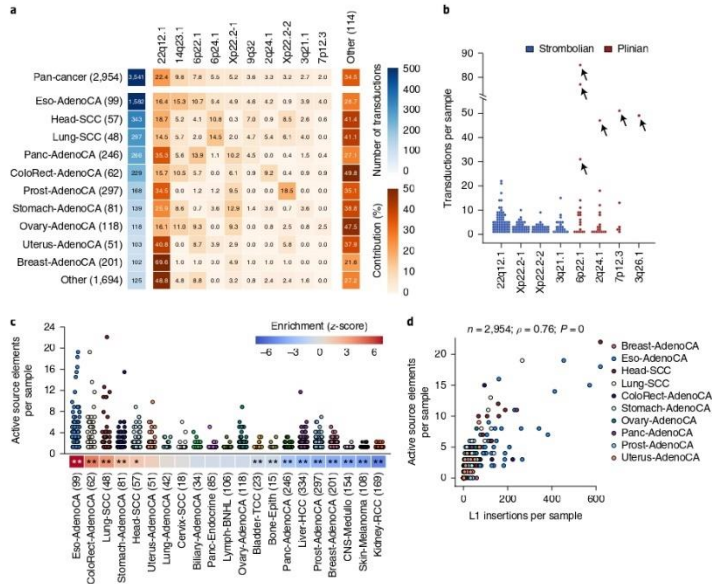


Fig. 3 | The dynamics of L1 source-element activity in human cancer. a, The total number of transductions identified for each cancer type is shown as a blue-colored scale. The sample size for each tumor type is shown in parentheses. Contribution of each source element is defined as the proportion of the total number of transductions from each cancer type that is explained by each source locus. Only the top ten contributing source elements are shown, while the remaining are grouped into the category ‘Other’. **b**, Two extreme patterns of hot-L1 activity, Strombolian (blue) and Plinian (red), were identified. Dots show the number of transductions promoted by each source element in a given tumor sample. Arrows highlight violent eruptions (that is, strong peaks of somatic activity) in particular samples. **c**, Number of active germline L1 source elements per sample, across cancer types with source element activity. A source element is considered to be active in a given sample if it promotes at least one transduction. The enrichment or depletion of the number of active source elements for each tumor type together with the level of significance (zero-inflated negative binomial regression) is shown. * $P < 0.05$, ** $P < 0.01$. The number of samples analyzed for each tumor type is shown in parentheses. **d**, Correlation between the number of somatic L1 insertions and the number of active germline L1 source elements in PCAWG samples. Each dot represents a tumor sample and colors match cancer types. Sample sizes (n), together with Spearman’s ρ and P values are shown above the panel.

in size from around 0.5kb to 53.4Mb (Fig. 4c and Supplementary Table 6). The reconstruction of the sequence at the breakpoint junctions in each case supports the presence of an L1-element—or L1-transduction—sequence and its companion polyadenylate tract, indicative of passage through an RNA intermediate. No target site duplication was found, which is also the typical pattern for L1-mediated deletions¹⁷. One potential mechanism for these events is that a molecule of L1 cDNA pairs with a distant 3′ overhang from a pre-existing double-strand DNA break generated upstream of the initial integration site, and the DNA region between the break and the original target site is subsequently removed by aberrant repair¹⁷ (Fig. 4d). Indeed, in 75% (47 out of 63) of L1-mediated deletions with a 5′-end breakpoint characterized to base-pair resolution, the analysis of the sequences at the junction revealed short (1–5 bp long, with median at 3 bp) microhomologies between the pre-integration site and the 5′ L1 sequence integrated right there (Supplementary Table 6). Furthermore, we found 14% (9 out of 63) instances in

which short insertions (1–33 bp long, with median at 9 bp) are found at the 5′-breakpoint junction of the insertion. Both signatures are consistent with a non-homologous end-joining mechanism¹⁸, or other type of microhomology-mediated repair, for the 5′-end attachment of the L1 cDNA to a 3′ overhang from a pre-existing double-strand DNA break located upstream. L1-mediated deletions in which microhomologies or insertions are not found may follow alternative models^{17,29–31}.

To confirm that these rearrangements are mediated by the integration of a single intervening retrotransposition event, we explored the PCAWG dataset for somatic L1-mediated deletions in which the L1 sequences at both breakpoints of the deletion could unequivocally be assigned to the same L1 insertion. These include small deletions and associated L1 insertions that were shorter than the library size, allowing sequencing read pairs to overlay the entire structure. For example, in a lung tumor sample, SA313800, we identified a deletion involving a 1-kb region of 19q12 with hallmarks of

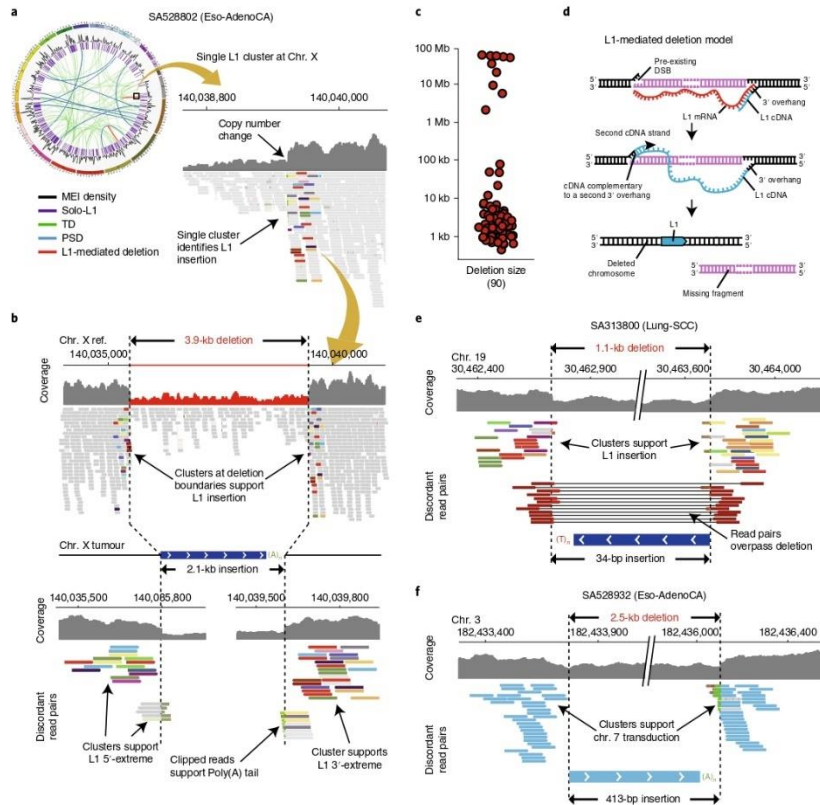


Fig. 4 | The hallmarks of somatic L1-mediated deletions revealed by copy-number and paired-end mapping analysis. a, In esophageal adenocarcinoma sample SA528802, we found a single cluster of reads on chromosome X, which is associated with one breakpoint of a copy-number loss, and for which the mates unequivalently identified one extreme of a somatic L1 integration. Paired-end reads are colored by the chromosome on which their mates can be found. Different colors for different reads from the same cluster indicate that mates are mapping a repetitive element. **b**, Analysis of the associated copy-number change on chromosome X identifies the missing L1 reciprocal cluster at the second breakpoint of the copy-number loss, and reveals a 3.9-kb deletion that occurs in conjunction with the integration of a 2.1-kb L1 somatic insertion. (A)_n and (T)_n represent poly(A) and poly(T) tails, respectively. **c**, Model of L1-mediated deletion. The integration of an L1 mRNA starts with L1-endonuclease cleavage promoting a 3' overhang for reverse transcription. The cDNA (–) strand invades a second 3' overhang from a pre-existing double-strand break upstream of the initial integration site. **d**, Distribution of the sizes of 90 L1-mediated deletions identified in the PCAWG dataset. **e**, In lung squamous carcinoma sample SA313800, a 34-bp truncated L1 insertion promotes a 1.1-kb deletion on chromosome 19. Because the L1 insertion was so short, we also identified discordant read pairs that span the L1 event and support the deletion. **f**, In esophageal adenocarcinoma sample SA528932, the integration on chromosome 3 of a 413-bp orphan L1 transduction from chromosome 7 causes a 2.5-kb deletion, which is supported by two clusters of discordant read pairs for which the mates map onto the transduced region of chromosome 7.

being generated by an L1 element (Fig. 4e). In this rearrangement, we found two different types of discordant read pairs at the deletion breakpoints: one cluster that supported the insertion of an L1

element and a second that spanned the L1 event and supported the deletion. Another type of L1-mediated deletion that could unequivocally be assigned to a single L1 insertion event is represented

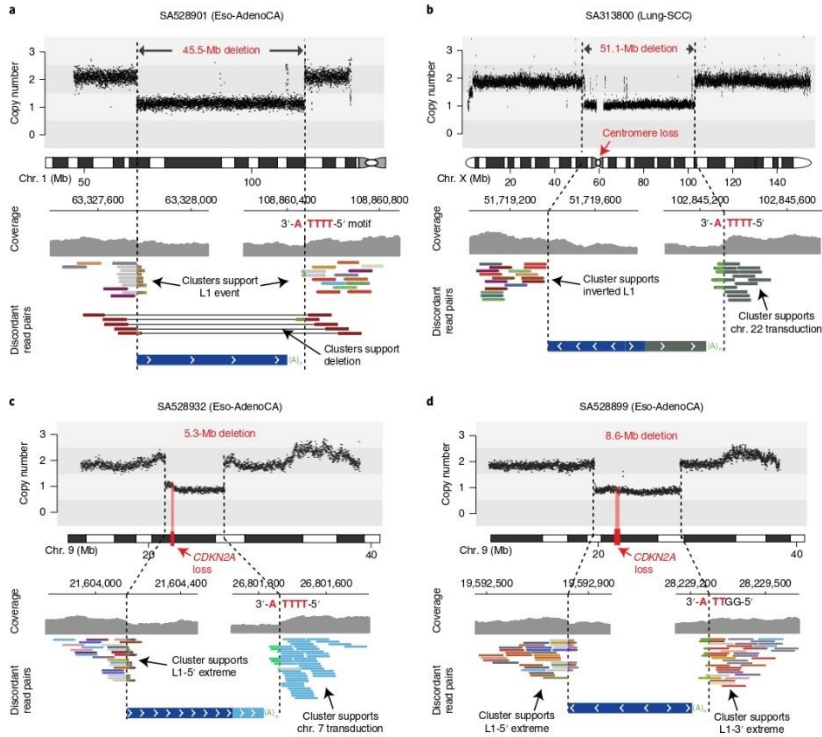


Fig. 5 | Somatic integration of L1 causes loss of megabase-size interstitial chromosomal regions in cancer. a In esophageal adenocarcinoma sample SA528901, a 45.5-Mb interstitial deletion on chromosome 1 is generated after integration of a short L1 event. We observed a pair of clusters of discordant read pairs for which the mates support both extremes of the L1 insertion. Because the L1 element size is smaller than the library insert size, we also identified read pairs that span the L1 event and support the deletion. The L1-endonuclease 5'-TTTT/A-3' motif identifies a target-primed reverse transcription (TPRT) L1-integration mechanism. **b** In esophageal tumor sample SA313800, a partnered transduction¹ (that is, the transduced region and its companion L1 source element) from chromosome 22 is integrated on chromosome X, promoting a 51.1-Mb deletion that removes the centromere. One negative cluster (green reads) supports a small region transduced from chromosome 22. Other negative clusters promote the loss of tumor-suppressor genes. In esophageal tumor sample SA528932, the somatic integration on chromosome 9 of a partnered transduction from chromosome 7, promotes a 5.3-Mb deletion that involves the loss of one copy of the tumor-suppressor gene *CDKN2A*. We observed a positive cluster of reads for which the mates map onto the 5' extreme of an L1, and a negative cluster that contains split reads that match a poly(A) region and for which the mates map onto a region that is transduced from chromosome 7 (light blue). **d** In a second esophageal adenocarcinoma sample, SA528899, the integration of an L1 retrotransposon generates an 8.6-Mb deletion that involves the same tumor-suppressor gene, *CDKN2A*. The sequencing data reveal two clusters—positive and negative—for which the mates support the L1 event.

by those deletions generated by the integration of orphan L1 transductions. These transductions represent fragments of unique DNA sequence located downstream of an active L1 locus, which are mobilized without the companion L1 (refs. 21^b). For example, in one esophageal tumor sample, SA528932, we found a deletion of 2.5 kb on chromosome 3 mediated by the orphan transduction of a sequence downstream of an L1 locus on chromosome 7 (Fig. 4f).

Owing to the unavailability of PCAWG DNA specimens, we performed a validation of 16 additional somatic L1-mediated deletions that were identified by TraFiC-mem in two head-and-neck cancer cell lines with high retrotransposition rates, NCI-H2009 and NCI-H2087. We carried out two independent validation approaches, including PCR followed by single-molecule sequencing of amplicons, and Illumina whole-genome sequencing

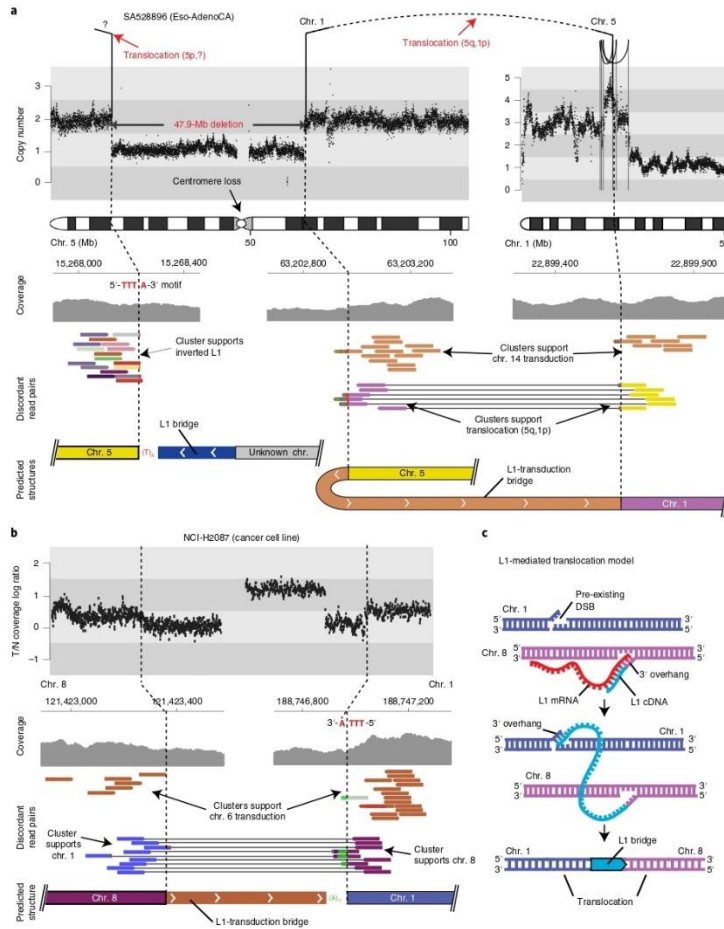


Fig. 6 | Somatic L1 integration promotes translocations in human cancers. a. In esophageal adenocarcinoma sample SA528896, two separate L1 events mediate interchromosomal rearrangements. In the first, an L1 transduction from a source element on chromosome 14q23.1 bridged an unbalanced translocation from chromosome 1p to 5q. A second somatic retrotransposition event bridged from chromosome 5p to an unknown part of the genome, completing a 47.9-Mb interstitial copy-number loss on chromosome 5 that removes the centromere. **b.** In a cancer cell line, NCI-H2087, we found an interchromosomal translocation, between chromosomes 8 and 1, mediated by a region transduced from chromosome 6, which acts as a bridge and joins both chromosomes. We observed two read clusters, positive and negative, that demarcate the boundaries of the rearrangement, for which the mates support the transduction event. In addition, two reciprocal clusters span the insertion breakpoints, supporting the translocation between chromosomes 8 and 1. **c.** A model for megabase-size L1-mediated interchromosomal rearrangements. L1-endonuclease cleavage promotes a 3' overhang in the negative strand, retrotranscription starts and the cDNA (–) strand invades a second 3' overhang from a pre-existing double-strand break on a different chromosome, leading to translocation.

using mate-pair libraries with long insert size (3 kb and 10 kb). The results confirmed the somatic status of the rearrangements and a single L1-derived retrotransposition as the cause of the associated copy-number loss (Supplementary Figs. 10–12 and Supplementary Table 7).

Analysis of L1 3'-extreme insertion breakpoint sequences from L1-mediated deletions found in the PCAWG dataset revealed that 82% (74 out of 90) of the L1 events that caused deletions preferentially inserted into sequences that resemble L1-endonuclease consensus cleavage sites (for example, 5'-TTT/A-3' and related sequences³³) (Supplementary Table 6). This confirms that the L1 machinery, through a target-primed reverse-transcription mechanism, is responsible for the integration of most of the L1 events that cause neighboring DNA loss³². Notably, in 16% (14 out of 90) of the events endonucleolytic cleavage occurred at the phosphodiester bond between a T and G instead of between the standard T and A site. In addition, we observed 8% (7 out of 90) instances in which the endonuclease motif was not found and the integrated element was truncated at both the 5' and 3' ends, suggesting that a small fraction of L1-associated deletions are the consequence of an L1-endonuclease-independent insertion mechanism^{36,37}. Whatever mechanism of L1 integration is effective in each case, taken together, these data indicate that the somatic integration of L1 elements induces the associated deletions.

Megabase-size L1-mediated deletions cause loss of tumor-suppressor genes. Most L1-mediated deletions ranged from a few hundred to thousands of base pairs, although occasionally megabase-long regions of a chromosome were deleted (Fig. 4c and Supplementary Table 6). For example, in esophageal tumor sample SA528901, we found a 45.5-Mb interstitial deletion that involved the p31.3–p13.3 regions of chromosome 1 (Fig. 5a), in which both breakpoints of the rearrangement showed the hallmarks of a deletion mediated by integration of an L1 element. Here, the L1 element is 5' truncated, which generated a small L1 insertion, allowing a fraction of the sequencing read pairs to span both breakpoints of the rearrangement. This unequivocally supports the model that the observed copy-number change is indeed a deletion mediated by retrotransposition of an L1 element. Similarly, in a lung tumor sample, SA313800, we found an interstitial L1-mediated deletion that induced the loss of 51.1 Mb from chromosome X, which included the centromere (Fig. 5b).

L1-mediated deletions were, on occasion, driver events and caused the loss of tumor-suppressor genes. In esophageal tumor sample SA528932, the integration of an L1 transduction from chromosome 7p12.3 to the short arm of chromosome 9 caused a 5.3-Mb clonal deletion that involved the 9p21.3–9p21.2 region. This led to the loss of one copy of a key tumor-suppressor gene, *CDKN2A* (Fig. 5c), which is deleted in many cancer types including esophageal tumors^{33–36}. Notably, the sequencing data revealed a somatic transduction that arose from this L1 element at its new insertion site, demonstrating that L1 events that promote deletions can be competent for retrotransposition (Supplementary Fig. 13). In a

second esophageal tumor sample, SA528899, an L1 element integrated into chromosome 9 promoted an 8.6-Mb clonal deletion that encompasses the 9p22.1–9p21.1 region that removes one copy of the same tumor-suppressor gene, *CDKN2A* (Fig. 5d). Thus, L1-mediated deletions have clear oncogenic potential.

L1 retrotransposition generates other types of structural variation in human tumors. Somatic retrotransposition can also be involved in mediating or repairing more complex structural variants. In one esophageal tumor sample, SA528896, two separate L1-mediated structural variants were present within a complex cluster of rearrangements (Fig. 6a). In the first, an L1 transduction from a source element on chromosome 14q23.1 bridged an unbalanced translocation from chromosome 1p to 5q. A second somatic retrotransposition event bridged from chromosome 5p to an unknown part of the genome, completing a large interstitial copy-number loss on chromosome 5 that involves the centromere. This case suggests that retrotransposon transcripts and their reverse-transcriptase machinery can mediate breakage and repair of complex dsDNA breaks, spanning two chromosomes.

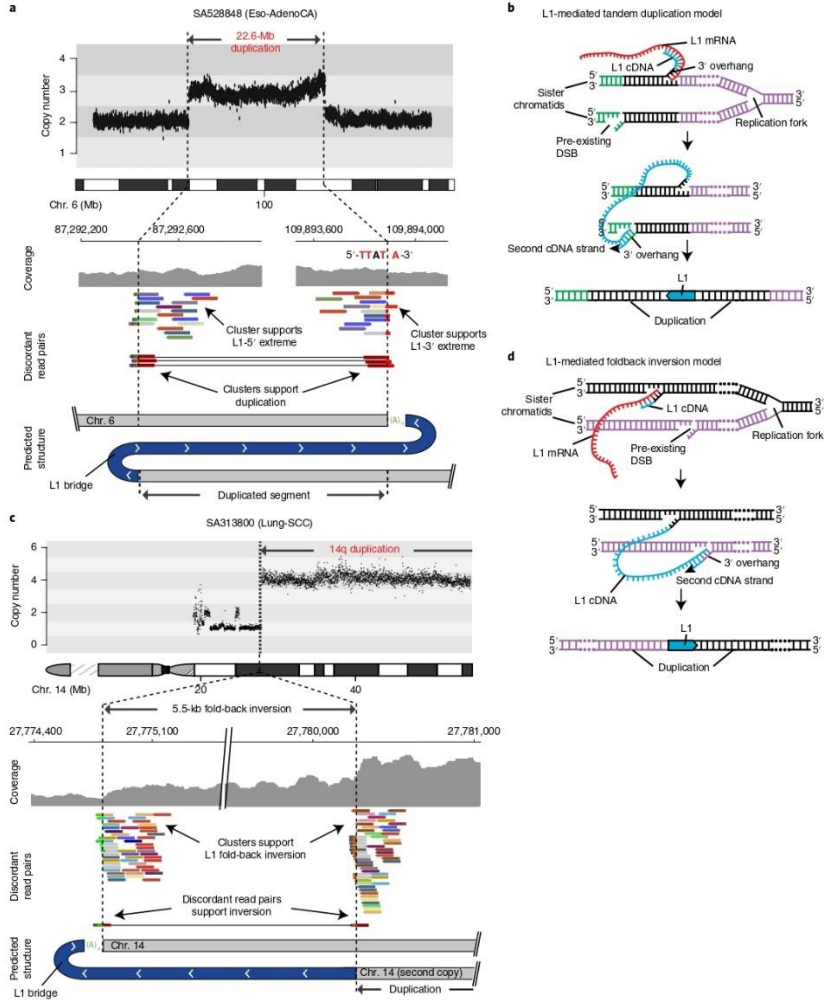
To explore this further, we identified single-L1 clusters with no reciprocal cluster in the cancer cell lines that were sequenced by using mate pairs with 3 kb and 10 kb inserts. Such events may correspond to hidden genomic translocations leading to the linkage of two different chromosomes, in which L1 retrotransposition is involved. One of the samples, NCI-H2087, showed translocation breakpoints at 1q31.1 and 8q24.12, both of which had the hallmarks of L1-mediated deletions, for which the mate-pair sequencing data identified an orphan L1 transduction from chromosome 6p24 that bridged both chromosomes (Fig. 6b). The configuration has also been confirmed by using long-read single-molecule sequencing (Supplementary Fig. 11). This interchromosomal rearrangement is likely mediated by the aberrant operation of L1-integration mechanism, in which the L1-transduced cDNA is wrongly paired with a second 3' overhang from a pre-existing double-strand break generated in a second chromosome³² (Fig. 6c).

We also found evidence that L1 integrations can cause duplications of large genomic regions in human cancer. In esophageal tumor sample SA528848 (Fig. 7a), we identified two independent read clusters that support the integration of a small L1 event, coupled with a coverage drop at both breakpoints. Copy-number analysis revealed that the two L1 clusters demarcate the boundaries of a 22.6-Mb duplication that involves the 6q14.3–q21 region, suggesting that the L1 insertion could be the cause of such rearrangement by bridging sister chromatids during or after DNA replication (Fig. 7b). The analysis of the rearrangement data at the breakpoints identified read pairs that traverse the length of the L1 insertion breakpoint, and the L1-endonuclease motif is the L1 3' insertion breakpoint, both confirming a single L1 event as the cause of a tandem duplication (Fig. 7a). Notably, this duplication increases the copy number of the cyclin C gene, *CCNC*, which is dysregulated in some tumors³⁷.

Fig. 7 | Somatic L1 integration promotes duplications of megabase-scale regions in human cancers. **a**, In esophageal adenocarcinoma sample SA528848, we found a 22.6-Mb tandem duplication on the long arm of chromosome 6. The analysis of the sequencing data at the boundaries of the rearrangement breakpoints reveals two clusters of discordant read pairs for which the mates support the involvement of an L1 event. Because the L1 element was shorter than the library size, we also found two reciprocal clusters that aligned 22.6 Mb apart on the genome and in opposite orientation, spanning the insertion breakpoints and confirming the tandem duplication. An L1-endonuclease 5'-TTT/A-3' degenerate motif was found. **b**, Large direct tandem duplications can be generated if the cDNA (–) strand invades a second 3' overhang from a pre-existing double-strand break that occurred on a sister chromatid, and downstream to the initial integration site locus. **c**, In lung tumor sample SA313800, a small L1 insertion causes a 79.6-Mb duplication of the 14q arm through the induction of a fold-back inversion rearrangement. The analysis of the sequencing data at the breakpoint revealed two clusters of discordant read pairs (multi-colored reads) with the same orientation, aligning close together (5.5 kb apart) and demarcating a copy-number change for which the sequencing density is much greater on the right half of the rearrangement than the left. Both clusters of multi-colored reads support the integration of an L1. **d**, L1-mediated fold-back inversion model.

L1-mediated rearrangements can induce breakage–fusion–bridge cycles that trigger oncogene amplification. L1 retrotranspositions can also induce genomic instability by triggering breakage–fusion–bridge cycles. This form of genetic instability starts with end-to-end

fusion of broken sister chromatids, and lead to a dicentric chromosome that forms an anaphase bridge during mitosis. Classically, the end-to-end chromosome fusions are thought to arise from telomere attrition^{38–40}. We found, however, that somatic retrotransposition



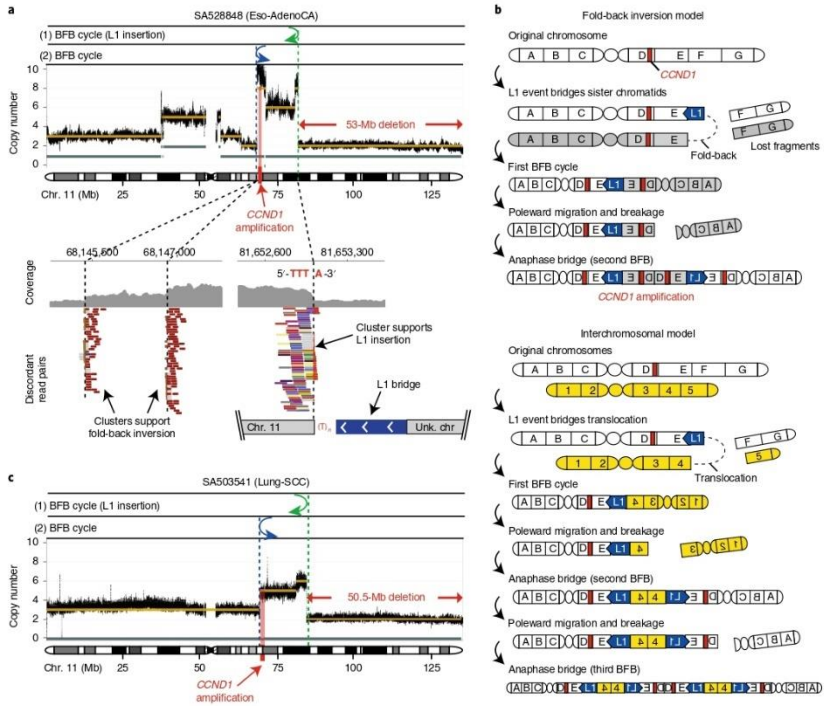


Fig. 8 | Somatic integration of L1 can trigger breakage-fusion-bridge cycles that lead to oncogene amplification. a, In esophageal adenocarcinoma sample SA528848, a single cluster of discordant reads (multi-colored reads) together with an L1-endonuclease cleavage site motif 5'-TTT'A-3' supports the integration of an L1 event that demarcates a 53-Mb telomeric (that is, including the telomere) deletion, from a region of massive amplification that involves *CCND1*. Around 14 Mb upstream of the breakpoint of the deletion, we observed the presence of two clusters of read pairs (brown reads) that align close together and in the same orientation, which demarcate a change in copy number; this is a distinctive pattern of a fold-back inversion^{62,63}, a rearrangement typically found to be associated with breakage-fusion-bridge (BFB) repair. In this fold-back inversion, the coverage shows much greater density on the right half of the rearrangement than the left, indicating that the abnormal chromosome is folded back on itself leading to duplicated genomic sequences in a head-to-head (inverted) orientation. The patterns described here suggest two independent breakage-fusion-bridge cycles, marked with (1) and (2). The copy-number plot shows the consensus total copy numbers (gold band) and the minor allele copy numbers (gray band). **b**, Models for the patterns described in **a**. The fold-back inversion model involves two breakage-fusion-bridge cycles, one induced by L1-mediated fold-back inversion (see Fig. 7d), and a second induced by standard breakage-fusion-bridge repair. The interchromosomal rearrangement model involves an interchromosomal rearrangement mediated by an L1, followed by one extra cycle of breakage-fusion-bridge repair. **c**, In lung cancer sample SA503541, the integration of an L1 retrotransposon is associated with a 50-Mb loss on 11q that includes the telomere, and activates breakage-fusion-bridge repair, which leads to the amplification of *CCND1*.

can induce the first inverted rearrangement that generates end-to-end fusion of sister chromatids. In lung tumor sample SA313800 (Fig. 7c), we found a small L1 event inserted on chromosome 14q that demarcates a copy-number change that involves a 79.6-Mb amplification of the 14q arm. Analysis of the sequencing data at the breakpoint revealed two discordant read clusters with the same orientation, which are 5.5 kb apart and support the integration of an L1. Both discordant clusters demarcate an increment

of the sequencing coverage, for which the density is much greater in the right cluster. The only genomic structure that can explain this pattern is a fold-back inversion in which the two sister chromatids are bridged by an L1 retrotransposon in head-to-head (inverted) orientation (Fig. 7d).

In the example described above (Fig. 7c,d), no further breaks occurred, and the L1 retrotransposon generated an isochromosome (14q). In addition, we found examples in which the fusion of

two chromatids by an L1 bridge induced further cycles of breakage–fusion–bridge repair. In esophageal tumor sample SA528848, we identified a cluster of reads on the long arm of chromosome 11 that had the typical hallmarks of an L1-mediated rearrangement (Fig. 8a). Copy-number data analysis showed that this L1 insertion point demarcated a 53-Mb deletion, which involved the loss of the telomeric region, from a region of massive amplification on chromosome 11. The amplified region on chromosome 11 contains the *CCND1* oncogene, which is amplified in many human cancers⁴¹. The other end of this amplification was bound by a conventional fold-back inversion rearrangement (Fig. 8a), which is indicative of breakage–fusion–bridge repair^{24,42}.

These patterns suggest the following sequence of events. During or soon after S phase, a somatic L1 retrotransposition bridges across sister chromatids in inverted orientation, breaking off the telomeric ends of 11q, which are then lost to the clone during the subsequent cell division (fold-back inversion model, Fig. 8b). The chromatids bridged by the L1 insertion now produce a dicentric chromosome. During mitosis, the two centromeres are pulled to opposite poles of the dividing cell, creating an anaphase bridge, which is resolved by further dsDNA breakage. This induces a second cycle of breakage–fusion–bridge repair, albeit not one mediated by L1 retrotransposition. These cycles lead to rapid-fire amplification of the *CCND1* oncogene. Alternatively, an interchromosomal rearrangement mediated by L1 retrotransposition (interchromosomal rearrangement model, Fig. 8b) followed by two cycles of breakage–fusion–bridge repair could generate similar copy-number patterns with telomere loss and amplification of *CCND1*.

Our data show that L1-mediated retrotransposition is an alternative mechanism of creating the first dicentric chromosome that induces subsequent rounds of chromosomal breakage and repair. If this occurs near an oncogene, the resulting amplification can provide a powerful selective advantage to the clone. We searched the PCAWG dataset for other rearrangements that included copy-number amplifications from telomeric deletions that were mediated by L1 integration. We found four more such events across three cancer samples (Supplementary Fig. 14). In a lung tumor sample, SA503541, we found almost identical rearrangements to the one described above (Fig. 8c). In this case, a somatic L1 event also generated telomere loss that induced a second cycle of breakage–fusion–bridge repair. The megabase-size amplification of chromosomal regions also targeted the *CCND1* oncogene, in which the boundaries were demarcated by the L1 insertion breakpoint and a fold-back inversion, which indicates breakage–fusion–bridge repair. The independent occurrence of these patterns, which involve the amplification of *CCND1*, in two different tumor samples (SA528848 and SA503541) demonstrates a mutational mechanism mediated by L1 retrotransposition, which likely contributes to the development of human cancer.

Discussion

Here we characterize the patterns and mechanisms of cancer retrotransposition on a multidimensional scale, across 2,954 cancer genomes, integrated with rearrangement, transcriptomic and copy-number data. Our analyses provide a new perspective on the long-standing question of whether the activation of retrotransposons is relevant in human oncogenesis. Our findings demonstrate that major restructuring of cancer genomes can sometimes emerge from aberrant L1 retrotransposition events in tumors with high retrotransposition rates, particularly in esophageal, lung and head-and-neck cancers. L1-mediated deletions can promote the loss of megabase-scale regions of a chromosome that may involve centromeres and telomeres. It is likely that the majority of such genomic rearrangements would be harmful for a cancer clone. However, occasionally, L1-mediated deletions may promote cancer-driving rearrangements that involve the loss of tumor-suppressor genes

and/or the amplification of oncogenes, representing another mechanism by which cancer clones acquire new mutations that help them to survive and grow. We expect that structural variants induced by somatic retrotransposition in human cancer are more frequent than we could unambiguously characterize here, given the constraints on the fragment sizes of paired-end sequencing libraries. Long-read sequencing technologies should be able to provide a more comprehensive picture of how frequent such events are. Relatively few germline L1 loci in a given tumor, typically one to three copies, are responsible for such marked structural remodeling. Given the role that these L1 copies may have in some cancer types, this work underscores the importance of characterizing cancer genomes for patterns of L1 retrotransposition.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-019-0562-0>.

Received: 21 September 2017; Accepted: 26 November 2019;

Published online: 5 February 2020

References

- International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Kazazian, H. H. Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
- Sassaman, D. M. et al. Many human L1 elements are capable of retrotransposition. *Nat. Genet.* **16**, 37–43 (1997).
- Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285 (2003).
- Beck, C. R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
- Menendez, L., Benigno, B. B. & McDonald, J. F. L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. *Mol. Cancer* **3**, 12 (2004).
- Tubio, J. M. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251–1254 (2014).
- Hölmès, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. Jr. A new retrotransposable human L1 element from the L1RE2 locus on chromosome 1q produces a chimeric insertion. *Nat. Genet.* **7**, 143–148 (1994).
- Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
- Kazazian, H. H. Jr. Processed pseudogene insertions in somatic cells. *Mob. DNA* **5**, 20 (2014).
- Cooke, S. L. et al. Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5**, 3644 (2014).
- Ewing, A. D. et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14**, R22 (2013).
- Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
- Helman, E. et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
- Solyom, S. et al. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2528–2538 (2012).
- Symer, D. E. et al. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327–338 (2002).
- Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315–325 (2002).
- Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.* **19**, 1583–1591 (2016).
- Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
- Kazazian, H. H. Jr. & Moran, J. V. Mobile DNA in health and disease. *N. Engl. J. Med.* **377**, 361–370 (2017).
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* <https://doi.org/10.1038/s41586-020-1969-6> (2020).
- Wylie, A. et al. p53 genes function to restrain mobile elements. *Genes Dev.* **30**, 64–77 (2016).

23. Jung, H., Choi, J. K. & Lee, E. A. Immune signatures correlate with L1 retrotransposition in gastrointestinal cancers. *Genome Res.* **28**, 1136–1146 (2018).
24. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
25. Mita, P. et al. LINE-1 protein localization and functional dynamics during the cell cycle. *eLife* **7**, e30058 (2018).
26. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
27. Kimberland, M. L. et al. Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum. Mol. Genet.* **8**, 1557–1560 (1999).
28. Chang, H. H. Y., Panunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **18**, 495–506 (2017).
29. Han, K. et al. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res.* **33**, 4040–4052 (2005).
30. Sen, S. K., Huang, C. T., Han, K. & Batzer, M. A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res.* **35**, 3741–3751 (2007).
31. Farkash, E. A., Kao, G. D., Horman, S. R. & Prak, E. T. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res.* **34**, 1196–1204 (2006).
32. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol. Cell Biol.* **25**, 7780–7795 (2005).
33. Zhou, C., Li, J. & Li, Q. CDKN2A methylation in esophageal cancer: a meta-analysis. *Oncotarget* **8**, 50071–50083 (2017).
34. The Cancer Genome Atlas Research Network Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
35. The Cancer Genome Atlas Network Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
36. The Cancer Genome Atlas Research Network Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
37. Xu, W. & Ji, J. Y. Dysregulation of CDK8 and cyclin C in tumorigenesis. *J. Genet. Genomics* **38**, 439–452 (2011).
38. Artandi, S. E. et al. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406**, 641–645 (2000).
39. O'Hagan, R. C. et al. Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell* **2**, 149–155 (2002).
40. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **18**, 175–186 (2017).
41. Beroukhi, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
42. Campbell, P. J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
43. Li, Y. et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98–102 (2014).

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.
© The Author(s) 2020

¹Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. ²Biomedical Research Centre (CINBIO), University of Vigo, Vigo, Spain. ³Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. ⁴Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. ⁵The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. ⁶Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. ⁷Genome Data Science, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ⁸Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ⁹The Francis Crick Institute, London, UK. ¹⁰Department of Human Genetics, University of Leuven, Leuven, Belgium. ¹¹Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. ¹²Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. ¹³Cancer Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Cambridge, UK. ¹⁴Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain. ¹⁵Galicja Sur Health Research Institute, Vigo, Spain. ¹⁶Faculty of Science and Technology, University of Vic—Central University of Catalonia (UVic-UCC), Vic, Spain. ¹⁷Barcelona Supercomputing Center (BSC), Barcelona, Spain. ¹⁸Experimental Cancer Genetics, Wellcome Sanger Institute, Cambridge, UK. ¹⁹Oxford Big Data Institute, University of Oxford, Oxford, UK. ²⁰DNA Repair and Genome Integrity, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. ²¹Department of Biochemistry and Molecular Biology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. ²²Medical Research Council (MRC) Cancer Unit, University of Cambridge, Cambridge, UK. ²³Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Baltimore, MD, USA. ²⁴University of Cambridge, Cambridge, UK. ²⁵Li Ka Shing Centre, Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. ²⁶European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. ²⁷Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. ²⁸Department of Urology, Charité Universitätsmedizin Berlin, Berlin, Germany. ²⁹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ³⁰Department of Genetics and Informatics Institute, University of Alabama at Birmingham (UAB) School of Medicine, Birmingham, AL, USA. ³¹University of Texas MD Anderson Cancer Center, Houston, TX, USA. ³²Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ³³The Broad Institute of Harvard and MIT, Cambridge, MA, USA. ³⁴Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. ³⁵Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. ³⁶Harvard Medical School, Boston, MA, USA. ³⁷Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. ³⁸Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. ³⁹Department of Computer Science, Yale University, New Haven, CT, USA. ⁴⁰Department of Human Genetics, McGill University, Montreal, Québec, Canada. ⁴¹Department of Oncology and Onco-Hematology, University of Milan, Milan, Italy. ⁴²Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. ⁴³Canadian Center for Computational Genomics, McGill University, Montreal, Québec, Canada. ⁴⁴Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ⁴⁵Ludwig Center at Harvard, Boston, MA, USA. ⁴⁶Oxford NIHR Biomedical Research Centre, Oxford, UK. ⁴⁷European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. ⁴⁸Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Baltimore, MD, USA. ⁴⁹McKusick-Nathans Institute of Genetic Medicine, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁵⁰A full list of members and their affiliations appears at the end of the paper. ⁵¹Department of Haematology, University of Cambridge, Cambridge, UK. ⁵²A full list of members and their affiliations appears online. ⁵³These authors contributed equally: Eva G. Alvarez, Adrian Baez-Ortega, Jorge Zamora, Fran Supek. ⁵⁴These authors jointly supervised this work: Peter J. Campbell, Jose M. C. Tubio. ⁵⁵e-mail: pc8@sanger.ac.uk; jose.mc.tubio@usc.es

Detection of early seeding of Richter transformation in chronic lymphocytic leukemia



OPEN

Detection of early seeding of Richter transformation in chronic lymphocytic leukemia

Ferran Nadeu^{1,2,21}✉, Romina Royo^{3,21}, Ramon Massoni-Badosa^{4,21}, Heribert Playa-Albinyana^{5,12,21}, Beatriz Garcia-Torre^{1,21}, Martí Duran-Ferrer^{6,1,2}, Kevin J. Dawson⁷, Marta Kulis¹, Ander Diaz-Navarro^{2,6}, Neus Villamor^{1,2,7}, Juan L. Melero⁸, Vicente Chapaprieta⁹, Ana Dueso-Barroso³, Julio Delgado^{1,2,7,9}, Riccardo Moia¹⁰, Sara Ruiz-Gil⁴, Domenica Marchese⁴, Ariadna Giró^{1,2}, Núria Verdaguer-Dot¹, Mónica Romo¹, Guillem Clot^{1,2}, Maria Rozman^{1,7}, Gerard Frigola⁹, Alfredo Rivas-Delgado^{1,7}, Tycho Baumann^{2,7,20}, Miguel Alcoceba^{2,11}, Marcos González^{2,11}, Fina Climent¹², Pau Abrisqueta¹³, Josep Castellvi¹³, Francesc Bosch¹³, Marta Aymerich^{1,2,7}, Anna Enjuanes¹, Sílvia Ruiz-Gaspà¹, Armando López-Guillermo^{1,2,7,9}, Pedro Jares^{1,2,7,9}, Sílvia Beà^{1,2,7,9}, Salvador Capella-Gutierrez³, Josep Ll. Gelpi⁹, Núria López-Bigas^{14,15,16}, David Torrents^{3,16}, Peter J. Campbell⁵, Ivo Gut^{4,15}, Davide Rossi¹⁷, Gianluca Gaidano¹⁰, Xose S. Puente^{2,6}, Pablo M. Garcia-Roves^{9,18}, Dolores Colomer^{1,2,7,9}, Holger Heyn^{4,15}, Francesco Maura¹⁹, José I. Martín-Subero^{1,2,9,16} and Elías Campo^{1,2,7,9}✉

Richter transformation (RT) is a paradigmatic evolution of chronic lymphocytic leukemia (CLL) into a very aggressive large B cell lymphoma conferring a dismal prognosis. The mechanisms driving RT remain largely unknown. We characterized the whole genome, epigenome and transcriptome, combined with single-cell DNA/RNA-sequencing analyses and functional experiments, of 19 cases of CLL developing RT. Studying 54 longitudinal samples covering up to 19 years of disease course, we uncovered minute subclones carrying genomic, immunogenetic and transcriptomic features of RT cells already at CLL diagnosis, which were dormant for up to 19 years before transformation. We also identified new driver alterations, discovered a new mutational signature (SBS-RT), recognized an oxidative phosphorylation (OXPHOS)^{hi}-B cell receptor (BCR)^{low}-signaling transcriptional axis in RT and showed that OXPPOS inhibition reduces the proliferation of RT cells. These findings demonstrate the early seeding of subclones driving advanced stages of cancer evolution and uncover potential therapeutic targets for RT.

Clonal evolution¹ drives cancer initiation, progression and relapse due to the stepwise acquisition and/or selection of fitter subclones^{2,3}. The understanding of tumor evolution is hampered by the analysis of bulk tumor cell populations at low resolution and at single or limited time points of the disease course in most studies⁴. A better knowledge of this process might translate into anticipation-based treatment strategies⁵. RT in CLL represents a paradigmatic model of cancer evolution occurring rarely in treatment-naïve patients with CLL but found in 4–20% of patients after chemoimmunotherapy (CIT) and targeted therapies⁶. RT sometimes occurs within the first months after treatment

initiation^{7–9}, suggesting selection of pre-existing subclones¹⁰. Nonetheless, the genomic/epigenomic mechanisms driving RT after CIT^{11–17} or targeted agents^{18–21} are not well known. The aims of the present study were to reconstruct the evolutionary history of RT and to reveal the molecular processes underlying this transformation.

Results

Genomic characterization of RT. We sequenced 53 whole genomes and 1 whole exome of synchronous or longitudinal samples of 19 patients (up to six time points per patient) in whom CLL transformed into diffuse large B cell lymphoma (RT-DLBCL; $n = 17$),

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. ²Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. ³Barcelona Supercomputing Center (BSC), Barcelona, Spain. ⁴CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ⁵Wellcome Sanger Institute, Hinxton, UK. ⁶Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. ⁷Hospital Clínic de Barcelona, Barcelona, Spain. ⁸OmniScope, Barcelona, Spain. ⁹Universitat de Barcelona, Barcelona, Spain. ¹⁰Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy. ¹¹Biología Molecular e Histocompatibilidad, IBSAL-Hospital Universitario, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain. ¹²Hospital Universitari de Bellvitge-Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. ¹³Department of Hematology, Vall d'Hebron Institute of Oncology, Vall d'Hebron University Hospital, Barcelona, Spain. ¹⁴Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. ¹⁵Universitat Pompeu Fabra (UPF), Barcelona, Spain. ¹⁶Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ¹⁷Oncology Institute of Southern Switzerland, Bellinzona, Switzerland. ¹⁸Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. ¹⁹Myeloma Service, Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA. ²⁰Present address: Hospital Universitario 12 de Octubre, Madrid, Spain. ²¹These authors contributed equally: Ferran Nadeu, Romina Royo, Ramon Massoni-Badosa, Heribert Playa-Albinyana, Beatriz Garcia-Torre. ✉e-mail: nadeu@recercaclinic.cat; ecampo@clinic.cat

plasmablastic lymphoma (RT-PBL; $n=1$) or prolymphocytic leukemia (RT-PLL; $n=1$). Nontumor samples were available in 12 patients. RT occurred simultaneously with CLL at diagnosis ($n=3$) or after up to 19 years following different lines of treatment with CIT ($n=6$) and targeted therapies ($n=10$; BCR inhibitors, ibrutinib $n=6$; duvelisib $n=2$; idelalisib $n=1$; and BCL2 inhibitor, venetoclax $n=1$). All instances of RT were clonally related to CLL, 15 tumors had unmutated IGHV (U-CLL) and 4 had mutated IGHV (M-CLL). Whole-genome sequencing (WGS) data were integrated with bulk epigenetic and transcriptomic analyses as well as single-cell DNA and RNA sequencing (Fig. 1a, Extended Data Fig. 1 and Supplementary Tables 1 and 2).

The WGS and epigenome of CLL and RT revealed a concordant increased complexity from CLL diagnosis to relapse and RT (Fig. 1b, Extended Data Fig. 2a and Supplementary Tables 3–8). The RT genomes carried a median of 1.8 mutations per megabase, 18 copy number alterations (CNAs) and 37 structural variants (SVs) that contrasted with 1.1 mutations per megabase, 4 CNAs and 5 SVs observed at CLL diagnosis. No major differences were seen among RT occurring after different therapies (Fig. 1b and Extended Data Fig. 2b). We discovered new driver genes and mechanisms in RT, expanding previous observations^{2,18,21–24} (Fig. 1c, Extended Data Fig. 2c–e, Supplementary Fig. 1 and Supplementary Tables 9 and 10). The main alterations involved cell-cycle regulators (17 of 19, 89%), chromatin modifiers (79%), MYC (74%), nuclear factor (NF)- κ B (74%) and NOTCH (32%) pathways. These aberrations were simultaneously present in most cases but alterations in MYC and NOTCH pathways only co-occurred in 2 of 19 cases (Fig. 1c). Aberrations in genes such as *TP53*, *NOTCH1*, *BIRC3*, *EGR2* and *NFKBIE* were usually present and clonally dominant after the first CLL sample, whereas others were only detected at RT or during the disease course (for example *CDKN2A/B*, *CDKN1A/B*, *ARID1A*, *CREBBP*, *TRAF3* and *TNFAIP3*) (Fig. 1c). New alterations included deletions of *CDKN1A* and *CDKN1B* in five cases of RT associated with downregulation of their expression, one immunoglobulin (IG)-*CDK6* translocation and one *CCND2* mutation already present at CLL diagnosis, and *CCND3*-IG and *MYCN*-IG translocations acquired at RT in two different cases (Fig. 1d,e, Extended Data Fig. 3a,b and Supplementary Table 11). Most chromatin remodelers were affected by deletions with reduced gene expression. New alterations in this group were deletions of *ARID4B* and truncations of *CREBBP*²⁵ and *SMARCA4* (ref.¹⁰) by translocations and chromoplexy (Fig. 1f and Extended Data Fig. 3c–e). We also identified recurrent *IRF4* alterations in RT, which have been linked to increased MYC levels in CLL²⁶. *BTK/PLCG2* or *BCL2* mutations were not detected in any RT after treatment with BCR or BCL2 inhibitors, respectively. Notably, the two cases of M-CLL developing RT after targeted therapies carried the IGLV3–21^{B10} mutation, which triggers cell-autonomous BCR signaling²⁷ (Fig. 1c).

In addition to the high frequency of CNAs previously identified in RT^{18,24}, we observed a high number of complex structural alterations (Fig. 1c). Chromothripsis was found in eight RT tumors targeting *CDKN2A/B* and the new *CDKN1B* in five and one cases, respectively, and *MYC*, *MGA*, *SPEN*, *TNFAIP3* and chromatin remodeling genes in additional cases (Fig. 1g and Extended Data Fig. 3f–j).

Altogether, our analyses expand the catalog of driver genes, pathways and mechanisms involved in RT and recognize a similar distribution of these alterations in RT after different therapies, suggesting that treatment-specific pressure is not a major determinant of the driver genomic landscape of these tumors.

New mutational processes in RT. To understand the increased mutational burden of RT, we explored the mutational processes re-shaping the genome of CLL and RT. An unsupervised analysis showed that the mutational profile of RT was notably different

from M-CLL and U-CLL before therapy (ICGC-CLL cohort, $n=147$)²⁸ or at post-treatment relapse (independent cohort of 27 CLL post-treatment samples) (Fig. 2a). We identified 11 mutational signatures distributed genome-wide and 2 in clustered mutations (Extended Data Fig. 4 and Supplementary Tables 12–14). Among the former, we extracted a new signature characterized by (T>A)A and, to a lesser extent, (T>C/G)A mutations not recognized previously in any cancer type, including CLL and DLBCL^{28–33}. We named this single-base substitution signature, SBS-RT (Fig. 2b). SBS-RT was present in the RT sample of 7 of 18 patients, 1 of 6 after CIT and 6 of 10 after multiple therapies, including targeted agents and detected in all subtypes of transformation (RT-DLBCL, RT-PBL and RT-PLL) (Fig. 2c and Supplementary Table 15). It was also present in CLL samples before RT in patients 12 and 3,299 but was not identified in the reanalysis of our ICGC-CLL or post-treatment CLL cohorts. None of the patients in these two additional cohorts had evidence of RT (median follow-up 9.8 years, range 0.2–30.4) (Fig. 2c, Extended Data Fig. 5a and Supplementary Table 15). Further characterization of this new signature showed (1) a modest correlation between SBS-RT and total number of mutations ($R=0.79$, $P=0.11$); (2) SBS-RT mutations present in all different chromatin states and early/late replicating regions although with a moderate enrichment in heterochromatin/late replication; and (3) lack of replication and transcriptional strand bias (Extended Data Fig. 5b–f and Supplementary Table 16).

Among the remaining ten genome-wide signatures, five were previously identified in CLL and DLBCL (SBS1 and SBS5 (clock-like), SBS8 (unknown etiology), SBS9 (attributed to polymerase ϵ) and SBS18 (possibly damage by reactive oxygen species)); three had been only found in DLBCL (SBS2 and SBS13 (APOBEC enzymes) and SBS17b (unknown)); and two have been recently described related to treatments with melphalan³⁴ or ganciclovir³⁵, which were named here as SBS-melphalan and SBS-ganciclovir, respectively (Fig. 2b,c and Extended Data Fig. 4). SBS-melphalan was found in three RT cases, two had received melphalan as a conditioning of their allogeneic stem-cell transplant 1.9 and 4.2 years before RT, respectively. SBS-ganciclovir was found in the RT sample of one patient that had received valganciclovir (prodrug of ganciclovir) due to cytomegalovirus reactivation (Fig. 2c,d and Extended Data Fig. 1a). Notably, all cases with the new SBS-RT at time of RT had been treated with the alkylating agents bendamustine ($n=5$) or chlorambucil ($n=2$) during their CLL history at a median of 2.9 years (range 0.7 to 6.8) before RT. Contrarily, RT cases lacking the SBS-RT had never received these drugs (Fig. 2c,d and Extended Data Fig. 1a).

To time the activity of each mutational process, we reconstructed the phylogenetic tree for the 11 patients with multiple synchronous ($n=2$) or longitudinal ($n=9$) samples and germline available and measured the contribution of each signature to the mutational profile of each subclone. The major subclone at time of transformation was named ‘RT subclone’ (Supplementary Table 17). As expected, clock-like mutational signatures were present all along the phylogeny (constantly acquired), whereas SBS9 was found only in the trunk of the two M-CLL tumors (patients 365 and 19; early events). DLBCL-related signatures, SBS-ganciclovir, SBS-melphalan and SBS-RT were found in single RT subclones in six cases while two cases carried two simultaneous subclones with SBS-RT (patients 12 and 19) (Fig. 2e). SBS-RT represented 28.6% of the mutations acquired in RT (mean 679, range 499–1,167) and it was occasionally associated with coding mutations in driver genes (*EP300* and *CHTA*) (Fig. 2f, Extended Data Fig. 5g and Supplementary Table 16). By applying a high-coverage, unique molecular identifier (UMI)-based next-generation sequencing (NGS) approach in longitudinal samples of patients 12, 19 and 63 (Supplementary Table 18), we observed that mutations of the RT subclones found in the main peaks of the SBS-RT were mainly identified in samples collected after bendamustine or chlorambucil therapy, whereas

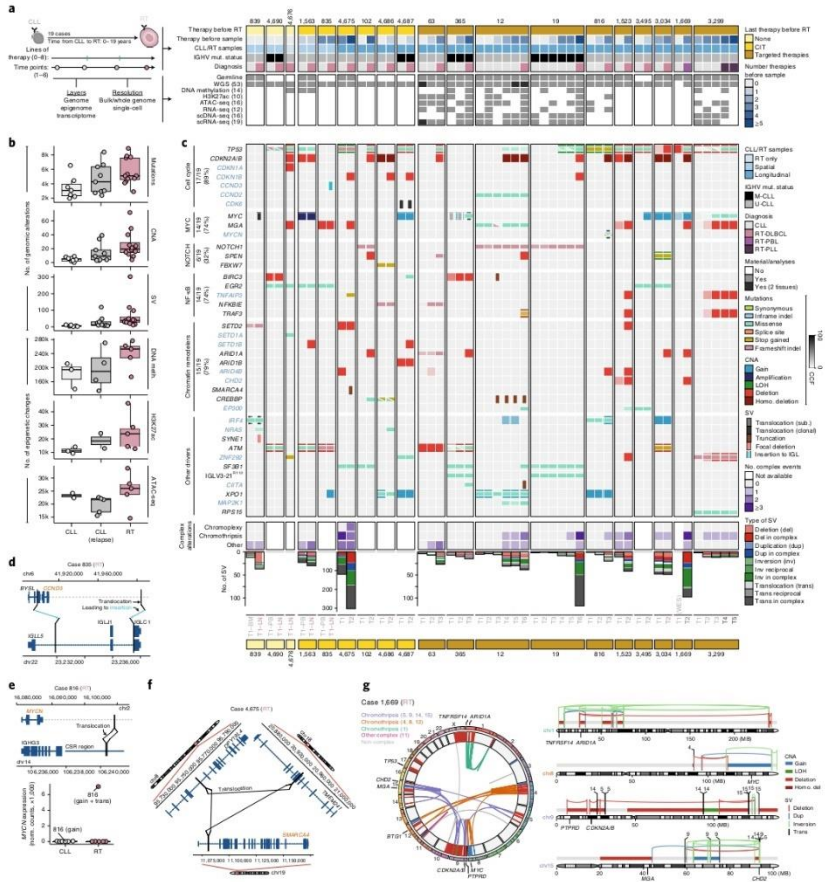


Fig. 1 | The genomic landscape of RT. a, Summary of the study. mut., mutation. **b**, Increase in genomic alterations and epigenetic changes compared to healthy naive and memory B cells over the disease course. Center line indicates median; box limits indicate upper and lower quartiles; whiskers indicate 1.5x interquartile range; and points indicate individual samples. **c**, Driver alterations of CLL and RT. New drivers in RT are labeled in blue. Each column represents a sample and genes are represented in rows. The transparency of the color of mutations and CNAs indicates the cancer cell fraction (CCF). The number of tumors harboring an alteration at the time of transformation is indicated for each biological group of drivers (left). Complex structural alterations are shown below, together with the total number of SVs. LOH, loss of heterozygosity. **d**, Schema of the *CCND3* insertion next to the constant region *IGLC1* in the RT sample of patient 835. **e**, Reciprocal translocation between *MYCN* and class-switch recombination (CSR) region of *IGHG3* in the RT sample of patient 816 (top). *MYCN* expression based on bulk RNA-seq (bottom). **f**, Chromoplexy disrupting *SMARCA4* in the RT sample of patient 4,675. **g**, The circos plot (left) displays the SVs (links) and CNAs (inner circle) found in the RT sample of patient 1,669. CNAs are colored by type and SVs are colored according to their occurrence within specific complex events. Target driver genes are annotated. Chromosome-specific plots (right) illustrate selected complex rearrangements affecting one or multiple driver genes with CNAs and SVs colored by type.

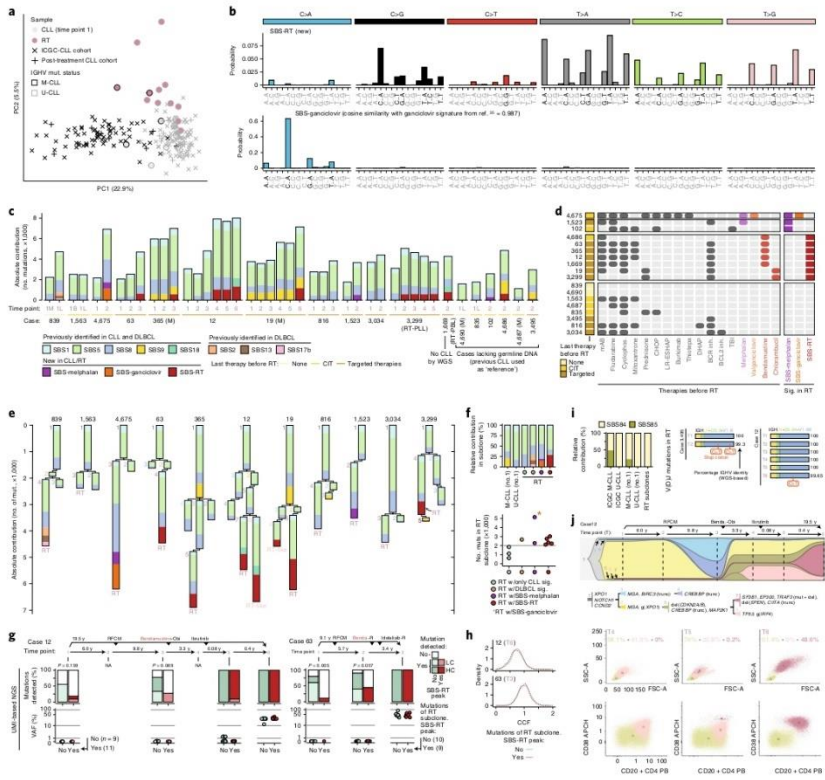


Fig. 2 | Mutational processes in RT. **a**, Principal component analysis (PCA) of the 96-mutational profile of CLL and RT. **b**, Signatures identified de novo in CLL/RT not reported in COSMIC. The main peaks of each signature are labeled in black. **c**, Contribution of mutational processes in CLL/RT. RT time points are marked in a rose color. B, peripheral blood; L, lymph node; M, bone marrow; (M), M-CLL. **d**, Therapies received before RT and presence/absence of SBS-melphalan, SBS-ganciclovir and SBS-RT at time of RT for each patient. mAb, monoclonal antibody; TBI, total body irradiation; Inh., inhibitor; Sig., signatures. **e**, Phylogenetic relationship of subclones and contribution of each mutational signature to their mutational profile. **f**, Relative contribution of mutational processes in CLL (no. 1) and RT subclones (top). Number of mutations (mut) and variant allele frequency (VAF) (bottom) of mutations assigned to the RT subclone during the disease course in patients 12 and 63 by high-coverage UMI-based NGS. Mutations are grouped according to the main peaks of SBS-RT. P values were obtained by Fisher's test. LC, low confidence; HC, high confidence; NA, not available. **g**, Distribution of the CCF of the single-nucleotide variants (SNVs) assigned to the RT subclone based on WGS and stratified according to the main peaks of the SBS-RT. **h**, Relative contribution of mutational processes in regions of kataegis in CLL and RT (left). Two cases acquiring mutations in the immunoglobulin genes at time of RT (right). **i**, Clonal evolution along the disease course in patient 12 inferred from WGS. Abbreviations for treatment regimens are detailed in Extended Data Fig. 1a. Each subclone is depicted by a different color and number and its CCF is proportional to its height in each time point (vertical line). The phylogeny of the subclones with the main driver alterations is shown (top). Flow cytometry analysis for time points (T) 4, 5 and 6 (bottom). The size of the cells (forward scatter (FSC) versus side scatter (SSC), first row) and the expression levels of CD20 and CD38 (second row) differentiated CLL cells (yellowish) and the two larger size tumor populations (pale and dark rose color, respectively). Numbers along axes are divided by 1,000.

mutations not associated with SBS-RT were detected earlier during the disease course (Fig. 2g and Extended Data Fig. 5h). These results suggest a causal link between the exposure to these drugs and SBS-RT. The finding of SBS-melphalan, SBS-ganciclovir and SBS-RT in RT argues in favor of a single-cell expansion model for RT; a single cell that can carry the footprints of cancer therapies (Fig. 2h). Contrarily, the lack of SBS-RT in the 27 post-treatment CLL samples (7 patients treated with bendamustine or chlorambucil) suggests that CLL relapse might be driven by the simultaneous expansion of different subclones, hindering the detection of SBS-RT through bulk sequencing^{4,36}.

RT subclones also acquired kataegis, mainly within the immunoglobulin loci, attributed to activation-induced cytidine deaminase (AID) activity (SBS84 and SBS85)^{29,32} (Fig. 2i and Extended Data Fig. 4). These kataegis led to the acquisition of mutations in the rearranged V(D)J gene in five RT cases (one after CIT and four targeted therapies) (Fig. 2i, Extended Data Fig. 5i,j and Supplementary Table 19). This canonical AID activity in RT is concordant with the acquisition of SBS9 mutations in two RT samples (4,686 (CIT) and 3,495 (targeted therapies)) and SVs mediated by aberrant class-switch recombination or somatic hypermutation in six RT (one before therapy, two CIT and three new agents), which targeted *MYC*, *MYCN*, *TRAF3* and *CCND3* (Fig. 1c and Supplementary Table 2).

SBS-RT mutations were found in CLL samples before the transformation in patient 3,299 although it was only present in the RT subclone (Fig. 2c,e). SBS-RT was also found in two different subclones in case 12 and 19. We speculated that these secondary subclones with SBS-RT (named 'RT-like' subclones) could correspond to the single-cell expansion of a 'transformed' cell that could have been missed by the routine analysis (Fig. 2c). The reanalysis of flow cytometry data available for case 12 detected two cell populations at time point (T) 4 differing in size and surface markers (likely CLL and RT-like subclones), whereas at T5 we detected an additional population of large cells (RT subclone, 0.2% cells) that expanded at T6, substituting the previous large cell population (RT-like subclone) (Fig. 2j and Extended Data Fig. 5k–m). WGS analysis showed that the RT-like and RT subclones diverged from a cell carrying a deletion of *CDKN2A/B* and truncation of *CREBBP*, each acquiring more than 2,100 specific mutations (Fig. 2e,j).

Altogether, these findings show that RT may arise simultaneously from different subclones and that such subclones can be detectable time before their final expansion and clinical manifestation. The identification of mutations in RT associated with early-in-time CLL therapies demonstrates that RT emerges from the clonal expansion of a single cell previously exposed to these therapies.

Dormant seeds of RT at CLL diagnosis. The WGS-based subclonal phylogeny of the nine patients with fully characterized longitudinal samples predicted that the RT subclone was present at low cancer cell fraction (CCF) in the preceding CLL samples in five (56%) patients and only detected at time of transformation in the remaining four (44%) (Fig. 3a). Indeed, the RT subclone was detected at time of CLL diagnosis in three of five patients, remained stable at a minute size (<1%) for 6–19 years of natural and treatment-influenced CLL course and expanded at the moment of clinical manifestations (patients 12, 19 and 63) (Fig. 3a). In the other two patients, the RT subclone was also detected in the first CLL sample analyzed but rapidly expanded driving the RT 0.6 and 3.5 years later in patients 3,034 and 3,299 (RT-PLL), respectively (Fig. 3a and Extended Data Fig. 6).

We next performed single-cell DNA sequencing (scDNA-seq) of 32 genes in 16 longitudinal samples of 4 patients (12, 19, 365 and 3,299) to validate these evolutionary histories of RT (202,210 cells passing filters, mean of 12,638 cells per sample; Fig. 1a, Supplementary Fig. 2 and Supplementary Table 20). Focusing on patient 19 with a time lapse of 14.4 years from diagnosis to RT (Fig. 3b), the RT subclone (subclone 5) at transformation (T6)

carried *CDKN2A/B* and *TP53* (p.G245D) alterations, whereas the main CLL subclones driving the relapse after therapy at T4 and T5 harbored a different *TP53* mutation (p.I195T; subclones 3 and 4). The WGS predicted the presence of all these subclones at CLL diagnosis (T1). Using scDNA-seq we identified two small populations accounting for 0.1% of cells carrying the *TP53* p.I195T and p.G245D mutations, respectively, at T1, which were also detected at relapse 7.2 years later (T3). The subclone carrying *TP53* p.I195T expanded to dominate the second relapse after 3.7 years at T4 and T5 but was substituted by the subclone carrying *TP53* p.G245D at T6 in the RT 14.4 years after diagnosis. All these subclones carried the *SF3B1* and *NOTCH1* mutations of the initial CLL subclone (Fig. 3c and Supplementary Table 20). The scDNA-seq of the three additional cases also corroborated the phylogenies and most of the dynamics inferred from WGS (Extended Data Fig. 6a). These results suggest that CLL evolution to RT is characterized by an early driver diversification probably generated before diagnosis, consistent with the early immunogenetic and DNA methylation diversification previously reported in CLL^{27–29} and that RT may emerge by a selection of pre-existing subclones carrying potent driver mutations rather than a de novo acquisition of leading clones.

As we identified five cases of RT carrying specific mutations in the immunoglobulin genes by WGS (Fig. 2i), we analyzed whether these immunoglobulin-based RT subclones were already present at CLL diagnosis using high-coverage NGS in patients 12 and 3,495 (Supplementary Table 21). Focusing on patient 3,495, for which the lack of germline material precluded our phylogenetic analyses, the RT occurring after treatment with ibrutinib harbored two new V(D)J mutations generating an unproductive IGH gene. NGS identified 0.002% sequences carrying the same two mutations at CLL diagnosis 1.72 years before (Fig. 3d). We also observed the expansion of additional unproductive subclones accounting for 11.8% of all sequences at time of RT, suggesting that BCR-independent subclones may have a proliferative advantage under therapy with BCR inhibitors (Fig. 3d). Similar results were found in patient 12 in which the V(D)J sequence of RT carrying a new mutation was already identified at CLL diagnosis 19.5 years before at DNA and RNA level (Fig. 3e). As the immunogenetic features represent a faithful imprint of the B cell of origin, the early identification of the same immunogenetic subclone provides further evidence for an early seeding of RT.

We finally tracked RT subclones during the disease course using single-cell RNA sequencing (scRNA-seq) of 19 longitudinal samples of five patients (24,800 tumor cells passing filters, mean of 1,305 cells per sample; Fig. 1a and Supplementary Table 22). As expected, RT and CLL cells had remarkably different gene expression profiles (Fig. 3f and Extended Data Fig. 7a–d). The transcriptome of CLL cells was dominated by three main clusters identified across patients and characterized by different expression of *CXCR4*, *CD27* and *MIR155HG*, respectively, which may represent the recirculation of CLL cells between peripheral blood and lymph nodes^{40–42} (Fig. 3f,g and Extended Data Fig. 7a–d). Contrarily, RT intracлонаl heterogeneity was mainly related to distinct proliferative capacities with a cluster of cells showing high *MKI67* and *PCNA* expression as well as high S and G2M cell-cycle phase scores. The remaining RT clusters were characterized by the expression of different marker genes among patients, including *CCND2*, *MIR155HG* and *TP53INP1* (Fig. 3f–h and Extended Data Fig. 7a–d). When considering each time point separately, we detected RT cells in all CLL samples before transformation in patient 12, 19, 63 and 3,299 but not in patient 365 (Fig. 3i and Extended Data Fig. 7a–i). The presence and dynamics of these RT subclones according to their transcriptomic profile recapitulated the findings obtained by WGS, scDNA-seq and immunoglobulin analyses in all five patients, suggesting that they captured the same cells. Indeed, using scRNA-seq we could identify the CNAs involved in simple and complex structural alterations found at time

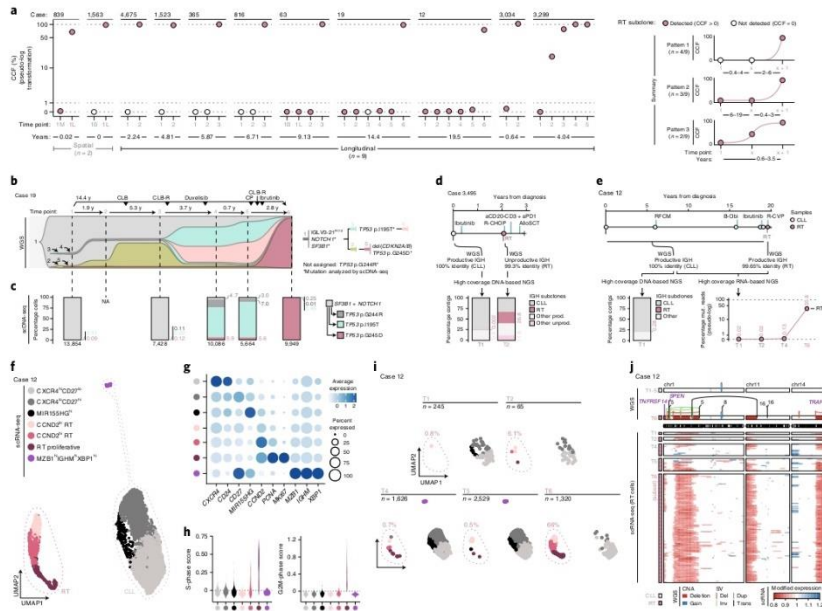


Fig. 3 | Early seeding of RT. **a**, Evolution of the RT subclone along the disease course based on WGS. Time lapse between the first and last sample analyzed (bottom). RT time points are marked in a rose color. Summary of the three patterns observed (right). **b**, Fish plot showing the clonal evolution along the course of the disease in patient 19 inferred from WGS analysis. Each subclone is depicted by a different color and number and its CCF is proportional to its height at each time point (vertical lines). Phylogeny of the subclones and main driver events (right). **c**, Mutation tree reconstructed by scDNA-seq for case 19 together with the fraction of cells carrying each specific combination of mutations in each time point. The total number of cells per sample is shown at the bottom. The number of cells assigned to each subclone is shown in Supplementary Table 20. **d**, Schematic representation of the clinical course and samples analyzed for patient 3,495 together with the size of the IGH subclones identified using high-coverage NGS analyses. Abbreviations for treatment regimens are detailed in Extended Data Fig. 1a. **e**, Clinical course and IGH subclones identified by DNA- and RNA-based NGS in patient 12. **f**, Uniform Manifold and Projection (UMAP) plot for case 12 based on the scRNA-seq data of all time points colored by annotation. **g**, Expression of key marker genes in each cluster identified in case 12 with the fraction of RT cells annotated. 'n', number of cells. **h**, Distribution of cell-cycle phase scores for each cluster based on scRNA-seq in case 12. **i**, UMAP visualization split by time point in case 12 with the fraction of RT cells annotated. 'n', number of cells. **j**, Chromosomal alterations detected by WGS in chromosomes 1, 11 and 14 in CLL and RT samples of patient 12 (top). Copy number profile of RT cells detected at the different time points according to scRNA-seq. Only a subset of RT cells from time point 6 (time of diagnosis of RT) was included for illustrative purposes (bottom).

of RT by WGS already in the dormant RT cells at CLL diagnosis and subsequent time points before their final expansion (Fig. 3) and Extended Data Fig. 8). These findings suggest an early acquisition of SVs, including chromothripsis and transcriptomic identity in RT.

To validate our observations, we reanalyzed the longitudinal scRNA-seq dataset from Penter et al.¹³ consisting of nine patients with CLL, one of which developed RT. In this case, we identified RT cells in the CLL sample collected 1.6 years before the RT (Extended Data Fig. 7). Overall, our integrative analyses uncovered a widespread early seeding of RT cells up to 19 years before their expansion and clinical manifestation.

OXPHOS^{hi}-BCR^{low} transcriptional axis of RT. To understand the transcriptomic evolution from CLL to RT and its epigenomic

regulation, we integrated genome-wide profiles of DNA methylation, chromatin activation (H3K27ac) and chromatin accessibility (ATAC-seq) with bulk RNA-seq and scRNA-seq of multiple longitudinal samples of six patients treated with BCR inhibitors (Fig. 1a). The DNA methylome of RT mainly reflected the naive and memory-like B cell derivation of their CLL counterpart, whereas chromatin activation and accessibility were remarkably different upon transformation (Fig. 4a). We identified 150 regions with increased H3K27ac and 426 regions that gained accessibility in RT (Fig. 4b, Extended Data Fig. 9a and Supplementary Tables 7 and 8). These *de novo* active regions were enriched in transcription factor (TF) families different from those known to modulate the epigenome of CLL⁴¹. Among them, 24 were enriched and upregulated in RT (Supplementary Table 7). The top TF was TEAD4, which

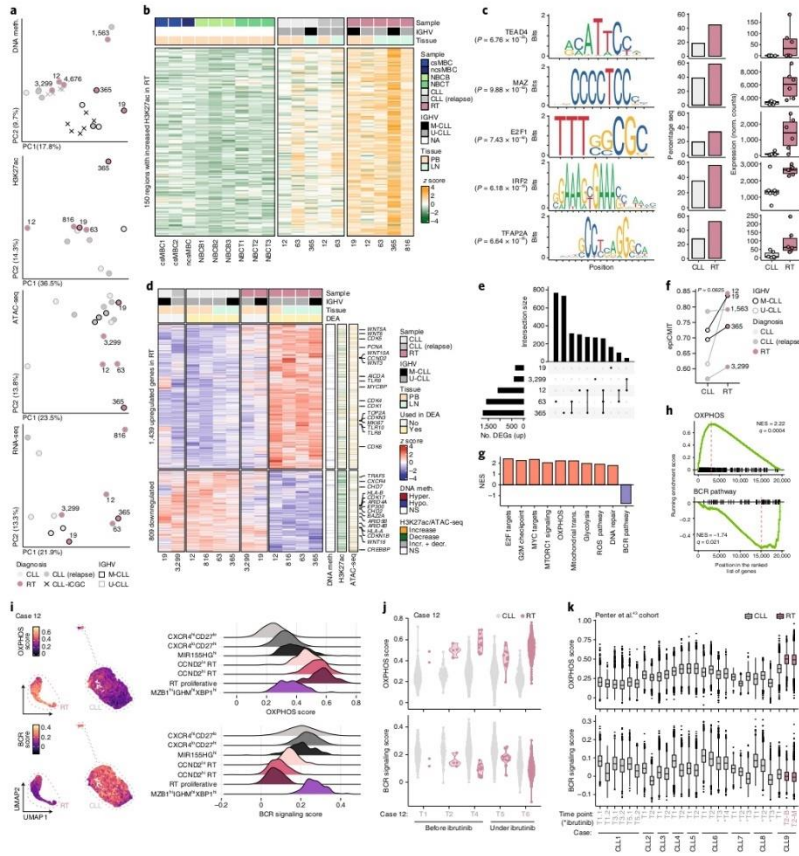


Fig. 4 | Proliferation, OXPHOS and BCR pathways dominate the epigenome and transcriptome of RT. a, PCA of the bulk epigenetic and transcriptomic layers analyzed. **b**, Heat map showing 150 regions with increased H3K27ac levels in RT. **c**, TF enriched within the ATAC peaks identified in the regions of increase H3K27ac in RT. The motif, percentage of RT-specific active regions and regions with increased H3K27ac in CLL that contained the motif and TF expression (bulk RNA-seq) in CLL and RT are shown. Center line indicates median; box limits indicate upper and lower quartiles; whiskers indicate 1.5x interquartile range; points indicate individual samples. *P* values were derived using a one-tailed Wilcoxon rank-sum test. **d**, Heat map showing the DEGs between CLL and RT identified by bulk RNA-seq. Samples used in the differential expression analysis (DEA) are indicated. The overlap of DEGs with DNA methylation changes, H3K27ac and ATAC peaks is shown on the right. Selected genes are annotated. **e**, Intersection of upregulated genes in RT compared to CLL in scRNA-seq analyses. **f**, epicMIT evolution from CLL to RT. *P* values were derived by paired Wilcoxon signed-rank test. **g**, Summary of the main gene sets modulated in RT based on bulk RNA-seq. NES, normalized enrichment score; ROS, reactive oxygen species. **h**, Gene set enrichment plot for OXPHOS and BCR signaling (bulk RNA-seq). Ridge plots show the OXPHOS and BCR score across clusters (right). **i**, OXPHOS and BCR signaling scores of CLL and RT cells are highlighted (left). Ridge plots show the OXPHOS and BCR score across clusters (right). **j**, OXPHOS and BCR signaling scores of CLL and RT cells of patient 12 across time points by scRNA-seq. **k**, Distribution of OXPHOS and BCR signaling scores at a single-cell level across different time points of nine cases included in the study of Penter et al.⁴³. Center line indicates median; box limits indicate upper and lower quartiles; whiskers indicate 1.5x interquartile range; points indicate outliers. B, peripheral blood; M, bone marrow. *Sample collected under treatment with ibrutinib.

1668

NATURE MEDICINE | VOL 28 | AUGUST 2022 | 1662–1671 | www.nature.com/naturemedicine

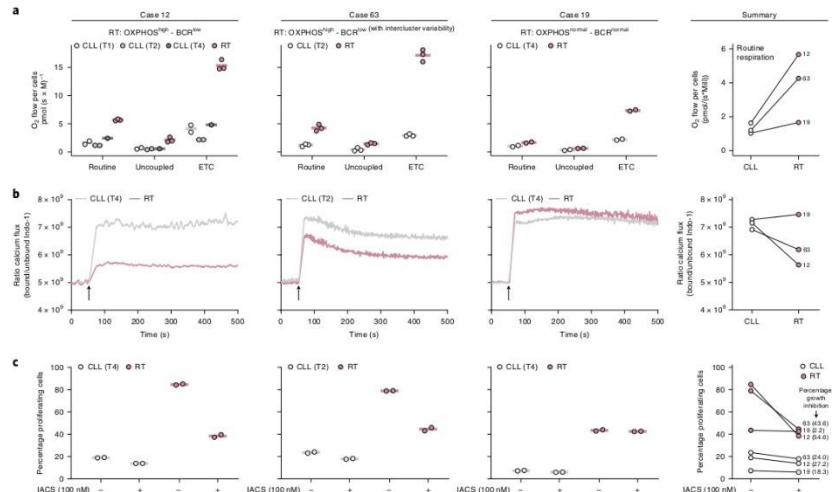


Fig. 5 | Cellular respiration, BCR signaling and OXPHOS inhibition in RT cells. a, Oxygen consumption of intact CLL and RT cells of three patients at routine respiration (routine), oligomycin-inhibited leak respiration (uncoupled) and uncoupler-stimulated ETC. Each dot represents a technical replicate. The mean of the replicates is shown using a horizontal line (left). Summary of the routine respiration of CLL and RT cells of the three patients collapsed (right). **b**, Calcium kinetics of tumoral cells (CD19⁺, CD5⁻) upon stimulation with 4-hydroxytamoxifen (4-OHT) and anti-BCR (black arrow). Basal calcium was adjusted at 5×10^3 Indo-1 ratio for 60 s before cell stimulation with F(ab')₂ anti-human IgM + H₂O₂ at 37 °C. Then, Ca²⁺ flux was recorded up to 500 s (left). Summary of the calcium release after BCR stimulation of CLL and RT cells. Average mean fluorescence after stimulation is represented (right). **c**, Cell proliferation after 72-h incubation with or without IACS-010759 (IACS) at 100 nM. Percentage of proliferating cells was determined by carboxyfluorescein succinimidyl ester (CFSE) cell tracer. Two technical replicates of each sample were performed (left). Summary of the proliferation for each CLL and RT cells with or without IACS treatment after 72 h. The normalized percentage of growth inhibition is indicated (right).

activates genes involved in oxidative phosphorylation (OXPHOS) through the mTOR pathway¹¹ and co-operates with MYCN¹⁶. Additional TFs were related to MYC (MAZ), proliferation/cell cycle (E2F family) or IRF family, among others (Fig. 4c). Notably, high IRF4 levels seem to attenuate BCR signaling in CLL¹⁷, whereas they are necessary to induce MYC target genes, OXPHOS and glycolysis in activated healthy B cells¹⁸.

The RNA-seq analysis, excluding cases 19 and 3,299 (RT-PLL) due to their intermediate transcriptomic profile, identified 2,248 differentially expressed genes (DEGs) between RT and CLL (1,439 upregulated and 809 downregulated) (Fig. 4a,d,e, Extended Data Fig. 10a and Supplementary Tables 11 and 23). A remarkable fraction of upregulated/downregulated genes overlapped with regions with the respective increase/decrease of H3K27ac (20%) and chromatin accessibility (16%) at RT (Fig. 4d and Extended Data Fig. 9b). Contrarily, only 4% of the DEGs overlapped with any of the 2,341 differentially methylated CpGs (DMCs) between RT and CLL, emphasizing the limited effect of DNA methylation on gene regulation⁹. Most DMCs were hypomethylated at RT (2,112 of 2,341; 90%), found in open sea and intergenic regions and correlated with the proliferative history of the cells measured by the epiCMT score¹⁹ (1,681; 72%), which increased during CLL evolution and at RT (Fig. 4d,f, Extended Data Fig. 9c–g and Supplementary Table 6).

Genes upregulated in RT involved pathways that seem independent of BCR signaling such as Wnt (WNT5A and others)²⁰, Toll-like

receptors (TLR9 among others)²¹ and a number of cyclin-dependent kinases. Downregulated genes included, among others, CXCR4, HLA-A/B and chromatin remodelers also targeted by genetic alterations in some cases (Fig. 4d and Extended Data Fig. 10b,c). Gene sets modulated by gene expression in RT were in harmony with the identified chromatin-based changes and included upregulation of E2F targets, G2M checkpoints, MYC targets, MTORC1 signaling, OXPHOS, mitochondrial translation, glycolysis, reactive oxygen species and DNA repair pathways, among others. In addition, RT showed downmodulation of BCR signaling (Fig. 4g,h, Extended Data Fig. 10d and Supplementary Table 11). The OXPHOS^{high}-BCR^{low} pattern observed by bulk RNA-seq in RT was further refined using scRNA-seq: two of five tumors had OXPHOS^{high}-BCR^{low} (12 and 63, although the latter showed some intercluster variability), the two M-CLL carrying IGLV3-21^{R110} had RT with BCR expression similar to CLL and were OXPHOS^{high}-BCR^{normal} (365) or OXPHOS^{normal}-BCR^{normal} (19) and the RT-PLL (3,299) was OXPHOS^{low}-BCR^{low} (Fig. 4i, Extended Data Fig. 10e–j and Supplementary Table 23). In addition, the scRNA-seq analysis showed that the OXPHOS/BCR profiles of RT were already identified in the early dormant RT cells, suggesting that they might represent an intrinsic characteristic of RT cells rather than being modulated by BCR inhibitors (Fig. 4j and Extended Data Fig. 10g–j). To expand these observations, we measured the expression of OXPHOS and BCR pathways in the scRNA-seq dataset from Penter et al.¹¹. Case CLL9, which

developed RT in the absence of any therapy, showed a remarkably higher OXPPOS and slightly lower BCR expression at time of RT compared to CLL (Fig. 4k and Extended Data Fig. 10k,l).

Overall, the epigenome and transcriptome of RT converge to an OXPPOS^{high}-BCR^{low} axis reminiscent of that observed in the de novo DLBCL subtype characterized by high OXPPOS (DLBCL-OXPPOS) and insensitive to BCR inhibition^{32–54}. This axis might explain the selection and rapid expansion of small RT subclones under therapy with BCR inhibitors.

OXPPOS and BCR activity in RT. We next validated experimentally the OXPPOS and BCR activity of RT in samples of patients 12, 19 and 63. Respiriometry assays confirmed that OXPPOS^{high} RT cells (patients 12 and 63) had a 3.5-fold higher oxygen consumption at routine respiration and fivefold higher electron transfer system capacity (ETC) compared to CLL. In addition, OXPPOS^{normal} RT (patient 19) showed a routine oxygen consumption similar to CLL, although also had a relatively higher ETC than its CLL counterpart (Fig. 5a, Supplementary Fig. 3a–d and Supplementary Table 24). BCR signaling measured by Ca²⁺ mobilization upon BCR stimulation with IgM showed that BCR^{low} RT cells (patients 12 and 63) had a lower Ca²⁺ flux compared to CLL, which contrasted with the higher flux observed in the BCR^{normal} RT cells of patient 19, concordant with its IGLV3–21^{R110} mutation²⁷ (Fig. 5b, Supplementary Fig. 4a,b and Supplementary Table 25).

To determine the biological effect of OXPPOS^{high} in RT, we performed in vitro proliferation assays using IACS-010759 (100 nM), an OXPPOS inhibitor that targets mitochondrial complex I (Supplementary Figs. 3c and 4c and Supplementary Table 25). OXPPOS^{high} RT (patients 12 and 63) had a higher proliferation at 72 h compared to OXPPOS^{normal} RT (patients 19) and all of them were higher than their respective CLL. OXPPOS inhibition resulted in a marked decrease in proliferation in OXPPOS^{high} RT (mean 49.1%), which contrasted with that observed in OXPPOS^{normal} RT (2.2% decrease) and CLL (23.2% decrease) (Fig. 5c and Supplementary Fig. 4d). Overall, these results confirm the role of OXPPOS^{high} phenotype in high proliferation of RT and suggest its potential therapeutic value in RT as proposed for other neoplasms^{33–37}.

Discussion

The genome of RT is characterized by a compendium of driver alterations in cell cycle, MYC, NOTCH and NF- κ B pathways, frequently targeted in single catastrophic events and by the footprints of early-in-time, treatment-related, mutational processes, including the new SBS-RT potentially associated with bendamustine and chlorambucil exposure. A very early diversification of CLL leads to emergence of RT cells with fully assembled genomic, immunogenetic and transcriptomic profiles already at CLL diagnosis up to 19 years before the clonal explosion associated with the clinical transformation. RT cells have a notable shift in chromatin configuration and transcriptional program that converges into activation of the OXPPOS pathway and downregulation of BCR signaling, the latter potentially compensated by activating Toll-like, MYC and MAPK pathways^{17,31,38,39}. The rapid expansion of RT subclones under treatment with BCR inhibitors is consistent with its low BCR signaling, except when carrying the IGLV3–21^{R110} and further supported by the increased number of subclones carrying unproductive immunoglobulin genes and the development of RT with plasmablastic differentiation, a cell type independent of BCR signaling⁶⁰. Finally, we also uncovered that OXPPOS inhibition reduced the proliferation of RT cells in vitro, a finding worth exploring in future therapeutic strategies^{33–37}.

In conclusion, our comprehensive characterization of CLL evolution toward RT has revealed new genomic drivers and epigenomic reconfiguration with very early emergence of subclones driving late stages of cancer evolution, which may set the basis for

developing single-cell-based predictive strategies. Furthermore, this study also identifies new RT-specific therapeutic targets and suggests that early intervention to eradicate dormant RT subclones may prevent the future development of this lethal complication of CLL.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-022-01927-8>.

Received: 10 November 2021; Accepted: 1 July 2022;

Published online: 11 August 2022

References

1. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
3. Drento, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239–2254 (2021).
4. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
5. Ferrando, A. A. & López-Otín, C. Clonal evolution in leukemia. *Nat. Med.* **23**, 1135–1145 (2017).
6. Ding, W. Richter transformation in the era of novel agents. *Hematology* **2018**, 256–263 (2018).
7. Maddocks, K. J. et al. Etiology of ibrutinib therapy discontinuation and outcomes in patients with chronic lymphocytic leukemia. *JAMA Oncol.* **1**, 80 (2015).
8. Ahn, I. E. et al. Clonal evolution leading to ibrutinib resistance in chronic lymphocytic leukemia. *Blood* **129**, 1469–1479 (2017).
9. Jain, P. et al. Outcomes of patients with chronic lymphocytic leukemia after discontinuing ibrutinib. *Blood* **125**, 2062–2067 (2015).
10. Landau, D. A. et al. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat. Commun.* **8**, 2185 (2017).
11. Bea, S. et al. Genetic imbalances in progressed B-cell chronic lymphocytic leukemia and transformed large-cell lymphoma (Richter's syndrome). *Am. J. Pathol.* **161**, 957–968 (2002).
12. Scandurra, M. et al. Genomic profiling of Richter's syndrome: recurrent lesions and differences with de novo diffuse large B-cell lymphomas. *Hematol. Oncol.* **28**, 62–67 (2010).
13. Rossi, D. et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood* **117**, 3391–3401 (2011).
14. Fabbri, G. et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J. Exp. Med.* **210**, 2273–2288 (2013).
15. Chigrimova, E. et al. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood* **122**, 2673–2682 (2013).
16. Klintman, J. et al. Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood* **137**, 2800–2816 (2021).
17. Chakraborty, S. et al. B-cell receptor signaling and genetic lesions in TP53 and CDKN2A/CDKN2B cooperate in Richter transformation. *Blood* **138**, 1053–1066 (2021).
18. Anderson, M. A. et al. Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood* **129**, 3362–3370 (2017).
19. Miller, C. R. et al. Near-tetraploidy is associated with Richter transformation in chronic lymphocytic leukemia patients receiving ibrutinib. *Blood Adv.* **1**, 1584–1588 (2017).
20. Kadri, S. et al. Clonal evolution underlying leukemia progression and Richter transformation in patients with ibrutinib-relapsed CLL. *Blood Adv.* **1**, 715–727 (2017).
21. Herling, C. D. et al. Clonal dynamics towards the development of venetoclax resistance in chronic lymphocytic leukemia. *Nat. Commun.* **9**, 727 (2018).
22. Villamor, N. et al. NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia* **27**, 1100–1106 (2013).

23. De Paoli, L. et al. MGA, a suppressor of MYC, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leuk. Lymphoma* **54**, 1087–1090 (2013).
24. Rossi, D. et al. Different impact of NOTCH1 and SF3B1 mutations on the risk of chronic lymphocytic leukemia transformation to Richter syndrome. *Br. J. Haematol.* **158**, 426–429 (2012).
25. Chitalia, A. et al. Descriptive analysis of genetic aberrations and cell of origin in Richter transformation. *Leuk. Lymphoma* **60**, 971–979 (2019).
26. Benatti, S. et al. IRF4 L116R mutation promotes proliferation of chronic lymphocytic leukemia B cells inducing MYC. *Hematol. Oncol.* **39**, 707–711 (2021).
27. Minici, C. et al. Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukemia. *Nat. Commun.* **8**, 15746 (2017).
28. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukemia. *Nature* **526**, 519–524 (2015).
29. Kasir, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukemia evolution. *Nat. Commun.* **6**, 8866 (2015).
30. Maury, F. et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
31. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).
32. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
33. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
34. Rustad, E. H. et al. Timing the initiation of multiple myeloma. *Nat. Commun.* **11**, 1917 (2020).
35. de Kanter, J. K. et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, 1726–1739 (2021).
36. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
37. Gatti, F. et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukemia. *Nature* **569**, 576–580 (2019).
38. Gemenetzis, K. et al. Higher-order immunoglobulin repertoire restrictions in CLL: the illustrative case of stereotyped subsets 2 and 169. *Blood* **137**, 1895–1904 (2021).
39. Bagnara, D. et al. Post-transformation IGHV-IGHD-IGHJ mutations in chronic lymphocytic leukemia B cells: implications for mutational mechanisms and impact on clinical course. *Front. Oncol.* **11**, 1769 (2021).
40. Calissano, C. et al. In vivo intraclonal and interclonal kinetic heterogeneity in B-cell chronic lymphocytic leukemia. *Blood* **114**, 4832–4842 (2009).
41. Calissano, C. et al. Intraclonal complexity in chronic lymphocytic leukemia: features enriched in recently born/divided and older/quiescent cells. *Mol. Med.* **17**, 1374–1382 (2011).
42. Cui, B. et al. MicroRNA-155 influences B-cell receptor signaling and associates with aggressive disease in chronic lymphocytic leukemia. *Blood* **124**, 546–554 (2014).
43. Feister, L. et al. Longitudinal single-cell dynamics of chromatin accessibility and mitochondrial mutations in chronic lymphocytic leukemia mirror disease history. *Cancer Discov.* **11**, 3048–3063 (2021).
44. Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
45. Chen, C.-L. et al. Arginine is an epigenetic regulator targeting TEAD4 to modulate OXPHOS in prostate cancer cells. *Nat. Commun.* **12**, 2398 (2021).
46. Rajbhandari, P. et al. Cross-cohort analysis identifies a TEAD4-MYC-N positive feedback loop as the core regulatory element of high-risk neuroblastoma. *Cancer Discov.* **8**, 582–599 (2018).
47. Maffei, R. et al. IRF4 modulates the response to BCR activation in chronic lymphocytic leukemia regulating IKAROS and SYK. *Leukemia* **35**, 1330–1343 (2021).
48. Patterson, D. G. et al. An IRF4-MYC-mTORC1 integrated pathway controls cell growth and the proliferative capacity of activated B cells during B cell differentiation in vivo. *J. Immunol.* **207**, 1798–1811 (2021).
49. Duran-Ferrer, M. et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nat. Cancer* **1**, 1066–1081 (2020).
50. Hasan, M. K., Ghia, E. M., Rassenti, L. Z., Widhopf, G. F. & Kipps, T. J. Wnt5a enhances proliferation of chronic lymphocytic leukemia and ERK1/2 phosphorylation via a ROR1/DOCK2-dependent mechanism. *Leukemia* **35**, 1621–1630 (2021).
51. Noufa, S., Villa, M. G., Stamatopoulos, K., Ghia, P. & Muzio, M. Toll-like receptors signaling: a complex network for NF- κ B activation in B-cell lymphoid malignancies. *Semin. Cancer Biol.* **39**, 15–25 (2016).
52. Monti, S. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105**, 1851–1861 (2005).
53. Caro, P. et al. Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. *Cancer Cell* **22**, 547–560 (2012).
54. Norberg, E. et al. Differential contribution of the mitochondrial translation pathway to the survival of diffuse large B-cell lymphoma subsets. *Cell Death Differ.* **24**, 251–262 (2017).
55. Medina, J. R. et al. An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nat. Med.* **24**, 1036–1046 (2018).
56. Vangapandu, H. V. et al. Biological and metabolic effects of IACS-010759, an OxPhos inhibitor, on chronic lymphocytic leukemia cells. *Oncotarget* **9**, 24980–24991 (2018).
57. Zhang, L. et al. Metabolic reprogramming toward oxidative phosphorylation identifies a therapeutic target for mantle cell lymphoma. *Sci. Transl. Med.* **11**, eaui1167 (2019).
58. Varano, G. et al. The B-cell receptor controls fitness of MYC-driven lymphoma cells via GSK3 β inhibition. *Nature* **546**, 302–306 (2017).
59. Dadashian, E. L. et al. TLR signaling is activated in lymph node-resident CLL cells and is only partially inhibited by ibrutinib. *Cancer Res.* **79**, 360–371 (2019).
60. Chan, K.-L. et al. Plasmablastic Richter transformation as a resistance mechanism for chronic lymphocytic leukemia treated with BCR signalling inhibitors. *Br. J. Haematol.* **177**, 324–328 (2017).

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



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

© The Author(s) 2022

Methods

Consent and sample processing. Written informed consent was obtained from all patients. The study was approved by the Hospital Clinic of Barcelona Ethics Committee. Tumor DNA was extracted from tumor cells purified from fresh/frozen preserved mononuclear cells, frozen lymph nodes or formalin-fixed paraffin-embedded (FFPE) tissue ($n=1$, CLL sample of patient 1,669). Germline DNA was obtained from the non-tumoral purified cell fraction in 12 cases. In two patients (1,523 and 4,675) who had received allogeneic stem-cell transplant before RT, germline DNA of the donor was also collected. All extractions were performed using appropriate QIAGEN kits (QIAamp DNA Blood Maxi kit, cat. no. 51194; QIAamp DNA Mini kit, cat. no. 51304; and AllPrep DNA/RNA FFPE kit, cat. no. 80234). Tumor RNA was obtained from tumor cells purified from fresh/frozen preserved mononuclear cells with TRIzol reagent (Invitrogen, cat. no. 15596026).

A specific flow cytometry analysis was conducted on peripheral blood samples of patient 12, which were stained with the Lymphocyte Screening Tube according to EuroFlow protocols (<https://www.euroflow.org/protocols>). At least 100,000 cells were acquired in a FACSCanto II instrument. Analysis was conducted using the Infinicyt 2.0 software. The sequential gating analysis was as follows: singlet identification in a FSC-W versus FSC-H plot; leucocyte identification in SSC-A versus CD45 (V500-C) plot and FSC-A versus SSC-A; lymphocytes identified as SSC-A low and CD45 high and back-gated in FSC-A versus SSC-A to exclude monocytes; in the lymphocyte gate, T cells were identified as CD3⁺ cells in SSC-A versus CD3 (APC) followed by sequentially distinguishing TCR β ⁺ T cells, CD4 T cells and CD8 T cells after excluding T cells. B cells were selected in a SSC-A versus CD19 (PE-Cy7), followed by inspection of CD19 (PE-Cy7) versus CD20 (PacB), CD5 (PerCPy5.5) versus CD20 (PacB) and CD20 (PacB) versus CD38 (APC-H7) plots to evaluate the expression of these B cell markers and the assignment of κ and λ expression in a plot of Igk (PE) versus Igl. (FITC); after excluding B cells, natural killer cells were identified in a SSC-A versus CD56 (PE) plot followed by SSC-A versus CD38 (APC-H7) plot.

WGS and WES. Library preparation and sequencing. All samples available were subjected to WGS except the FFPE CLL, which was analyzed by whole-exome sequencing (WES). WGS libraries were performed using the Kapa Library Preparation kit (Roche, cat. no. 07961901001), TruSeq DNA PCR-Free kit (Illumina, cat. no. 20015963) or TruSeq DNA Nano protocol (Illumina, cat. no. 20015965) and sequenced on a HiSeq 2000/4000/X Ten (2×126 bp or 2×151 bp) or NovaSeq 6000 (2×151 bp) instrument (Illumina). WES was performed using the SureSelect Human All Exon V5 (Agilent Technologies, cat. no. 5190-6209 and G9611B) coupled with a KAPA Hyper Prep kit (Roche, cat. no. 0796236001) for the DNA pre-capture library. Sequencing was performed on a HiSeq 2000 (2×101 bp). We also included WGS of three published CLL/germline pairs (patients 12, 19 and 63)³ (Supplementary Table 1).

General considerations. Overall, 12 patients had a complete dataset (germline, CLL and RT samples), 6 patients lacked germline DNA and 1 patient had only the RT sample (case 4,676). We conducted tumor versus normal analyses in cases with a complete dataset. For the six patients lacking the germline sample, we used the CLL samples as 'normal' to identify SNV acquired at RT for mutational signature analyses. In addition, tumor-only analyses were conducted in these CLL and RT samples, as well as in the patient with only a RT sample available, to identify driver gene mutations and genome-wide CNAs (Supplementary Table 1).

Read mapping and quality control. Reads were mapped to the human reference genome (GRCh37) using the BWA-MEM algorithm (v0.7.15)³⁶. BAM files were generated and optical/PCR duplicates flagged using biobambam2 (v2.0.65, <https://github.com/german.tischler/biobambam2>). FastQC (v0.11.5, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and Picard (v2.10.2, <https://broadinstitute.github.io/picard/>) were used to extract quality control metrics. Mean coverage was $33 \times$ and $119 \times$ for WGS and WES, respectively (Supplementary Table 1).

Immunoglobulin gene characterization. Immunoglobulin gene rearrangements were characterized using IgCaller (v1.2)³⁷. The rearranged sequences obtained were reviewed on the Integrative Genomics Viewer (IGV; v2.9.2)³⁸ and annotated using IMGT/QUEST (https://www.imgt.org/IMGT_quest/) and ARRES/AssignSubsets (<http://bat.linfipire.org/arrest/assignsubsets>).

Tumor versus normal SNVs and indel calling. SNVs were called using Sidrón³⁹, CaVEMan (cgpCaVEManWrapper, v1.1.2.0)⁴⁰, Mutect2 (Genome Analysis Toolkit (GATK) v4.0.2.0)⁴¹ and MuSE (v1.0 rc)⁴² and normalized using bcftools (v1.8)⁴³. Variants detected by CaVEMan with more than half of the mutant reads clipped (CLPM > 0) and with supporting reads with a median alignment score (ASMD) < 90, < 120 or < 140 for sequencing read lengths of 100, 125 or 150 bp, respectively, were excluded. Variants called by Mutect2 with MMQ < 60 were eliminated. Mutations detected by at least two algorithms were considered. Short insertions/deletions (indels) were called by SMuFin (v0.9.4)⁴⁴, Pindel (cgpPindel, v2.2.3)⁴⁵, SvABA (v7.0.2)⁴⁶, Mutect2 (GATK v4.0.2.0)⁴¹ and Platypus (v0.8.1)⁴⁷. The somaticMutationDetector.py script (<https://github.com/andyrimmer/Platypus/blob/master/extensions/Cancer/somaticMutationDetector.py>) was used to identify somatic indels called by Platypus. Indels were left-aligned and normalized using bcftools⁴³. Indels with MMQ < 60, MQ < 60 and MAPQ < 60 for Mutect2, Platypus and SvABA, respectively, were removed. Only indels identified by at least two algorithms were retained. Annotation of mutations was performed using snpEff/snpSift (v4.3e)⁴⁸ and GRCh37.p13.RefSeq as a reference. This approach showed a 93% specificity and 88% sensitivity when benchmarked against the mutations found at a VAF > 10% in our previous high-coverage NGS study³.

Tumor-only SNVs and indel calling. Tumor-only variant calling was restricted to coding regions of 243 genes described as drivers in CLL and other B cell lymphomas (Supplementary Table 10). Minp-BAM files were obtained using Picard tools and variant calling was performed using Mutect2 (GATK v4.0.4.0)⁴¹, VarScan2 (v2.4.3)⁴⁹, VarDictJava (v1.4)⁵⁰, LoFreq (v2.1.3.1)⁵¹, outlyzer (v1.0)⁵² and freebayes (v1.1.0, <https://github.com/freebayes/freebayes>). Variants were normalized using bcftools (v1.9)⁴³ and annotated using snpEff/snpSift (v4.3e)⁴⁸. Only non-synonymous variants that were identified as PASS by ≥ 2 algorithms were considered. Variants reported in 1000 Genomes Project, ExAC or gnomAD with a population frequency > 1% or reported as germline in our ICGC database of 506 WES/WGS³ were considered as polymorphisms.

Tumor versus normal CNA calling. CNAs were called using Battenberg (cgpBattenberg, v3.2.2)⁵³ and ASCAT (ascatNgs, v4.1.0)⁵⁴. CNAs within any of the immunoglobulin loci were not considered. We used the tumor purities obtained by Battenberg in downstream analyses. The median tumor cell content was 91.5% (Supplementary Table 1).

Tumor-only CNA calling. CNAs were extracted using CNVkit (v0.9.3)⁵⁵. CNAs < 500 kb, with an absolute log₂ copy ratio (log₂ CR) < 0.3 or located within any of the immunoglobulin loci were removed. CNAs were classified as gains if log₂ CR > 0.3, deletions if log₂ CR < -0.3, high-copy gains if log₂ CR > 1.1 and homozygous deletions if log₂ CR < -1.1. The log₂ CR cutoff was set to 0.15 for two samples with low tumor cell content (102-01-01TD and 4690-03-01BD). To avoid a high segmentation of the CNA profile, CNAs belonging to the same class were merged if they were separated by < 1 Mb and had an absolute log₂ CR difference < 0.25.

Array-based CNA calling in FFPE. CNAs were examined in the FFPE CLL sample using the OncoScan CNV FFPE Assay kit (Thermo Fisher Scientific, cat. no. 902695) and analyzed using Nexus 9.0 software (Biodiscovery).

Tumor versus normal SV calling. SVs were extracted using SMuFin (v0.9.4)⁴⁴, BRASS (v6.0.5)⁵⁶, SvABA (v7.0.2)⁴⁶ and DELLY2 (v0.8.1)⁵⁷. SVs identified were intersected considering a window of 300 bp around break points. We kept for downstream analyses the SVs identified by at least two programs if at least one of the algorithms called the alteration with high quality (MAPQ ≥ 90 for BRASS, MAPQ = 60 for SvABA and DELLY2). In addition, IgCaller (v1.2)³⁷ was used to call SVs within any of the immunoglobulin loci. All SVs were visually inspected using IGV³⁸. SVs were categorized into simple or complex events. Chromothripsis⁵⁸ was defined as ≥ 7 oscillating changes between two or three copy number states or the presence of > 7 SV break points occurring in a single chromosome and supported by additional criteria⁵⁹. Chromoplexy was determined by the presence of ≥ 3 chained chromosomal rearrangements, where chains were identified using a window of 50 kb⁵⁸. Cycles of templated insertions were defined as copy number gains in ≥ 3 chromosomes linked by SVs⁶⁰. Breakage-fusion bridge cycles were defined as patterns of focal copy number increases and fold-back inversions, together with telomeric deletions. Chains of rearrangements having > 2 SVs and not fulfilling any of the previous criteria were classified as 'other complex events'. Chromothripsis and 'other complex events' were subcategorized according to the number of chromosomes involved. The longitudinal nature of our dataset allowed us to refine the obtained classification based on the presence of the involved alterations in each time point analyzed.

Patients who underwent allogeneic stem-cell transplant. In these patients, we conducted tumor versus patient's germline and tumor versus donor's germline variant calling in parallel. Only the intersection of variants identified was considered.

Rescue of alterations based on longitudinal information. SNVs called in one sample were automatically added to the samples of additional time point(s) if at least one high-quality read with the mutation was found in the BAM file (alleleCounter v4.0.0, parameters: min_map_qual = 35; and min_base_qual = 20). Similarly, indels and SVs detected in one sample were added in the additional time point(s) if any of the algorithms detected the alteration, regardless of its filters.

WGS-based subclonal reconstruction. A Markov chain Monte Carlo sampler for a Dirichlet process mixture model was used to infer putative subclones, to assign mutations to subclones and to estimate the subclone frequencies in each sample from the SNV read counts, copy number states and tumor purities (Supplementary Table 17)⁶⁰. Clusters with < 100 mutations were excluded. The phylogenetic relationships between subclones were identified following the

'pigeonhole principle' which was relaxed using a case-specific 'tolerated error'³⁸. Clusters not assigned to the reconstructed phylogenetic tree were excluded. Fish plots were generated using the TimeScape R package (v1.6.0). The CGF of indels was calculated integrating read counts, CNAs and tumor purity³⁹. Driver indels subjected to validation by scDNA-seq and/or relevant to the tumor phylogeny were manually assigned to subclones. Similarly, driver CNAs relevant to the phylogeny were manually assigned. Seven SNVs found in *TP53/ATM* overlapping with CNAs were manually assigned to the most likely subclone as they were not automatically assigned by the Dirichlet process and were subjected to scDNA-seq (Supplementary Table 9).

Mutational signatures. We studied mutational signatures acting genome-wide and in localized regions (inter-mutation distance $\leq 1\text{kb}$)⁴⁰. We integrated the mutations identified in this CLL/RT cohort together with those of 147 CLL treatment-naïve samples (ICGC-CLL)⁴¹ and 27 new CLL collected at relapse post-treatment (mean coverage 31.5x; Supplementary Table 15). The WGS of these two additional cohorts was (re-)analyzed using our current bioinformatic pipeline (Supplementary Table 12). Mutational signatures were analyzed for SNVs or single-base substitutions (SBSs) according to their 5' and 3' flanking bases following three steps⁴²:

1. Extraction: de novo signature extraction was performed using a hierarchical Dirichlet process (HDP, v0.1.5; <https://github.com/nicolaroberts/hdp>), SignatureAnalyzer (v0.0.7)⁴³, SigProfiler (SigProfilerExtractor, v1.0.8)⁴⁴ and sigfit (v2.0.0; <https://github.com/ignacio/sigfit>). HDP was run with four independent posterior sampling chains, followed by 20,000 burn-in iterations and the collection of 200 posterior samples of each chain with 200 iterations between each. SigProfiler was run with 1,000 iterations and a maximum of ten extracted signatures. Similarly, sigfit was run to extract five signatures with 10,000 burn-in iterations and 20,000 sampling iterations.
2. Assignment: each extracted signature was assigned to a given COSMIC signature (v3.2)⁴⁵ if their cosine similarity was >0.85 . Otherwise, the extracted signature was decomposed into n COSMIC signatures using an expectation maximization (EM) algorithm⁴⁶. The EM algorithm was first run using the COSMIC signatures identified in the previous step. If their cosine similarity was <0.85 , we ran the EM algorithm, including all signatures reported in COSMIC and by Kucal et al.⁴⁷ (55 mutational signatures related to environmental agents). Three exceptions were made: (1) we combined two HDP signatures that together constituted COSMIC signature SBS5 to avoid splitting of signatures (Extended Data Fig. 4a); (2) APOBEC signatures (SBS2 and SBS13) were favored to be assigned to one of the signatures extracted by HDP and SignatureAnalyzer although it was not the best EM solution probably because they were only found in one sample, which impaired a clean extraction of the signatures (Extended Data Fig. 4f); and (3) one signature extracted by HDP and SignatureAnalyzer was directly assigned to the mutational signature associated with ganciclovir treatment⁴⁸ (cosine similarity 0.987 and 0.993, respectively) (Extended Data Fig. 4). The new SBS-RT extracted by HDP was considered for downstream analyses as it had less background noise than the one extracted by SignatureAnalyzer, favoring a higher specificity during the fitting step. Similarly, the SBS-ganciclovir extracted by HDP was used in downstream analyses (Extended Data Fig. 4). We also performed a detailed review to remove signatures susceptible of being originated due to sequencing artifacts (Supplementary Table 13).
3. Fitting: we used a fitting approach (MutationalPatterns, v3.0.1) to measure the contribution of each mutational signature in each sample. Based on (1) the de novo identification of the therapy-related SBS-ganciclovir and (2) that two patients received melphalan before RT, the mutational signature associated with melphalan therapy⁴⁹ was also included in this step. To avoid the so-called inter-sample bleeding effect⁵⁰, we iteratively removed the less-contributing signature if its removal decreased the cosine similarity between the original and reconstructed 96-profile <0.01 (ref.⁵¹). SBS1 and SBS5 were added if addition improved the cosine similarity⁵². Similarly, SBS9 was added in CLL/RT samples classified as M-CLL if addition improved the cosine similarity. We also ran mSigAct (v2.1.1; <https://github.com/stevenoren/mSigAct>) to confirm the presence/absence of SBS-melphalan (Supplementary Table 15). To assess the contribution of each signature to each subclone we followed the same fitting strategy but (1) considered only the signatures that were present in the corresponding sample and (2) removed the final step of adding SBS9 in M-CLL to avoid its addition in multiple subclones with low evidence.

Genomic locations and strand bias. We assessed the contribution of SBS-RT to coding SNVs in RT subclones (also including cases in which the CLL sample was used as a 'germline') by calculating the probability that a given mutation was caused by SBS-RT. To perform this calculation, we considered the signatures present in the subclone/sample and their signature profile⁴⁵. The reference epigenomes of CLL⁵³ were used to explore the contribution of the mutational processes in different regulatory regions. We simplified the described chromatin states in four categories: heterochromatin (H3K9me3_Repressed, Heterochromatin Low_Signal), polycomb

(Posed_Promoter, H3K27me3_Repressed), enhancer/promoter (Active_Promoter, Strong_Enhancer1, Weak_Promoter, Weak_Enhancer, Strong_Enhancer) and transcription (Transcription_Transition, Weak_Transcription, Transcription_Elongation). We also mapped the activity of mutational processes in early/late replication regions of the genome considering peaks/valleys of early/late replication as those regions of $\geq 1\text{kb}$ with absolute replication timing >0.5 (ref.⁵⁴). All SNVs of the CLL and RT subclones were classified in any of the four chromatin states and early/late replication regions before fitting mutational signatures. A cutoff of 0.05 was used to remove the less-contributing signature during the fitting step. We also generated replication and transcriptional strand bias profiles of the RT-specific mutations using the MutationalPatterns R package⁴⁵. The replication strand was annotated based on the left/right replication direction of the timing transition regions⁵⁵. The transcriptional strand was annotated using the TxDb.Hsapiens.UCSC.hg19.knownGene R package (v3.2.2). Finally, kataegis was defined as a genomic region having six or more mutations with an average inter-mutation distance $\leq 1\text{kb}$.

High-coverage, UMI-based gene mutation analysis. Data generation. A high-coverage, UMI-based NGS was performed to track 77 mutations identified by WGS (Supplementary Table 18). Molecular-barcoded and target-enriched libraries were prepared using a Custom CleanPlex UMI NGS Panel (Paragon Genomics) and CleanPlex Unique Dual-Indexed PCR Primers for Illumina (Paragon Genomics, cat. no. 716011 and 716013). Libraries were sequenced on a MiSeq and/or NextSeq 2000 instrument (2 × 150 bp, Illumina).

Data analysis. Raw reads were trimmed using cutadapt (<https://cutadapt.readthedocs.io>; v1.15 with parameters: `-g CCTACGACGAGCTCTCCGATCT -a AGATCGGAAGAGCACACGTCTGAA -a AGATCGGAAGAGCGCTGTGTA GG -G TTCGACGTGCTCTCCGATCT -e 0.1 -O 9 -m 20 -n 2`). Trimmed FASTQ reads were converted to unpaired BAM using Picard's FastqToSam tool (v2.10.2). UMI information was extracted and stored as a tag using igbio ExtractUMIsFromBam (<http://fulcrumgenomics.github.io/igbio/>; v1.3.0 with parameters: `--read structure = 16M+T 16M+T, --single-tag = RX, --molecular-index-tags = ZA ZB`). Template read was converted to FASTQ with Picard's SamToFastq; Template reads were mapped against the human reference genome (GRCh37) and reads were merged with the UMI information using Picard's MergeBamAlignment. Finally, reads were grouped by UMI and a consensus was called using igbio GroupReadsByUmi (parameters were: `--strategy = adjacency, --edits = 1, --min-map = 10`) and CallMolecularConsensusReads (parameters were: `--min-reads = 3`), respectively. A minimum of three reads was required to create a UMI-based final read. Final reads were converted back to FASTQ using Picard's SamToFastq and mapped against the reference genome using BWA-MEM (v0.7.15)⁵⁶. Mean coverage was determined using Picard's CollectTargetedPeMetrics (parameters: `CLIP_OVERLAPPING_READS = true, MINIMUM_MAPPING_QUALITY = 15 MINIMUM_BASE_QUALITY = 15`). Read counts were collected at all targeted genomic positions for all samples using bcftools mpileup (v1.8, parameters: `-B -Q 13 -q 10 -d 100,000 -a FORMAT/DP,FORMAT/AD,FORMAT/ADE,FORMAT/ADR -O v`)⁵⁷. Allele positions lacking mutations by WGS were used to model the background sequencing noise, which was unified according to the trinucleotide context of each possible mutation. Mutations of interest were annotated as high confidence when their frequency was above the background noise with a probability of 95%.

High-coverage immunoglobulin gene characterization. DNA-based. The LymphoTrack IGHV Leader Somatic Hypermutation Assay Panel, MiSeq (Invivoscribe Technologies, cat. no. 71210069) was performed in samples of two patients (Supplementary Table 21). Libraries were sequenced on a MiSeq instrument (2 × 301 bp, Illumina). Clonotypes were defined as IGHV-IGHD-IGHJ gene rearrangements with the same IGHV gene and IGH CDR3 amino acid sequence within a sample. Clonotypes with different nucleotide substitutions within the FRI-CDR1-FR2-CDR2-FR3 sequence of the rearranged IGHV gene were defined as subclones. Raw FASTQ files were trimmed using Trimmomatic (v0.36)⁵⁸ to keep only high-quality reads and bases (parameters were: `LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:100`). Trimmed, paired-end FASTQ files were analyzed using the LymphoTrack Software, MiSeq (v2.3.1, Invivoscribe Technologies, cat. no. 75000099), which combines forward and reverse reads to generate full-length sequences. Identical full-length sequences were grouped and reported together with their cumulative frequency. The reported full-length sequences were annotated using IMGIT/HighV-QUEST (v1.8.3; <https://www.imgt.org/HighV-QUEST>). Finally, we (1) selected the sequences that belonged to the dominant productive clonotype; (2) kept only sequences with complete V-region (missing bases and indels within the V-region were not allowed); and (3) merged sequences that shared the exact V-region nucleotide sequence.

RNA-based. For patient 12, cryopreserved samples collected at four different time points were thawed and malignant cells were enriched using the EasySep Human B Cell Enrichment kit II without CD43 depletion (Stemcell Technologies, cat. no. 17923). Next, 1–2 million tumor cells were used to perform the Omniscope BCR VDJ sequencing assay (<https://www.omniscope.ai>). Cells

were lysed and the RNA was reverse transcribed to complementary DNA with UMIs before amplification of the V(D)J region using BCR-specific multiplex PCR. Following sequencing, reads were aligned using STARsolo (v2.7.9a; <https://github.com/alexdelin/STAR/bbb/master/docs/STARsolo.md>) to the hg38 human genome. IGV³³ was used to review and quantify the mutation of interest (chr14:106714886C>T).

DNA methylation. *Data generation and processing.* DNA methylation data of 39 samples was generated using EPIC BeadChips (Illumina). These samples included different healthy B cell subpopulations (naive B cells (NBCs), $n=2$; germinal center B cells (GCs), $n=1$; memory B cells (MBCs), $n=3$; tonsillar plasma cells (tPCs), $n=1$); CLL samples without evidence of RT ($n=12$) and longitudinal CLL/RT samples ($n=20$) (Supplementary Table 6). Rand core Bioconductor packages, including minfi (v1.34.0)³⁴, were used to integrate and normalize DNA methylation data³⁵. We removed non-CpG probes, CpGs representing single nucleotide polymorphisms, CpGs with individual-specific methylation previously reported in B cells, CpGs in sex chromosomes and CpGs with a detection P value >0.01 in $>10\%$ of the samples. The data were normalized using the SWAN algorithm and CpGs were annotated using the IlluminaHumanMethylationEPICanno.lml1084.hg19 package (v0.6). Tumor cell content of each sample was inferred from DNA methylation³⁶ and samples with a tumor cell content $<60\%$ were excluded. After all filtering criteria, we retained 33 samples (NBCs, $n=2$; GCs, $n=1$; MBCs, $n=3$; tPCs, $n=1$; CLL controls, $n=12$; CLL/RT samples, $n=14$ (six patients); Supplementary Table 6).

Differential analyses, CLL epitopes and epICMIT. We compared the DNA methylation status of each CpG to the mean of each CpG in NBCs to calculate the number of hyper- and hypomethylation changes per CLL/RT sample. Changes in each sample were defined based on a minimum difference of 0.25 methylation. To perform a differential analysis between CLL and RT, we compared the DNA methylation of each CpG in each CLL sample (first available time point used) versus their respective RT sample. Differentially methylated CpGs were considered as those showing a minimum difference of 0.25 in at least four of the five longitudinal cases of RT versus CLL analyzed (Supplementary Table 6). The epigenetic subtypes (epitopes) and epICMIT score for each CLL and RT sample were calculated³⁷.

ChIP-seq of H3K27ac and ATAC-seq. *Data generation.* ChIP-seq of H3K27ac and ATAC-seq data were generated as described in <http://www.blueprint-epigenome.eu/index.cfm?p=7BF8A4B6-F4FE-861A-2AD57A08D63D0B58> (antibody anti H3K27ac, Diagenode, cat. no. C15410196/pAb-196-050, lot A1723-0041D; Supplementary Tables 7 and 8). Libraries were sequenced on Illumina machines aiming at 60 million reads/sample (Supplementary Tables 7 and 8).

Read mapping and initial data processing. FASTQ files were aligned to the reference genome (GRCh38) using BWA ALN (v0.7.7, parameter: $-q\ 5^*$), duplicated reads were marked using Picard tools (v2.8.1) and low-quality and duplicated reads were removed using SAMtools (v1.3.1, parameters: $-b\ -F\ -q\ 5\ -b\ -F\ 1,024^*$). PhantomPeakQualTools (v1.1.0) were used to generate wiggle plots and for extracting the predominant insert-size. Peaks were called using MACS2 (v2.1.1.20160309, parameters for H3K27ac: $-g\ hs\ -q\ 0.05\ -keep\ -dup\ all\ -nomodel\ -extsize\ insert\ -size\ -shift\ -96\ -extsize\ 200\ no\ input\ control^*$). Peaks with q values $<1 \times 10^{-8}$ were included for downstream analyses. For each mark separately, a set of consensus peaks, including regions within chromosomes 1–22 and present in published healthy B cells³⁸ and CLL samples was generated by merging the locations of the separate peaks per individual sample. For ChIP-seq, the numbers of reads per sample per consensus peak were calculated using the genocov function (bedtools, v2.25.0). For ATAC-seq, the number of Tn5 transposase insertions per sample per consensus peak was calculated by first determining the estimated insertion sites (shifting the start of the first mate 4bp downstream) before using the genocov function. Variance stabilizing transformation (VST) values were calculated for all consensus peaks using DESeq2 (v1.28.1)³⁹, which were then corrected for the consensus SPOT score (the percentage of reads that fall within the consensus peaks) using the ComBat function (sva R package, v3.36.0). To that purpose, the cell condition (tumor and different healthy B cell subtypes) was assigned to each sample and samples were clustered in 20 bins of 5% according to their consensus SPOT score. The bins on the extremes, which contained fewer than five samples, were joined with their neighboring bins to ensure that each bin contained five samples or more. PCA was generated using the corrected VST values of peaks that were present in more than one sample.

Detection of differential epigenetic regions and RT-specific changes. We first determined the regions with stable epigenetic profiles in the healthy B cell counterparts (NBCs and MBCs) by applying a threshold of $s.d. <0.8$ with respect to the mean value. For all these NBC/MBC stable regions, we then calculated the log₂FC between the mean of VST-corrected healthy B cell values and each of the tumor samples. Due to the data distribution variability, we applied slightly different thresholds of log₂FC for each case (Supplementary Tables 7 and 8). To identify

regions changing in RT for each case individually, we selected the regions that presented substantial epigenetic changes as compared to the normal counterpart and to the previous CLL (absolute log₂FC >1). The ATAC-seq RT-specific signature encompassed differential regions common to two or more cases of RT, whereas the H3K27ac RT-specific signature included differential regions common in three or more cases. Potential protein-coding target genes were assigned to each of the RT-specific regions using two strategies. To identify close target genes, we took the overlap with the regions of genes of interest adding 2 kb upstream of their transcription start site. To identify distant target genes, we used Hi-C data from the GM12878 cell line and selected all genes located within the same topologically associated domain as the region of interest. We only considered DEGs identified by bulk RNA-seq (Supplementary Tables 7 and 8).

Transcription factor analysis. Enrichment for TF-binding sites was analyzed in chromatin accessible regions within the RT-specific active chromatin regions. Accessible peaks were determined as regions with presence of ATAC peaks in two or more RT cases. Enrichment analysis of known TF-binding motifs was performed using the AME tool (MEME suite) considering the non-redundant *Homo sapiens* 2020 Jasp database and applying one-tailed Wilcoxon rank-sum tests with the maximum score of the sequence, a 0.01 FDR cutoff and a background formed by reference GRCh38 sequences extracted from the consensus ATAC-seq peaks (91,671 regions). We then established the occupancy of these motifs in RT and CLL by calculating the percentage of the target RT-specific active regions and of the regions with increased H3K27ac in CLL, respectively, which contained these motifs. Finally, we selected TFs presenting an occupancy difference between RT and CLL $\geq 10\%$ and overexpressed in RT (bulk RNA-seq, log₂FC >0 , adjusted P value <0.01).

Bulk RNA-seq. *Data generation.* Bulk RNA-seq data of six patients with paired CLL and RT samples were analyzed. Libraries were prepared using the TruSeq Stranded mRNA Library Prep kit (Illumina, cat. no. 20020595) or the Stranded mRNA Library Prep, Ligation kit (Illumina, cat. no. 20040534) and sequenced on a HiSeq 4000 (2 \times 76bp, Illumina) or NextSeq 2000 (2 \times 100bp, Illumina). All samples had a tumor purity $\geq 92\%$ as assessed by flow cytometry (Supplementary Table 11).

Data analysis. Ribosomal RNA reads were filter out using SortMeRNA (v3.2.2)⁴⁰. Non-ribosomal reads were trimmed using Trimmomatic (v0.38)⁴¹. Gene-level counts (GRCh38.p13, Ensembl release 100) were calculated using kallisto (v0.46.1)⁴² and tximport (v1.14.2). A paired DEA was conducted using DESeq2 (v1.26.0)³⁹. Adjusted P value <0.01 and absolute log₂(fold change) >1 were used to identify DEGs. Gene set enrichment analysis (GSEA) was conducted using a pre-ranked gene list ordered by $-\log_{10}(P) \times (\text{sign of fold change})$ using the 'GSEA' function (clusterProfiler R package, v3.14.3). We focused on G2 (curated) and Hallmark gene sets from the Molecular Signatures Database (v7.4) with a minimal size of 10 and maximal size of 250. Gene ontology (GO) GSEA was conducted using the pre-ranked gene list as input of the 'gseGO' function (clusterProfiler) focusing on biological processes. Redundancy in the output list of GO terms was removed using the 'simplify' function (cutoff of 0.35).

Single-cell DNA-seq. *Data generation.* scDNA-seq was performed for 16 samples of 4 patients using the Tapestry Platform (Mission Bio, cat. no. 191335) and a commercial 32-gene panel (Tapestry single-cell DNA CLL panel, Mission Bio, cat. no. MB53-0011_J01). Cryopreserved cells were thawed on 5 ml of fetal bovine serum (FBS; Fisher Scientific, cat. no. 10082147) and incubated at 37 °C for 5 min. Then, cells were washed twice with 1 ml phosphate buffered saline (PBS; Thermo Fisher, cat. no. 20012-019) with 4% bovine serum albumin (BSA; Miltenyi Biotec, cat. no. 130-091-376) and centrifuged at 400g for 4 min. Cell concentration and viability were verified by counting with a TC20T Automated Cell Counter (Bio-Rad Laboratories, cat. no. 1450102). After a final centrifugation step, supernatant was removed and cells were resuspended in an appropriate volume of Mission Bio cell buffer to obtain a final cell density of 3,000–4,000 cells μl^{-1} . Encapsulation, lysis and barcoding of cells were performed following the exact manufacturer's instructions. Afterwards, PCR products were digested and cleaned up with AMPure XP Reagent (Beckman Coulter, cat. no. 100-265-900), followed by quantification of PCR products using a High-Sensitivity dsDNA 1x Qubit kit (Qubit, Invitrogen, cat. no. Q32851). Final library preparation consisted of a Target Library PCR with the V2 Index Primer for ten cycles and a library cleanup with AMPure XP Reagent (Beckman Coulter). Quality control and final quantification were performed on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies, cat. no. 5067-4626). Libraries were sequenced on a NovaSeq 6000 instrument (Illumina) aiming for 1,300 reads per cell (Supplementary Table 20).

Data analysis. FASTQ files were analyzed through the Tapestry Pipeline (v1, Mission Bio), which trims adaptor sequences, aligns reads to the human genome (hg19) using BWA aligner, performs barcode correction, assigns sequence reads to cell barcodes and performs genotype calling using GATK (v3.7). Loom files generated were analyzed using the Tapestry Insights (v2.2, Mission Bio). For each patient (considering all time points together), genotypes with quality <30 , read depth <10 or allele frequency $<20\%$ were marked as missing. Similarly, for each

patient, variants genotyped in <50% of the cells or mutated in <1% of the cells were removed. Cells with <50% of genotypes present were removed. Mutations identified in bulk WGS analysis were used as a whitelist. A list of variants not identified in COSMIC and present at low frequency (1–10% of cells) in all samples analyzed by scDNA-seq was used to remove potential artifacts. The analysis was restricted to coding and splice-site mutations. Genotypes of the selected mutations were exported from Tapestry Insights and used as input of ooSCITE (<https://github.com/cbg-ethz/ooSCITE>). Genotypes were encoded as zero for wild-type, one for heterozygous mutation, two for homozygous mutation and three for missing data. ooSCITE was used to find the mutation tree that best fitted the genotypes observed and to assign cells into subclones. ooSCITE was run using a global sequencing error rate (false-positive rate) of 1%, an estimated rate of non-mutated sites called as homozygous mutations of 0% and a patient-specific estimated rate of the allele dropout rate (false-negative rate). For each patient, the estimated rate of missed heterozygous mutations (dropout of the mutated allele) and the estimated rate of heterozygous mutations called as homozygous mutations (dropout of the normal allele) were calculated from germline single-nucleotide polymorphisms reported in gnomAD with a population frequency > 1% and called as mutated in at least 75% of cells with a VAF per read count between 47% and 53% according to Tapestry Insights. Patient-specific allele dropout rates were calculated for all patients except for patient 365, which did not have any heterozygous polymorphisms fulfilling the previous criteria. In this case, we used an allele dropout rate of 0.07, which is within the range measured in the other cases. We ran ooSCITE with and without considering *NOTCH1* mutations and manually curated the result of patient 3,299 carrying an *RPS15* mutation due to the high allele dropout rate observed in these genes (Supplementary Fig. 2). We ran ooSCITE for each patient combining all time points and obtained time-point-specific subclone sizes by counting the cells assigned to each subclone in each sample¹⁰. Only cells uniquely assigned to one subclone were considered. Cells genotyped as wild-type for all selected mutations were considered as non-tumoral cells and were removed.

Single-cell RNA-seq. *Data generation.* scRNA-seq was performed on longitudinal samples of five patients using three different approaches:

1. Smart-seq2: full-length scRNA-seq libraries were prepared for samples of patient 63 using the Smart-seq2 protocol¹⁰ with minor modifications. Single cells were sorted into 96-well plates containing the lysis buffer (0.2% Triton-100, 1 μ l⁻¹ RNase inhibitor; Applied Biosystems, cat. no. N8080119). Reverse transcription was performed using SuperScript II (Thermo Fisher Scientific, cat. no. 18084014) in the presence of 1 μ M oligo-dT₃₀VN (IDT, cat. no. 22859789), 1 μ M template-switching oligonucleotides (QIAGEN, cat. no. PER-YC00075516) and 1 M betaine (Merck, cat. no. W423212-5KG-K). cDNA was amplified using the KAPA HiFi Hotstart ReadyMix (Kapa Biosystems, cat. no. 7958935001) and IS PCR primer (IDT, cat. no. 22859789), with 25 cycles of amplification. Following purification with Agencourt Ampure XP beads (Beckmann Coulter), product size distribution and quantity were assessed on a Bioanalyzer using a High Sensitivity DNA kit (Agilent Technologies). A total of 140 pg of the amplified cDNA was fragmented using Nextera XT (Illumina, cat. no. FC-131-1096) and amplified with Nextera XT indexes (Illumina, cat. no. 20027215). Products of each well of the 96-well plate were pooled and purified twice with Agencourt Ampure XP beads (Beckmann Coulter). Pooled sequencing was performed on a HiSeq 4000 (2x75bp, Illumina) to an average depth of 0.5 million reads per cell.
2. Cell hashing experiment and 10x Genomics: For each patient (12, 19, 365 and 3,299, experiment BCLL.ATLAS_10), samples obtained at different time points of the disease were labeled following a cell hashing protocol¹⁰. For each sample, 1–2 million cells were resuspended in 100 μ l of cell staining buffer (BioLegend, cat. no. 420201) and incubated for 10 min at 4°C with 5 μ l of Human TruStain FcX Fc Blocking reagent (BioLegend, cat. no. 422302). Next, a specific TotalSeq-A antibody-oligo conjugate (BioLegend, TotalSeq-A anti-human Hashtag 1–8, cat. no. 394601, 394603, 394605, 394607, 394609, 394611, 394613 and 394615) was added and incubated on ice for 30 min. Cells were then washed three times with cold PBS-0.05% BSA and centrifuged for 5 min at 500g at 4°C. Finally, cells were resuspended in an appropriate volume of 1x PBS-0.05% BSA to obtain a final cell concentration of 500–1,000 cells μ l⁻¹, suitable for 10x Genomics scRNA-seq. An equal volume of hashed cell suspension from each of the conditions was mixed and filtered with a 40- μ m strainer (pluriSelect, cat. no. 43-10040-70). Cell concentration was verified by counting with a TC20 Automated Cell Counter (Bio-Rad Laboratories, cat. no. 1450102). Cells were partitioned into Gel Bead In Emulsions with a Target Cell Recovery of 10,000 total cells. Sequencing libraries were prepared using the Chromium Next Gem Single Cell 3' GEM, Library & Gel Bead kit v3.1 (10x Genomics, cat. no. 1000121) with some adaptations for cell hashing, as indicated in TotalSeq-A Antibodies and Cell Hashing with 10x Single Cell 3' Reagent kit v3.1 Protocol by BioLegend. Briefly, 1 μ l of 0.2 μ M HTO primer (IDT, Hashtag Oligonucleotides; CTGACTGGAAGTTCAGACGTGTGCTCTC; *phosphorothioate bond) was added to the cDNA amplification reaction to amplify the hashtag oligonucleotides together with the full-length cDNAs. An SPRI selection cleanup was performed to separate messenger RNA-derived cDNA (>300 bp) from

antibody-oligonucleotide-derived cDNA (<180 bp), as described in the above-mentioned protocol. 10x cDNA sequencing libraries were prepared following 10x Genomics Single Cell 3' v3.1 mRNA kit protocol, whereas HTO cDNAs were indexed by PCR as follows: 5 μ l of purified hashtag oligonucleotide cDNA were mixed with 2.5 μ l of 10 μ M Illumina TruSeq D70X_s primer (IDT) carrying a different 17 index for each sample, 2.5 μ l of SI primer (10x Genomics, cat. no. 2000095), 50 μ l of 2x KAPA HiFi Hotstart ReadyMix (Kapa Biosystems, cat. no. 7958935001) and 40 μ l of nuclease-free water. HTO libraries were purified with 1.2x SPRI bead selection. Size distribution and concentrations of cDNA and HTO libraries were verified on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies, cat. no. 5067-4626). Finally, HTO and cDNA libraries were sequenced on a NovaSeq 6000 (Illumina) to obtain approximately 25,000 reads per cell.

3. Non-cell hashing experiment and 10x Genomics: Samples with a low number of cells in the previous experiment (samples of patient 365 and a subset of samples of patients 12 and 19) were analyzed using a non-cell hashing experiment (BCLL.ATLAS_29). Frozen samples were thawed and 1 ml of 37°C pre-warmed Hibernate-E (Thermo Fisher Scientific, cat. no. A1247601) supplemented with 10% FBS (Thermo Fisher Scientific, cat. no. 10082147) was added drop-wise with gently swirling of the sample. After 1 min of incubation at room temperature, 2,000 μ l of pre-warmed medium was added as mentioned before. Samples were again kept at room temperature for 1 min and 5,000 μ l pre-warmed medium was gently added. This step was conducted twice. Afterwards, samples were centrifuged at 500g for 5 min. Supernatant was removed and pellets were resuspended in 500 μ l 1x PBS supplemented with 0.05% BSA and stained with 4,6-diamidino-2-phenylindole (DAPI) (Thermo Fisher Scientific, cat. no. D1306) at 1 μ M final concentration. DAPI-negative live individual cells were sorted with a BD FACSAria Fusion Flow cytometer (BD Biosciences) in 1x PBS supplemented with 0.05% BSA. After FACS, cells were partitioned into Gel Bead In Emulsions by using the Chromium Controller system (10x Genomics, cat. no. 1000204) aiming at a Target Cell Recovery of 5,000 total cells. Sequencing libraries were prepared using the v3.1 single-cell 3' mRNA kit (10x Genomics). After GEM-RT cleanup, cDNAs were amplified during 14 cycles. cDNA quality control and quantification were performed on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies). Libraries were indexed by PCR using the Chromium7 Sample Index Plate (10x Genomics, cat. no. 220103). Size distribution and concentration were verified on an Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies, cat. no. 5067-4626). Finally, libraries were sequenced on a NovaSeq 6000 sequencer aiming for 40,000 reads per cell.

Read alignment. Raw reads were aligned to the GRCh38 human genome with Cell Ranger (v4.0.0), with the 'chemistry' parameter set to 'SCPv3' and the 'expect-cells' parameter set to 20,000 and 5,000 for cell-hashed and non-hashed libraries, respectively. The remaining parameters for cell-hashed libraries were specified as described in the 'Feature Barcode Analysis' pipeline of Cell Ranger. For Smart-seq2 libraries, alignment and quantification was performed using zUMIs (v9.9c)¹¹.

Demultiplexing of hashing oligonucleotides. Expression matrices were imported into R (v4.0.4) with the 'Read10X' function from Seurat (v4.0.3)¹². HTO counts were normalized with a centered log-ratio transformation applied across features. Each cell barcode was assigned to a specific time point of the disease with the function 'HTODemux' (positive.quantile = 0.99) of Seurat. Barcodes that were positive for two or more time points were labeled as doublets and discarded. Likewise, cell barcodes negative for all time points were excluded. Finally, Scrublet (v0.2.1)¹³ was run to aid in the detection of doublets.

Quality control, normalization and dimensionality reduction. Cells that possessed <90 UMIs, <250 expressed genes or a mitochondrial expression >2.5% were considered as poor quality and removed. Similarly, genes expressed in three or fewer cells were filtered out. Following data normalization and correction (Seurat and NormalizedData), we performed PCA (Seurat, RunPCA) using the scaled expression (Seurat and ScaleData) of the top 2,000 highly variable genes (Seurat: FindVariableFeatures, selection.method = VST). For Smart-seq2 data, we filtered out cells with <150,000 counts, <550 expressed genes or mitochondrial expression >18%. Cells with more than 700,000 counts or 3,750 detected genes were excluded. Similarly, genes expressed in three or fewer cells were filtered out. To separate neoplastic cells from the microenvironment, we corrected the top 30 principal components (PCs) for sample-specific variation using Harmony (v1.0)¹⁴, as implemented in the RunHarmony (group.by.vars = sample) function (SeuratWrappers package, v0.3.0). Subsequently, these 30 corrected PCs were used to embed cells in a UMAP (Seurat, RunUMAP) and in a 20-nearest neighbors graph (Seurat, FindNeighbors) for visualization and clustering, respectively. Following Louvain clustering (Seurat, FindClusters, resolution = 0.1), we focused our downstream analyses only on tumor B cells (CD79A) due to the low number of microenvironment cells.

Dealing with confounders. We observed batch effects between 10x Genomics experiments. To avoid batch effects within samples of the same patient, we focused

on the BCL2L1ATLAS_10 experiment for patients 12, 19 and 3,299. Conversely, as we did not obtain a clear signal-to-noise separation in the HTO demultiplexing of case 365, we analyzed the cells obtained with BCL2L1ATLAS_29. We also found some cell neighborhoods that harbored a high percentage of mitochondrial expression and a low number of detected genes. In such cases, we were more stringent with the thresholds or fetched and eliminated these clusters with FindClusters. We also excluded some clusters of doublets that expressed markers of microenvironment cells (erythroblasts, T cells or natural killer cells). Finally, for patient 3,299 in which one sample was obtained from peripheral blood (PB), whereas the others were obtained from bone marrow (BM), we focused solely on the BM samples to avoid misinterpretations. For patient 365, the CLL and RT time points were sampled from PB and lymph nodes, respectively. As the same RT sample profiled with bulk RNA-seq clustered with other RT samples from PB, we analyzed them jointly. After all the filtering, we recomputed the highly variable genes and PCAs. To avoid overcorrection, we used the top 20 PCs as input to RunUMAP and FindNeighbors, without rerunning Harmony.

Clustering and annotation. Louvain clustering was performed with the FindClusters function, adjusting the resolution parameter for each patient independently. To annotate each cluster, we ran a 'one-versus-all' DEA for each cluster (Seurat, FindAllMarkers, Wilcoxon rank-sum test), keeping only upregulated genes with a $\log_2FC > 0.3$ and a Bonferroni-adjusted P value < 0.001 . If markers were specific to a subset of the cluster, we further stratified it with the FindSubCluster function. On the contrary, if two clusters possessed similar markers, we merged them. The CellCycleScoring function was used to identify clusters of cycling cells.

DEA and GSEA. We conducted a DEA between RT and CLL clusters of each patient independently, merging cells from all time points (Seurat, FindMarkers, \log_2FC threshold = 0, only_pos = FALSE, Wilcoxon rank-sum test). To find finer-grained gene expression changes, only nonproliferative clusters were considered. Genes with a Bonferroni-adjusted P value < 0.05 were considered as significant. The resulting list of genes (sorted by decreasing \log_2FC) was used as input to the gseGO function of clusterProfiler (v.3.18.1, parameters: ont = 'BP', OrgDb = org.Hs.eg.db, keyType = 'SYMBOL', minGSSize = 10, maxGSSize = 250, seed = TRUE). We then removed redundancy in the output list of GO terms with the 'simplify' function (cutoff of 0.75) and filtered out GO terms with an adjusted P value < 0.05 . To convert the expression of specific GO terms of interest into a cell-specific score, we utilized the AddModuleScore function from Seurat.

CNA inference from scRNA-seq data. For each patient separately, we ran inferCNV (v.1.11.1) integrating all samples together. We used CLL cells as reference because (1) we aimed to identify CNAs acquired at RT and (2) CLL had flat copy number profiles in virtually all chromosomes according to WGS. CLL cells were downsampled to the number of RT cells. We initialized an 'infercnv' object (CreateInfercnvObject) using the raw expression counts and the gene-ordering file https://data.broadinstitute.org/Trinity/CAT/cnv/genodecode_v21_gen_pos.complete.txt. CNAs were predicted (infercnv, run, HMM = FALSE, denoise = FALSE) setting the cutoff parameter to 1 and 0.1 for Smart-seq2 and 10x data, respectively. We customized the plotting with the plot_cnv function.

Analysis of an external scRNA-seq dataset. We downloaded the expression matrices and metadata of the dataset from Penter et al.¹³ with the GEOquery (v.2.62.2) (Gene Expression Omnibus identifier GSE165087), created a single Seurat object with all cells from all samples and filtered poor-quality cells as specified in the original publication¹³. Dimensionality reduction, DEA, GSEA and gene signature scoring were performed as described above.

Cellular respiration. Cryopreserved cells were resuspended on RPMI-1640 (Gibco, cat. no. 21875034) with 10% FBS (Gibco, cat. no. 10270-106) and 1% Glutamax (Gibco, cat. no. 35050-061) at a concentration of 3 million cells ml⁻¹. After 1 h of incubation at 37°C, cellular respiration was performed using O₂-consumers (Oroboros Instruments). Two milliliters of cell suspension were added in each respirometer chamber. Cellular respiration was performed at 37°C at a stirrer speed of 750 rpm. Respiratory control was studied by sequential determination of routine respiration (oxygen consumption in living cells resuspended on RPMI-1640 with 10% FBS and 1% Glutamax), oligomycin-inhibited leak respiration (2 μ M ml⁻¹, Sigma-Aldrich, cat. no. O4876, CAS, 1404-19-9), uncoupler-stimulated ETC measured by the sequential titration of the ionophore carbonyl cyanide *m*-chlorophenyl hydrazone (Sigma-Aldrich, cat. no. C2759, CAS, 555-60-2) and residual oxygen consumption after inhibition of the electron transfer system by the addition into the chamber of rotenone (0.5 μ M, Sigma-Aldrich, cat. no. R8875, CAS, 83-79-4) and antimycin A (2.5 μ M, Sigma-Aldrich, cat. no. A8674, CAS, 1397-94-0). Data acquisition and real-time analysis were performed using the software DatLab 7.4 (Oroboros Instruments). Automatic instrumental background corrections were applied for oxygen consumption by the polarographic oxygen sensor and oxygen diffusion into the chamber¹⁶. The same experimental workflow was used to study cellular respiration in CLL and RT cells after 1 h of treatment with LACS-010759 (Selleckchem, cat. no. S8731, CAS, 1570496-34-2) at 100 nM.

Calcium flux analysis. Cryopreserved cells were resuspended on RPMI-1640 medium with 10% FBS, 1% Glutamax and 5% penicillin (10,000 IU ml⁻¹) streptomycin (10 mg ml⁻¹) (Thermo Fisher, cat. no. S8731) at 10⁶ cells ml⁻¹. After 6 h of incubation at 37°C and 5% CO₂, cells were centrifuged and resuspended on RPMI-1640 with 4 μ M Indo-1 AM (Thermo Fisher, cat. no. I1223) and 0.08% Pluronic F-127 (Thermo Fisher, cat. no. P3000MP) for 30 min at 37°C and 5% CO₂. Cells were subsequently labeled for 20 min at room temperature with surface marker antibodies CD19 (Super Bright 600; Invitrogen, cat. no. 63-0198-42) and CD5 (PE-Cy5; BD Biosciences, cat. no. 555354) for the identification of tumoral cells (CD19⁺CD5⁺). Next, cells were resuspended on RPMI-1640 before flow cytometry acquisition. Basal calcium was measured during 1 min before stimulation, then cells were incubated during 2 min at 37°C with or without 10 μ g ml⁻¹ anti-human F(ab')₂ IgM (Southern Biotech, cat. no. 2022-01) and 3.3 mM H₂O (Sigma-Aldrich, cat. no. H1009). Finally, 2 μ M 4-hydroxytamoxifen (4-OHT) (Sigma-Aldrich, cat. no. H6278) was added to all conditions before continue recording for up to 8 min. Intracellular Ca²⁺ release was measured on ISRFortessa (BD Biosciences) using BD FACSDiva software (v.8) by exciting with ultraviolet laser (355 nm) and appropriate filters: Indo-1 violet (450/50 nm) and Indo-1 blue (530/30 nm). Bound (Indo-1 violet) and unbound (Indo-1 blue) ratiometric was calculated with FlowJo software (v.10). Gating analysis was as follows: cell identification in FSC-A versus SSC-A plot, single identification in FSC-A versus FCS-H plot, tumoral cells (CD19⁺CD5⁺) in CD19 (Super Bright 600) versus CD5 (PE-Cy5) plot and Ca²⁺ release in time versus Indo-1 violet/Indo-1 blue plot using a kinetics tool. Optimized dilutions for the antibodies were 1:3 for CD19 and 1:10 for CD5.

Cell growth assays. Cryopreserved cells were resuspended on PBS at a concentration of 10⁶ cells ml⁻¹ and labeled with 0.5 μ M CFSE Cell Tracer (Thermo Fisher, cat. no. C34554) for 10 min. Cells were centrifuged and resuspended on enriched RPMI-1640 medium with 1% Glutamax, 15% FBS, 1 \times insulin-transferrin-selenium (Merk, cat. no. 13146), 10 mM HEPES (Fisher Scientific, cat. no. BP299), 50 μ M 2-mercaptoethanol (Gibco, cat. no. 21985-023), 1 \times Non-Essential Amino Acids (Gibco, cat. no. 11140-050), 1 mM sodium pyruvate (Gibco, cat. no. 11360-070) and 50 μ g ml⁻¹ gentamicin (Gibco, cat. no. 13710-066) at a concentration of 10⁶ cells ml⁻¹ supplemented with 0.2 μ M CpG DNA TL89 ligand (ODN2006-TL9; InvivoGen, cat. no. TLRL-2006) and 15 ng ml⁻¹ recombinant human IL-15 (R&D Systems, cat. no. 247-1L2-025)¹⁶. When indicated, cells were treated for 72 h with 100 nM IACS-010759. Cells were labeled for 20 min at room temperature with surface marker antibodies CD19 (Super Bright 600), CD5 (PE-Cy5) and annexin V (Life Technologies, cat. no. A35122) before acquisition in a ISRFortessa (BD Biosciences) using the BD FACSDiva software (v.8) and analyzed using FlowJo (v.10). Gating analysis for divided cells was as follows: cell identification in FSC-A versus SSC-A plot, single identification in FSC-A versus FCS-H plot, alive cells in annexin V (PacB) versus SSC-A plot, tumoral cells (CD19⁺CD5⁺) in CD19 (Super Bright 600) versus CD5 (PE-Cy5) plot and proliferating cells in the CFSE histogram. Optimized dilutions for the antibodies were 1:3 for CD19, 1:10 for CD5 and 1:3 for annexin V.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data are available from the European Genome-phenome Archive (<http://www.ebi.ac.uk/ega/>) under accession no. EGAS00001006327. scRNA-seq expression matrices, Seurat objects and corresponding metadata are available at Zenodo (<https://doi.org/10.5281/zenodo.6631966>).

Code availability

R markdown notebooks used for mutational signature, bulk RNA-seq, H3K27ac and ATAC-seq analyses can be found at <https://github.com/ferranadeu/>. Richter Transformation, R markdown notebooks to reproduce the scRNA-seq analyses can be accessed at https://github.com/massonix/richter_transformation. Code to normalize DNA methylation data can be found at https://github.com/Duran-FerrerM/DNAmeth_arrays. Code to calculate the tumor cell content, CLL epiotypes and epiCMT from DNA methylation data can be found at <https://github.com/Duran-FerrerM/Pan-B-cell-methylome>.

References

- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Nadeu, F. et al. Ig-Celler for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* **11**, 3390 (2020).
- Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
- Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1–15.10.18 (2016).

65. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
66. Fan, Y. et al. 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 (2016).
67. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* <https://doi.org/10.1093/gigascience/giab008> (2021).
68. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).
69. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinforma.* **52**, 15.7.1–12 (2015).
70. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
71. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
72. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly. (Austin)*. **6**, 80–92 (2012).
73. Nadeu, F. et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia* **32**, 645–653 (2018).
74. 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).
75. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
76. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
77. Müller, E. et al. OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget* **7**, 79485–79493 (2016).
78. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
79. Raine, K. M. et al. ascATngs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinforma.* **56**, 15.9.1–15.9.17 (2016).
80. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
81. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
82. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, 1333–1339 (2012).
83. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
84. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromotripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
85. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
86. Shen, M. M. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell* **23**, 567–569 (2013).
87. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
88. Maura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).
89. Drentos, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
90. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
91. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
92. Yang, F. et al. Chemotherapy and mismatch repair deficiency cooperate to fuel TP53 mutagenesis and ALL relapse. *Nat. Cancer* **2**, 819–834 (2021).
93. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
94. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
95. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
96. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
97. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
98. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
99. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
100. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
101. Kuipers, J., Jahn, K., Raphael, B. J. & Beerewinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27**, 1885–1894 (2017).
102. Morita, K. et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 5327 (2020).
103. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
104. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
105. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., Hellmann, I. & UMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* <https://doi.org/10.1093/gigascience/gyy059> (2018).
106. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
107. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
108. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
109. Gnaiger, E., Steinelchner-Maran, R., Méndez, G., Eberl, T. & Margreiter, R. Control of mitochondrial and cellular respiration by oxygen. *J. Bioenerg. Biomembr.* **27**, 583–596 (1995).
110. Mongini, P. K. A. et al. TLR-9 and IL-15 synergy promotes the in vitro clonal expansion of chronic lymphocytic leukemia B cells. *J. Immunol.* **195**, 901–923 (2015).

Acknowledgements

The authors thank the Hematopathology Collection registered at the Biobank of Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) and the Biobank HUB-ICO-IDIBELL (PT20/00171) for sample procurement, S. Martín, F. Arenas, the Genomics Core Facility of the IDIBAPS, CNAG Sequencing Unit, Mission Bio, Omniscope and Barcelona Supercomputing Center for the technical support and the computer resources at MareNostrum4 (RES activity, BCV-2018-3-0001). This study was supported by the la Caixa Foundation (CL/Evolucion-LCF/PR/HR/17/521500/17, Health Research 2017 Program HR17-00021, to E.C.), the European Research Council under the European Union's Horizon 2020 Research and Innovation Program (810287, BCLAtlas, to E.C., J.L.M.-S., H.H. and I.G.), the Instituto de Salud Carlos III and the European Regional Development Fund Una Manera de Hacer Europa (PMP15/00007 to E.C. and RTI2018-094584-B-I00 to D.C.), the American Association for Cancer Research (2021 AACR-Angen Fellowship in Clinical/Translational Cancer Research, 21-40-11-NADE to E.N.), the European Hematology Association (EHA Junior Research Grant 2021, RG-202012-00245 to E.N.), the Lady Tata Memorial Trust (International Award for Research in Leukemia 2021–2022, IATD, TATA, 21_3223 to E.N.), the Generalitat de Catalunya Support Groups de Recerca AGAUR (2017_SGR-1142 to E.C., 2017_SGR-736 to J.L.M.-S. and 2017_SGR-1099 to D.C.), the Accelerator award CRUK/AIRC/AECC joint funder partnership (AECC-AA17_SUBERO to J.L.M.-S.), the Fundació La Marató de TV3 (201924-30 to J.L.M.-S.), the Centre de Investigació Biomèdica en Red Càncer (CIBERONC; CB16/12/00225, CB16/12/00334, CB16/12/00236), the Ministerio de Ciencia e Innovación (PID2017-11185RB-I00 to X.S.P.), the Fundación Asociación Española Contra el Cáncer (FUNCAR-PRYG2112585UAR to X.S.P.), the Associazione Italiana per la Ricerca sul Cancro Foundation (AIRC 5×1000 no. 21198 to G.C.) and the CERCA Programme/Generalitat de Catalunya. H.P.-A. is a recipient of a predoctoral fellowship from the Spanish Ministry of Science, Innovation and Universities (FPU19/03110). A.D.-N. is supported by the Department of Education of the Basque Government (PRE_2017_1/0100). E.C. is an Academia Researcher of the Institut Catalana de Recerca i Estudis Avançats of the Generalitat de Catalunya. This work was partially developed at the Center Esther Koplowitz (Barcelona, Spain).

Author contributions

E.N. designed the study, collected samples and data, analyzed genomic, immunogenetic and transcriptomic data, interpreted data, designed the figures and wrote the manuscript. R.R. centralized data collection and analyzed and interpreted WGS and bulk RNA-seq data. R.M.-B. analyzed and interpreted scRNA-seq data. H.P.-A. performed and interpreted calcium flux and cell growth experiments and contributed to respiration experiments. B.G.-T. analyzed and interpreted H3K27ac and ATAC-seq data. M.D.-F. analyzed and interpreted DNA methylation data. K.J.D. provided code for the WGS-based subclonal reconstruction and interpreted the results. M.K., A.D.-N., J.L.M., V.C., A.D.-B., S.R.-G., A.G., D.M., N.V.-D., M. Romo, G.C., M. Rozman, G.F. and A.E. performed experiments, analyzed data and/or interpreted data. N.V. conducted flow

cytometry analyses. S.R.-G. provided logistical assistance. J.D., R.M., A.R.-D., T.B., M.A., M.G., F.C., P.A., J.C., F.B., M.A., D.R. and G.G. contributed samples and/or clinical data. A.L.G., P.J., S.B., S.C.-G., J.L.G., N.L.-B., D.T., P.J.C., I.G. and X.S.P. interpreted data. P.M.G.-R. designed, conducted and interpreted respiration experiments. D.C. supervised calcium flux and cell growth experiments and interpreted data. H.H. supervised single-cell experiments and analyses and interpreted data. F.M. contributed to the design and interpretation of WGS analyses. J.L.M.-S. supervised epigenomic experiments and analyses and interpreted data. E.C. designed the study, reviewed pathology, interpreted data, supervised the research and wrote the manuscript. All authors read, commented on and approved the manuscript.

Competing interests

EN. has received honoraria from Janssen and AbbVie for speaking at educational activities. J.L.M. is an employee of Omniscope. X.S.P. is cofounder of and holds an equity stake in DREAMgenics. H.H. is cofounder of Omniscope and consultant to MIRXES. E.C. has been a consultant for Takeda, NanoString, AbbVie and Illumina; has received honoraria from Janssen, EUSPharma and Roche for speaking at educational

activities; and is an inventor on a Lymphoma and Leukemia Molecular Profiling Project patent 'Method for subtyping lymphoma subtypes by means of expression profiling' (PCT/US2014/64161) not related to this project. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41591-022-01927-8>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-022-01927-8>.

Correspondence and requests for materials should be addressed to Ferran Nadeu or Elias Campo.

Peer review information *Nature Medicine* thanks Daniel Hodson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling editor: Anna Maria Ranzoni, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

...

