UNIVERSITAT DE BARCELONA

Development and application of methodologies and infrastructures for cancer genome analysis within Personalized Medicine

Romina Royo Garrido

Development and application of methodologies and infrastructures for cancer genome analysis within Personalized Medicine

Romina Royo Garrido

UNIVERSITAT DE BARCELONA
FACULTAT DE BIOLOGIA
DEPARTAMENT DE BIOQUÍMICA I BIOMEDICINA MOLECULAR

PROGRAMA DE DOCTORAT EN BIOMEDICINA. LINIA DE RECERCA BIOINFORMÀTICA

Development and application of methodologies and infrastructures for cancer genome analysis within Personalized Medicine

Memòria presentada per Romina Royo Garrido per optar al grau de doctora per la Universitat de Barcelona

DIRECTORS DE TESI

David Torrents Arenales Josep Lluís Gelpí Buchaca

DOCTORANDA

Romina Royo Garrido

TUTOR Josep Lluís Gelpí Buchaca

Acknowledgements

A tots amb els que he coincidit durant aquests anys, i els anteriors, moltes gràcies per ser-hi!

Iñaki i Carles gràcies per tot el feedback durant les comissions de seguiment, sempre constructiu i del que he après molt.

Alfonso, Josep, Fátima, Sílvia y Salva, muchas gracias, será un honor teneros como parte del tribunal.

Al tiet, jefe, i director, gràcies per cuidar-nos sempre, encara que de nens ja en tenim ben poc! No oblidarem mai tots els bons records, espero que tu tampoc... no crec que gaire gent li doni una pinyata per trencar al jefe ;-)

Al Torrents, per prioritzar el "bon rotllo" per damunt de tot, i per ensenyar-nos a presentar, escriure... i que les normes es poden trencar, encara que alguns potser no ho farem mai...

És una sort que sempre m'he trobat bons companys, impossible nombrar-los a tots.

Sempre és una alegria passar a saludar els computational genomics, i en especial els que et desitgen suerte en la vida! Ana, tot el que hem fet plegades m'ha encantat, gràcies per ser tan bona persona. I igualment Lorena, Mercè, sou super maques, és un regal trobar-vos a l'oficina.

I igualment a tots els INB. Els de tota la vida... INB Team!! Me n'emporto records molt macos i sempre us tinc presents... there are a lot of rabbits! Un plaer haver compartit tots aquests moments, i haver après el nom d'un poblet (Vilamacolum!) amb una gran persona. Azazelle, cada cosa que fas brillen les

estrelles a tot l'univers. Al que nunca le falta un chiste ruso y siempre se acuerda de las fechas señaladas. I als nous (o no tant nous) companys que sempre és maco veure'ls, sobretot al que em va dir que "m'he de deixar fluir", a la que m'acompanya en la no-socialitat, i, més recentment, al que dóna suport als 21ºC... Y gracias Salva por apadrinarme y por tu confianza. A les pachis, les més millors, moltes gràcies per la vostra amistat.

A tots els supporters del BSC, que com ningú intenteu cuidar a l'usuari, fins i tot quan demanem coses totalment fora de la vostra feina. Al dream-team dels sysadmins, Javi, Smore, Vavavavavalls, Ocaña i Fenoy (als tres últims, us trobem a faltar!). Gràcies Ferran i Pedro per totes les vegades que m'heu ressetejat el password caducat... :-) i al Toni per ser tan dicharachero.

Thanks to the ICGC, PCAWG, and ICGC-ARGO teams. A Xose, con quien hice mi primer alineamiento, siempre fue un placer trabajar contigo.

A l'Elias i el Ferran, sense paraules! Moltes gràcies per fer-me sentir sempre "com a casa" i per tot el que he après amb vosaltres, i no només de CLL. La vostra paciència, perseverança i amabilitat en tot el que feu és un exemple a seguir.

Als fibbers, hugonaueeeer sempre ready per deixar anar una bona xapa... really? You know... i als ixus, el més millor “pet” de tota la història (o.^^)´(“.o) i la més millor persona.

Als pares i la família que tot ho aguanten, i al millor company que encara no sé com l'he pogut trobar, per sort estava de rebaixes i no va poder fer devolución...

Als pacients, amb el desig que puguin viure més i millor, i a les seves famílies.

Abstract

Next-generation sequencing (NGS) has revolutionized biomedical sciences, especially in the area of cancer. It has nourished genomic research with extensive collections of sequenced genomes that are investigated to untangle the molecular bases of disease, as well as to identify potential targets for the design of new treatments. To exploit all this information, several initiatives have emerged worldwide, among which the Pan-Cancer project of the ICGC (International Cancer Genome Consortium) stands out. This project has jointly analyzed thousands of tumor genomes of different cancer types in order to elucidate the molecular bases of the origin and progression of cancer. To accomplish this task, new emerging technologies, including virtualization systems such as virtual machines or software containers, were used and had to be adapted to various computing centers. The portability of this system to the supercomputing infrastructure of the BSC (Barcelona Supercomputing Center) has been carried out during the first phase of the thesis. In parallel, other projects promote the application of genomics discoveries into the clinics. This is the case of MedPerCan, a national initiative to design a pilot project for the implementation of personalized medicine in oncology in Catalonia. In this context, we have centered our efforts on the methodological side, focusing on the detection and characterization of somatic variants in tumors. This step is a challenging action, due to the heterogeneity of the different methods, and an essential part, as it lays at the basis of all downstream analyses.

On top of the methodological section of the thesis, we got into the biological interpretation of the results to study the evolution of chronic lymphocytic leukemia (CLL) in a close collaboration with the group of Dr. Elías Campo from the Hospital Clínic/IDIBAPS. In the first study, we have focused on the Richter transformation (RT), a transformation of CLL into a high-grade lymphoma that

leads to a very poor prognosis and with unmet clinical needs. We found that RT has greater genomic, epigenomic and transcriptomic complexity than CLL. Its genome may reflect the imprint of therapies that the patients received prior to RT, indicating the presence of cells exposed to these mutagenic treatments which later expand giving rise to the clinical manifestation of the disease. Multiple NGS-based techniques, including whole-genome sequencing and single-cell DNA and RNA sequencing, among others, confirmed the pre-existence of cells with the RT characteristics years before their manifestation, up to the time of CLL diagnosis. The transcriptomic profile of RT is remarkably different from that of CLL. Of particular importance is the overexpression of the OXPHOS pathway, which could be used as a therapeutic vulnerability. Finally, in a second study, the analysis of a case of CLL in a young adult, based on whole genome and single-cell sequencing at different times of the disease, revealed that the founder clone of CLL did not present any somatic driver mutations and was characterized by germline variants in *ATM*, suggesting its role in the origin of the disease, and highlighting the possible contribution of germline variants or other non-genetic mechanisms in the initiation of CLL.

Abbreviations and acronyms

1+MG	European '1+Million Genomes' Initiative
AD	Alternate Depth
AID	Activation-induced cytidine deaminase
AML	Acute Myeloid Leukemia
API	Application Programming Interface
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
B1MG	Beyond 1 Million Genomes (H2020 project)
BCL-2	B-cell leukemia/lymphoma 2
BCR	B-cell Receptor Pathway
bp	Base pair
BSC-CNS	Barcelona Supercomputing Center-Centro Nacional de Supercomputación
BTK	Bruton's tyrosine kinase
BTKi	BTK inhibitors
BWA	Burrows-Wheeler Aligner
CCE	Cancer Core Europe
CCF	Cancer Cell Fraction
CDR3	Complementary-determining region 3
CGC	The Cancer Gene Census
CGI	Cancer Genome Interpreter
ChIP-seq	Chromatin immunoprecipitation sequencing
CIT	Chemoimmunotherapy
CIViC	Clinical Interpretation of Variants in Cancer
CKB	Clinical Knowledgebase
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CMP	Cloud Management Platforms
CNA	Copy Number Alteration
CNAG	Centro Nacional de Análisis Genómico
COSMIC	Catalog of Somatic Mutations in Cancer
CRG	Centro de Regulación Genómica
CSR	Class Switch Recombination
CWL	Common Workflow Language

DCC	Data Coordination Center
del(11q)	Deletion of q arm of chromosome 11
del(13q)	Deletion of q arm of chromosome 13
DLBCL	Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic acid
DP	Depth
EATRIS	European infrastructure for translational medicine
EGFR	Epidermal Growth Factor Receptor
EM	Expectation Maximization
EUCANCan	European-Canadian Cancer Network (H2020 project)
ExAC	Exome Aggregation Consortium
FF	Fresh Frozen
FFPE	Formalin-Fixed Paraffin-Embedded
GA4GH	Global Alliance For Genomics and Health
GATK	Genome Analysis Toolkit
GB	Giga byte
GC	Germinal Center
GIAB	Genome In A Bottle Consortium
gnomAD	Genome Aggregation Database
GPFS	General Parallel File System
GWAS	Genome-wide association studies
H-chain	Heavy chain
HDP	Hierarchical Dirichlet Process
HPC	High-performance computing
HSC	Hematopoietic Stem Cell
HTS	High-throughput sequencing
ICGC	International Cancer Genome Consortium
ICGC-ARGO	Acceleration Research in Genomic Oncology
IDIBAPS	Institut d'Investigacions Biomèdiques August Pi i Sunyer
IDIBELL	Institut d'Investigació Biomèdica de Bellvitge
Ig	Immunoglobulin
IgH	Immunoglobulin heavy chain
IGHV	Immunoglobulin heavy-chain variable region
IGV	Integrative Genome Viewer
IL	Interleukin

indel	Short insertion or deletion
ITH	Intratumor heterogeneity
Kb	Kilo base
L-chain	Light chain
LSF	Load Sharing Facility
M-CLL	CLL expressing mutated IGHV
MBL	Monoclonal B-cell Lymphocytosis
MCL	Mantle Cell Lymphoma
MCMC	Markov Chain Monte Carlo
MedPerCan	Medicina Personalitzada Catalunya Cancer (PERIS Project)
MN	MareNostrum
MNP	Multiple Nucleotide Polymorphism
MTBP	Molecular Tumor Board Portal
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear factor-κB
NGS	Next-generation sequencing
NICD	*NOTCH1* intracellular domain
NIST	National Institute of Standards
NMF	Non-negative matrix factorization
OS	Operating System
PB	Peta byte
PCAWG	PanCancer Analysis of Whole Genomes
PCT	Personalized Cancer Therapy
PI3Kδ	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta
PM	Personalized Medicine
QC	Quality Control
RAM	Random Access Memory
RDPC	Regional Data Processing Center
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RS	Richter Syndrome
RT	Richter Transformation
RT-DLBCL	Diffuse large B-cell lymphoma type of RT
RT-PBL	Plasmablastic lymphoma type of RT
RT-PLL	Prolymphocytic leukemia type of RT

SBS	Single Base Substitution
scDNA-seq	Single-cell DNA sequencing
scRNA-seq	Single-cell RNA sequencing
SEQC2	Sequencing Quality Control 2
SHM	Somatic Hypermutation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
STAR	Spliced Transcripts Alignment to a Reference
SV	Structural Variant
TB	Tera byte
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
tri12	Trisomy 12
U-CLL	CLL expressing unmutated IGHV
UMI	Unique Molecule Identifiers
US	United States
VAF	Variant Allele Frequency
VDJ	Variable (V), Diversity (D) and Join (J) regions
VM	Virtual Machine
WDL	Workflow Description Language
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WHO	World Health Organization

Table of contents

1 Introduction

1.1 Thesis trajectory

I would like to start by summarizing the context in which I started this thesis and the trajectory that we followed.

My first contact with genomic research was back in 2011, when the BSC (and myself) participated in the CLLGenome project, which was part of the International Cancer Genome Consortium (ICGC), an international initiative that coordinated worldwide efforts focused on the study of the genomic basis of cancer. There, we analyzed more than 500 genomes of chronic lymphocytic leukemia (CLL) and identified several biomarkers associated with the offset and progression of this tumor (Puente et al., 2011, 2015; Quesada et al., 2012). My role was to manage and prepare the data for their analyses and to execute them in high-performance computing (HPC). Even though my contribution was purely technical, it gave me the opportunity to hear about the first concepts of next-generation sequencing (NGS) analysis, and to foresee the giant wave of genomic data that was coming in the field, together with their computational demands.

Later on, as a natural evolution of these activities, the ICGC launched a new worldwide initiative called Pan-Cancer Analysis of Whole Genomes (PCAWG), where more 2,600 normal-tumor genome pairs, covering 38 cancer types, were analyzed to further elucidate the origin and evolution of cancer (Campbell et al., 2020). The BSC was one of the main data centers of the project, stored around 1PB of data, and performed the analysis of genomic sequences for the search of cancer-related mutations. Within this project, I dealt with the underlying challenges of large-scale genomic projects, where portability and reproducibility are fundamental to carry out distributed efforts among different data centers. Virtualization approaches were starting to gain attention within the field and the use of virtual machines, followed by docker containers, was mandatory to

guarantee these requirements. Data storage and management was a challenge for the project, but, above all, the biggest obstacle for our center was the use of those emerging technologies, which had to be accommodated in our (almost pure) HPC center at the time.

Next, my contribution to other projects addressing the generation of computational environments for the analysis and management of cancer genome data, such as MedPerCan or EUCANCan, gave me a deeper understanding of the methodologies that are used in cancer genomics, which I had previously been executing for many years. Within the MedPerCan project, a national initiative from the Pla de recerca i innovació from the Generalitat de Catalunya, we implemented a pilot circuit among sequencing centers, data analysis centers, and hospitals, to evaluate the impact that genomic analysis can have into the clinics. In that context, I could further explore and evaluate the myriad of programs that can be used to analyze different kinds of NGS data, including the most challenging settings and scenarios in cancer genomics (e.g., the analysis of tumor-only samples that lack the corresponding germline sample, or the analysis of formalin-fixed paraffin-embedded (FFPE) specimens, which are prone to present genomic artifacts). This is where I saw first-hand the vast heterogeneity of quality and scope that exists among the different analysis pipelines across research centers, which require harmonizing and benchmarking environments. In line with this, within the EUCANCan project, a federated network for the harmonized genomic and phenotypic data sharing in oncology, we are pursuing these concerning topics, defining strategies for the harmonization of variant calling results, and devoting efforts to create benchmarking protocols, as well as addressing the legal aspects of genomic and clinical data sharing.

Finally, to fulfill this path, and to follow my growing interest in the biological aspects, I have had the opportunity to apply all these methodologies and

strategies to answer specific biomedical questions about CLL. Here, I was involved in the generation and also in the interpretation of the results, gaining deeper insights into the biological and medical aspects of this tumor. This activity has been done in a close collaboration with Dr. Elías Campo's group at IDIBAPS (Barcelona) and Hospital Clínic, where we have jointly analyzed the genomic and molecular basis of particular and aggressive forms of CLL progression to high-grade lymphomas (Richter transformation).

Overall, during these years, I have covered from the more technical aspects of the cancer genomics field, up to the real application of these methodologies to different studies, which, in turn, lead to new discoveries of the biology of CLL evolution and provided me with a wide and comprehensive view of modern genomic projects in Biomedicine.

1.2 The biology of the genome and its relationship to disease

The biomedical community has long been trying to decipher the molecular basis of disease, mapping the traits to genes, exploring what parts of the genome are controlling the molecular processes in our cells and, above all, elucidating the genetic factors that can have an effect on our health.

1.2.1 Transformation of biomedical research

Biomedicine plays a very important role in modern health care. The roadmap of this field, centered in the understanding of the genetic and biological basis of disease, has evolved over the years, pushed by new biological insights, as well as the emergence of new technologies and methodologies. These new protocols have shifted the standpoint of our research from physical traits to the genomic

alterations potentially responsible for the different phenotypes that we can observe.

Traditionally, the starting point of human genetic studies was the physical manifestation of disease, followed by the identification and characterization of the causal cellular functions that pinpoint specific proteins which, in turn, had to be linked to candidate genes. The advent of new sequencing techniques and the increasing knowledge about our genome has allowed to simultaneously study the genome of hundreds and thousands of patients of a particular trait and identify alterations associated with it. With this new perspective, instead of tackling our biological questions from function to genetics, we can go the other way around: from genomics to function (Figure 1). After the sequencing process, we can detect the modifications in the genome and evaluate their recurrence and functional impact as a means to associate genomic genes or locus to molecular processes driving malignancy.

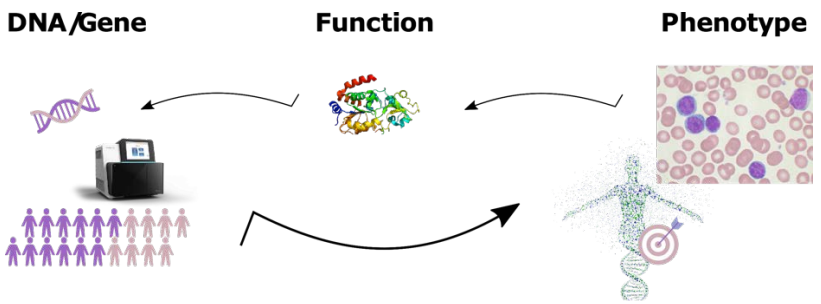


Figure 1. Traditional approach (up) from phenotype to genotype, and current approach (down) from DNA to functional implications.

This new approach comes with a remarkable change of methodologies, starting with the explosion of next-generation sequencing (NGS), and the consequent growth of biological data and bioinformatics applications. The extensive collection of tools is unceasingly increasing in order to analyze the vast

amounts of data generated. This data tsunami has shed light on how our genome works (C. A. Davis et al., 2018; Dunham et al., 2012), how specific alterations can be related to different kinds of diseases, and has also provided thousands of potential diagnosis and treatment markers. Altogether, it has fostered biomedical research, helping us understand the underlying mechanisms leading to disease, and favoring more accurate clinical decisions on diagnosis and treatment options, which corresponds to the ground of Personalized Medicine (PM).

At the same time, massive genome sequencing analyses also open up the need for computational infrastructures capable of meeting demanding resources, both in terms of data management and storage and computational capacities able to run the analyses. These technical challenges and existing means to address them will be described in more detail in section 1.4.3 in the context of cancer research.

1.2.2 Next-generation sequencing to study the genome

After the tremendous effort of the sequencing of the first human genome (Craig Venter et al., 2001; Lander et al., 2001) genomic research has taken a new direction. This endeavor, which starts from the very basics of directly exploring the DNA, has further elucidated our knowledge, and led us towards a better understanding and treatment of disease.

1.2.2.1 Data explosion

The advent of high-throughput sequencing (Bentley et al., 2008), or next-generation sequencing (NGS), has revolutionized genomic research, offering affordable sequencing both in terms of costs and time. It has improved up to the point where, since 2008, Moore's Law stopped being an accurate predictor of sequencing costs (Figure 2).

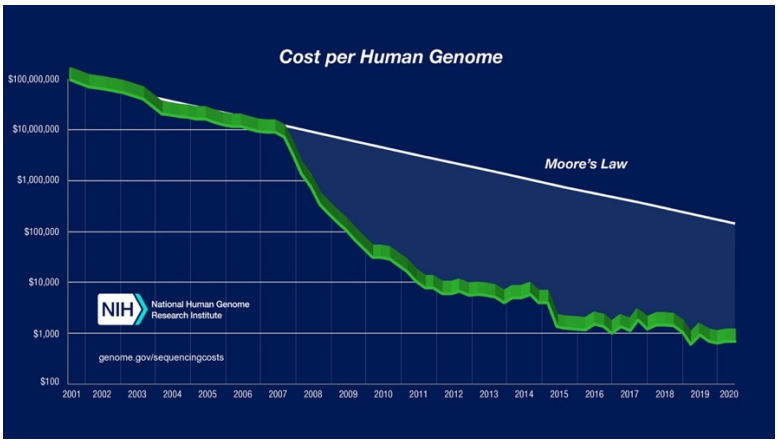


Figure 2. Cost per genome data – August 2020 (genome.gov). The observed drop in 2008 coincides with the wider use of NGS, offering affordable prices and leading to the consequent genome data growth.

NGS produces millions of DNA sequenced fragments, or reads, that can be computationally analyzed. Briefly, DNA samples are prepared as libraries of short fragments of 100-150 base pairs (bp) and sequenced. The output is a collection of short sequence reads (commonly known as "reads"), all mixed up and mapping at unknown genome locations. These sequences can be informatically assembled or aligned to an already known reference genome. Bioinformatics analysis usually proceeds with the identification of genomic variants that can be potentially related to mechanisms and processes behind the development of a disease. DNA sequencing techniques can be used to identify both germline and somatic variants, the latter using a normal reference sample to discern acquired mutations from the inherited genotype.

There are different kinds of assays that can be done, which can give us different and complementary perspectives of the underlying mechanisms leading to malignancy. The three main layers that can be investigated based on NGS are: the genome, the transcriptome, and the epigenome. Genomic methods include

three main techniques that vary in terms of cost and scope. Targeted sequencing of specific locus is the cheapest approach, it can be used to deeply explore or validate small regions of the genome. Whole-exome sequencing (WES) captures all protein-coding regions of the genome. It is cost-effective and it has been widely used to identify genes and variants associated with diseases, especially in the context of Mendelian diseases. Nonetheless, it only allows us to explore 1-2% of our genome. Whole-genome sequencing (WGS) gives us the chance to go further and study the remaining 99% non-coding region of the genome and has the highest cost. WGS provides orders of magnitude more point mutations than exomes, greater resolution to detect copy number alterations (CNAs), and the ability to call structural variants (SVs). Thus, it increases the breadth and depth of our analyses but, at the same time, moving from WES to WGS datasets for large studies increases data sizes up to the petabyte scale. The next layer is the transcriptome analysis (RNA-seq, short for RNA sequencing), which covers gene expression profiling and discovery of non-coding RNA and novel transcribed sequences. Finally, epigenomic studies address the modifications that affect gene expression without altering the DNA sequence, focusing on chromatin changes, and using methods such as DNase-seq, ATAC-seq, DNA methylation and histone modification ChIP-seq. Depending on the scope and budget, projects might include one or several of these techniques.

Faster and cheaper sequencing has fostered the generation of larger and larger collections of samples, growing at an unprecedented scale. This has defied bioinformaticians and computational scientists: the bottleneck is not on data generation anymore; it goes down to data processing. Consequently, strategies to analyze such data must adapt to new datasets that grow in size and number.

1.2.2.2 Bioinformatics applications, infrastructures, and data sharing

High-throughput technologies generate vast amounts of raw data, which puts the need for computational strategies able to manage and analyze them on the spot. As previously explained, the sequencing process starts with the DNA preparation. Briefly, the DNA is fragmented, short oligonucleotides to ligate the ends of DNA fragments of interest to the primers are linked, and all together form the library of fragments, which have a specific fragment size, and that will be sent to sequencing. In the paired-end sequencing strategy, the most common nowadays, the parts targeted for sequencing are located at both ends of the insert, have a selected read length, and are sequenced in opposed orientation. The non-sequenced region between both paired reads corresponds to their inner distance (Figure 3).

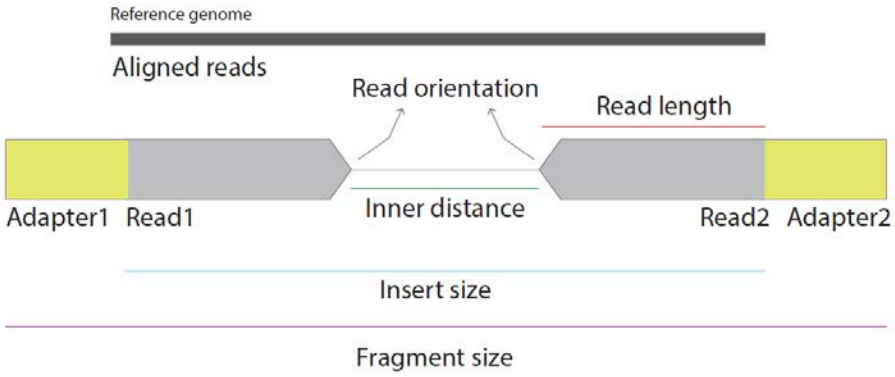


Figure 3. Sequencing and alignment concepts. View of an alignment of a pair-end read.

The result of DNA sequencing is a set of short nucleotide sequences (from 70bp to 150bp) called reads. They are encoded using the alphabet letters A, C, G, and T, to symbolize the nucleobases adenine, cytosine, guanine, and thymine, respectively. At this point, we are far beyond wet labs, and informaticians and computers come into place. The unsorted reads are stored in FASTQ formatted

files that are aligned against a reference genome to find the location of each sequence. The obtained alignments are usually stored on BAM formatted files, a zipped form of the SAM format (H. Li et al., 2009), or CRAM formatted files, which is an even more compressed format. The alignments, in turn, are the inputs of the variant calling phase, a crucial step that can affect all downstream analyses. Variant calling analysis provides a list of sequence variants with respect to a reference genome. This list can be used for multiple analysis, including complex algorithms to infer tumor evolution, mutational processes contributing to cancer, or the most direct question one might ask which is to determine the functional impact of genetic variants. In any case, the answer to the biological questions that we are trying to respond will greatly depend on the variant calling results. Thus, it is of the utmost importance that the detection of variants is as accurate as possible.

Bioinformatics methodologies are becoming essential not only in biomedical research but also in clinical applications. Over the past decade, this field has evolved together with the advances and requirements of the NGS-era, where terabytes and petabytes of data pile up. New bioinformatics methodologies and protocols are essential to manage, analyze, and interpret the continuously growing biological data in modern biomedicine, ultimately enhancing the discovery of new drug targets, and improving patient care. In this sense, biological data has entered into the world of big data analytics (Schadt, 2012) (Figure 4).

Tools, methods, and infrastructures are being developed and adapted to process voluminous datasets and extract their biologically relevant information. As we strengthen our comprehension of the relationship between alterations in our DNA and their functional and clinical impact, we can translate this knowledge into actionable clinical practice. In particular, we are now able to identify the specific variants in the genome of an individual, which is the ground of

personalized or precision medicine, where we can treat patients according to their particular alterations in their DNA.

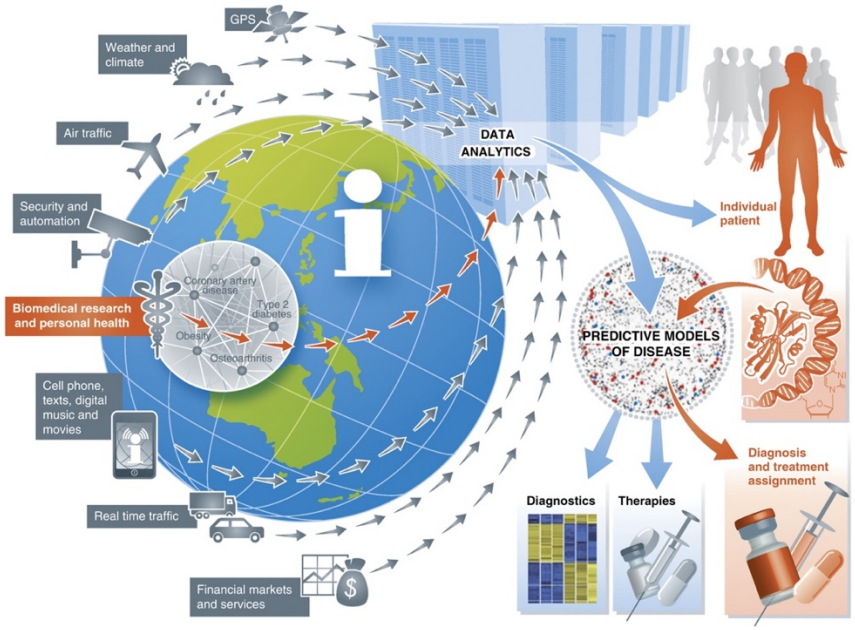


Figure 4. Life and biomedical sciences and the big data revolution. Image from (Schadt, 2012).

As a community, we can advance research faster and gain statistical power by joining together the insights of individual genomes and external datasets. This prompts the necessity to share sensitive data, which heralds barriers at many different levels. Data must be organized, raw data and associated metadata must be easily linked, found, and made available to the researchers and clinicians. In other words, data producers should follow the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson et al., 2016) to contribute to the broader community scientific advances and maximize transparency. Openness does not only apply to data, but also the computer code used to extract the biologically meaningful information. In this sense, the software used to analyze the data, from single tools to complex workflows, should also be

open access (Jiménez et al., 2017). Data sharing can benefit scientific advances at many levels, but despite encouraging initiatives there is still some reluctancy. To move towards this direction and persuade data producers to make their data and code available, many funding agencies and journals are implementing policies to request openness of their work (Popkin, 2019).

1.2.2.3 From bulk to single cell resolution

We have learned a lot from next-generation sequencing studies that taught us the molecular mechanisms and genetic determinants of disease. In the case of cancer, not only particular alterations driving tumorigenesis have been identified, but it has also been seen that tumors are not formed by a monolithic population of malignant cells, but rather multiple subpopulations with particular attributes that, in turn, will define different patterns of evolution. This heterogeneity can be quite complex and is a major problem when imposing treatment pressures.

Conventional bulk sequencing obscures the underlying genetic diversity within cell subpopulations, and signals may not be identified when they are analyzed altogether as a group. However, when they are analyzed individually, they can reveal the variation and differences on a cell-by-cell basis. An analogy often used to compare bulk and single-cell sequencing is that of a smoothie, where all ingredients are mingled, and a bowl of fruit, where each single fruit can be explored separately (Figure 5). Single-cell emerged as the perfect technology to dissect this heterogeneity and was selected Method of the Year 2013 (Eberwine et al., 2013; "Method of the Year 2013," 2013). Sequencing of DNA and RNA of single cells is poised to expand our knowledge of biology and medicine as it becomes more potent and broadly available. Diseases show heterogeneity at the level of individual cells, and single-cell studies can untangle the differences among them and lead to a better understanding of why they might have different drug

responses. Single-cell DNA sequencing (scDNA-seq) can identify somatic mutations in the genomes of individual cancer cells, and this information can be used to assess the subclonal architecture of tumors and to trace the evolution and spread of the disease. At the RNA level, single-cell transcriptome profiling (single-cell RNA sequencing, scRNA-seq) can determine phenotypic differences in cells that are biologically relevant, shedding light on the cellular differences with higher resolution.

Figure 5. Bulk sequencing vs single-cell sequencing. Bulk sequencing resembles a smoothie, where all ingredients are mixed, the same way cells are mingled together in bulk sequencing. On the other hand, single-cell sequencing can provide information on each individual cell, differentiating each particular cell the same way one can characterize different fruits from a bowl.

1.2.3 Genomic variation

The human genome is made of 3.2 billion bases, and it is estimated that 99.9 percent of them are identical across all human beings. The 0.1% difference is responsible for the diversity that makes us unique, and also for the differences among individuals from the point of view of their risk of developing diseases.

1.2.3.1 Germline and somatic variants

Genetic variants can either be inherited or generated and accumulated during our lifetime. Germline mutations, inherited from our parents, are present

in all the cells of our body, and they can be passed out to our children. On the other hand, somatic mutations are acquired later on during our life span and occur initially in a single cell. Thus, they only affect tissues derived from the mutated cell, and are not passed out to offspring.

Everyone is born with a set of genetic variants which constitutes our genotype. Our genetic background can determine many things in our life, and together with environmental factors, can predispose us to different kinds of diseases (Figure 6). There are some variants that can help us prevent disease, while others can be more damaging. Some germline variants do not have strong effects but predispose to disease, which can appear if not-so-healthy choices are added on top. This is the case of complex diseases such as type 2 diabetes. On the other hand, other variants that have strong and deleterious effects will lead to monogenic or rare diseases, regardless of our lifestyle. During our life, new variants, called somatic, can also be acquired independently, due to intrinsic or extrinsic factors, such as tobacco or ultraviolet radiation. Many of them can have a neutral effect, but others can have a deleterious effect leading to the formation of a tumor.

These three types of diseases require three different research approaches. Genome-wide association studies (GWAS) (Uffelmann et al., 2021) are used to identify single nucleotide polymorphisms (SNPs) that are enriched in a subgroup of individuals with a specific phenotype or disease. Large-scale sequencing projects, like the 1000 Genomes Project (1000G) (D. L. Altshuler et al., 2010; D. M. Altshuler et al., 2012; Auton et al., 2015; Sudmant et al., 2015), the GoNL project (Boomsma et al., 2014), or the UK10K project (Walter et al., 2015), generate haplotype maps of specific populations that are needed to achieve the statistical power required in this study of common complex diseases. The study of rare diseases has greatly benefited from NGS, especially the more cost-effective

whole-exome sequencing, that can be used to identify pathogenic variants in coding regions of the genome, which is where we expect high penetrant variants to occur. This complements current clinical protocols that, despite comprehensive clinical evaluation, are unable to find a definitive diagnosis (Worthey et al., 2011), and can thus improve the diagnosis of this type of disease. Finally, the analysis of somatic variants, which is mainly related to cancer, has been one of the areas with more sequencing and analysis of genomes.

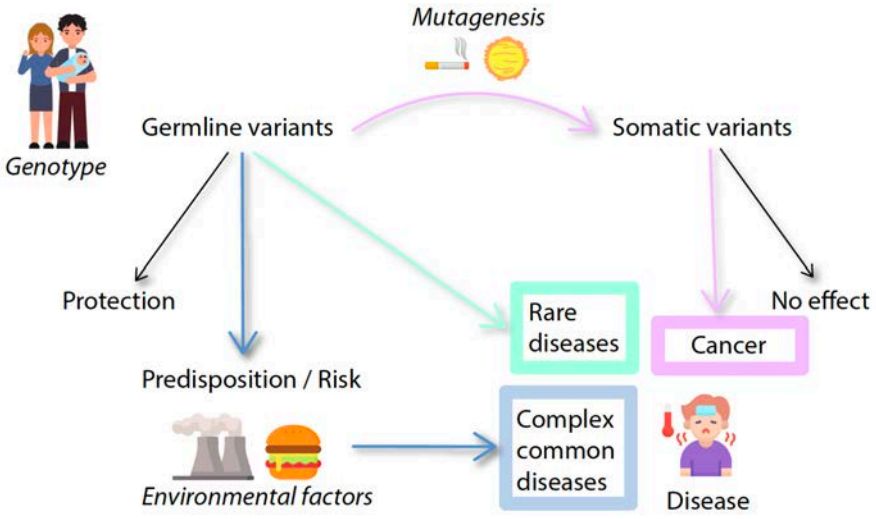


Figure 6. Types of genetic variants: germline variants are inherited from our parents, while somatic variants are acquired during our life. Both types of DNA alterations can confer a higher risk to develop different kinds of diseases: from rare diseases to complex common diseases and cancer, where lifestyle and environmental factors can play an important role.

In cancer research, these studies have demonstrated not only the role of DNA alterations in tumor formation and progression, but also their potential translation and actionability in the clinics. These discoveries led to a new paradigm where cancer researchers identify somatic genetic alterations, drugs targeting those cancer-specific alterations are developed, and patients are managed with treatments targeting their specific DNA alterations. There are many examples of

success; the first one was the use of imatinib to treat patients with chronic myeloid leukemia (CML) harboring a translocation that creates a BCR-ABL fusion kinase, followed by the use of epidermal growth factor receptor (EGFR) inhibitors for lung cancers bearing mutant EGFR, or BRAF inhibitors to treat melanomas bearing mutated BRAF, among others (Letai, 2017).

Overall, with the new aforementioned perspective, diseases can be categorized and treated according to their genomic and molecular basis, in addition to their macroscopic symptoms that have been used over the years.

1.2.3.2 Types of genomic variants

Genomic alterations can affect the sequence and the structure of the genome in different ways, and are usually classified according to type of DNA change (Figure 7):

- Single nucleotide variants (SNVs) are substitutions of one single nucleotide by another. They are the smallest and most common type of variant, and the most easily detectable.
- Small insertions and deletions (indels) are short insertions and deletions, usually up to 50bp.
- Copy number alterations (CNAs) encompass changes in the number of copies of a region in the genome, either due to duplications or deletions.
- Structural variants (SVs) are the most complex type of alterations where the DNA has been broken and reassembled elsewhere in the genome. They include deletions, insertions, where new DNA is acquired by exogenous sources like viruses; duplications, inversions, and translocations, where more than one chromosome is involved. A balanced translocation is an event where there is no loss of genetic material, whereas unbalanced translocation results in loss of DNA. Reciprocal translocations are two-way exchanges between two

non-homologous chromosomes. On the other hand, nonreciprocal translocations are one-way transfers of a chromosomal segment into a non-homologous chromosome.

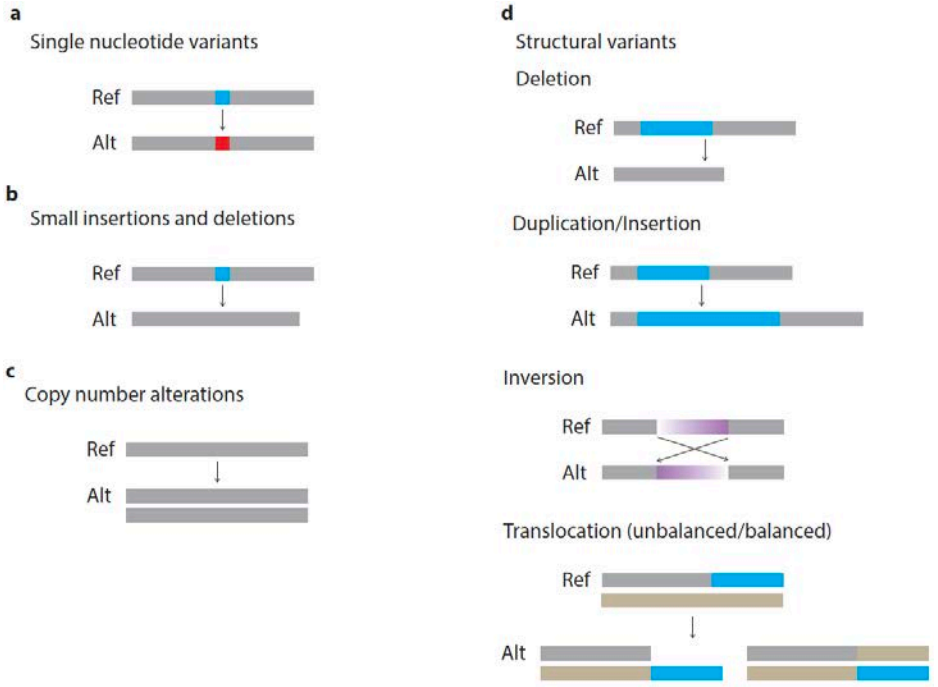


Figure 7. Types of DNA variants. Mutations in the genome can be classified according to the DNA sequence change into single nucleotide variants (a), small insertions and deletions (b), copy number alterations (c) and structural variants (d), where different types of alterations are included, namely deletions, duplications, insertions, inversions, and translocations, which can result in loss of genetic material (unbalanced, left) or not (balanced, right). Ref, reference allele. Alt, alternate allele harboring the alteration.

1.2.3.3 Complex genomic rearrangements, chromothripsis, and chromoplexy

Somatic alterations can be acquired individually, one at a time, or in a single catastrophic event that generates numerous alterations at the same time, leading to complex rearrangements that combine structural variants with copy number

alterations. Different types of complex rearrangements have been described based on clusters of structural variants in which multiple breakpoints occur close together, usually in time and in genomic space, implying that they might be mechanistically linked (Y. Li et al., 2020; Yi & Ju, 2018).

The first complex rearrangement was described in 2011. This event led to a massively reorganized chromosome in a single-hit event (Stephens et al., 2011). It was named chromothripsis that means "chromosome (chromo, which represents the chromosomes) shattering into pieces (thripsis in Greek)", which indeed describes the fragmentation of a chromosome into numerous segments that are wrongly repaired afterwards, supposedly by non-homologous end-joining (Figure 8). Using NGS, these defective rearrangements can be observed as numerous SV breakpoints (typically from 10 to 100) with similar proportions of all types (i.e., deletions, duplications, and inversions), clustered in one or a few chromosomal arms. Copy number alterations are also observed, including deletions, and loss-of-heterozygosity is frequent in minimum copy number regions. Chromothripsis is found in 2-3% of all cancers, and more frequently in bone cancers (25%) (Stephens et al., 2011).

There are two possible mechanisms behind this catastrophic hit: telomere crisis where telomeres are shortened and chromatids can be fused forming a chromatic bridge when they are stretched out during the anaphase of mitosis (Maciejowski & De Lange, 2017), and formation of aberrant nuclear structures (micronuclei) where the isolated genetic materials are massively broken into pieces and reassembled (C. Z. Zhang et al., 2015).

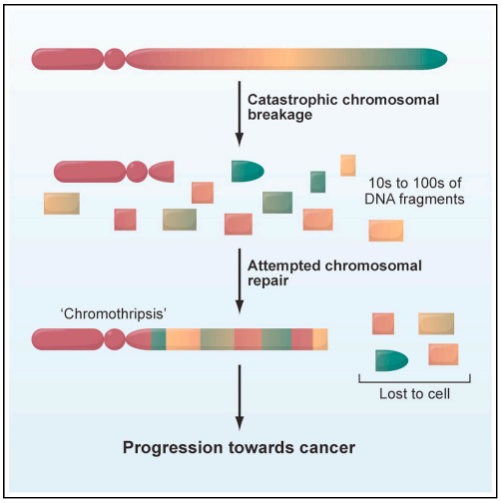


Figure 8. Chromothripsis representation. One or a few chromosome arms are shattered into pieces and wrongly reassembled by repair mechanisms in a single catastrophic event. Image from Stephens et al., 2011.

Chromoplexy is another type of complex rearrangement arising from a single catastrophic event. It results from multiple double-stranded DNA breaks in different chromosomes that are wrongly reassembled. It is characterized by interdependent SV breaks, mainly interchromosomal translocations that form balanced chains of rearrangements, involving three or more chromosomes, and it is usually copy number neutral, although small deletions can occur close to the breakpoints (Figure 9). It was first described in prostate cancer, where it is particularly prevalent (Baca et al., 2013). Chromoplexy events frequently disrupt tumor suppressor genes and/or active oncogenes. The mechanisms driving this phenomenon are not well understood, but the breakpoints distribution is enriched in open chromatin and active regions, suggesting that DNA injury might occur in transcriptional hubs occupied by co-regulated genomic regions from multiple chromosomes (Baca et al., 2013).

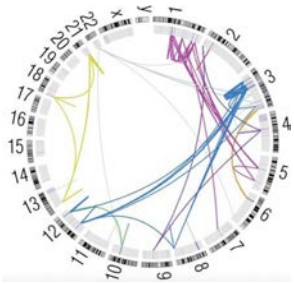


Figure 9. Chromoplexy events in prostate cancer. Circos plot showing chains of rearrangements in a prostate adenocarcinoma. Each independent chain is painted in a different color. The inner ring depicts the copy number alterations (red for deletions and blue for duplications). Image from Baca et al., 2013.

1.2.4 Large-scale computational technologies

The world of computing is constantly advancing. It is fed not only by new technologies, but also by the ever-growing interconnectedness of our society. The high-speed connectivity around the world has made cloud-based solutions very appealing to research and business centers. Large-scale computing is not only supported by local HPC clusters, but also by flexible cloud environments. Both systems offer different capabilities that make them more or less suitable depending on the type of work.

HPC aggregates computing power in order to deliver higher performance than traditional computers. A supercomputer is built on multiple computing nodes and fast storage devices that work together to complete tasks very efficiently. The nodes are networked at high speed and work in parallel with each other, boosting massive amounts of computing to be executed in a short period of time. HPC clusters are managed by batch queue systems that receive requests of jobs to run and schedule their execution according to the system load and pre-defined priorities.

Cloud computing represents a practical and cheaper way to scale compute capabilities that has gained popularity over the years. It also releases companies and research institutes from expensive hardware maintenance and software upgrades.

HPC processing power, together with its fast storage and network connections, will always outperform cloud computing in terms of speed. However, it comes with a cost, and its dedication to small or non-parallelizable tasks might be more expensive than necessary. Moreover, HPC regulations can be very stringent, and software developed within communities not dedicated to HPC might not be suitable or not even allowed (i.e., docker containers, or internet access requirements). In this case, cloud-based solutions might be more convenient.

1.2.4.1 HPC infrastructure at the BSC

The Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS) is the national supercomputing center in Spain. It specializes in HPC and has the MareNostrum (MN) saga of supercomputers at its base. It has an active role promoting HPC and providing HPC resources to the scientific community, including its four research departments (Computer Sciences, Life Sciences, Earth Sciences, and Computer Applications in Science and Engineering).

MareNostrum 3 (Figure 10) was the third supercomputer of the center. It was based on Intel SandyBridge processors, iDataPlex Compute Racks, and Infiniband interconnection, with a peak performance of 1,1 Petaflops. It had 52 racks, and 3,056 nodes (2x Intel SandyBridge 8-core, 2,752 nodes with 32GB of RAM, 128 nodes with 64GB of RAM, and 128 nodes with 256GB of RAM). Each computing rack had 1,344 cores and 2,688GB of memory. The compute nodes

were interconnected through a high-speed network based on Infiniband and had 500GB of local storage. It had Linux operating system (OS), and IBM LSF to manage the workload and schedule batch jobs. The storage infrastructure was managed by the General Parallel File System (GPFS), a high-performance clustered file system software developed by IBM.

Figure 10. MareNostrum 3 at the Barcelona Supercomputing Center.

The next supercomputer was MareNostrum 4 and achieved a peak performance of 13.9 Petaflops. Its computational power is distributed in two different blocks: a general-purpose section, and an emerging technologies section. The latter includes an IBM Power9 and NVIDIA Volta GPUs cluster, an AMD cluster, and a 64-bit ARMv8 processor prototype machine. The general-purpose part has 48 racks and 3,456 nodes. Each node has two Intel Xeon Platinum chips with 24 processors each and 96GB of memory. There are 216 high-memory nodes with 384GB of RAM. High-speed Omnipath network is used to interconnect all the computing nodes and other components. It has a disk capacity of 14 Petabytes. As its predecessor, Linux is the OS, and batch processing is administered by Slurm workload manager, a free and open-source job scheduler. GPFS was again used to manage the storage infrastructure.

Primarily, the BSC was born as an HPC center and, as such, it had, and still has, strict regulations. This includes rules that limit the kind of jobs that can be run on the machines and prohibitions at the security level, such as internet access or use of virtualized systems. The standard queues typically have a maximum wall time of 48 hours, or at most 72 hours. Non-parallel jobs are sent to the sequential queue, which might have some limitations (i.e., maximum number of running jobs per user). These requirements often clash with typical bioinformatics applications, which often consist of complex workflows that take a long time to complete, or simple independent tasks that are not very well parallelized. Embarrassingly parallel workloads consist of numerous tasks that can be run independently and are one of the most common kinds of workload that bioinformatics deal with. To pave their road to HPC, the BSC Support team developed a framework called GREASY that leverages the use of HPC resources for embarrassingly parallel tasks.

1.2.4.2 Workflows and virtualization

Virtualization refers to the creation of a virtual, or logical, instance of computing hardware, storage resources, or network devices. Hardware virtualization encompasses the creation of a virtual machine (VM) that acts as a real computer with its own operating system. The software executed on these VMs acts on the virtualized hardware and it is thus separated from the underlying bare metal. Therefore, virtualization becomes a reasonable solution to export and execute processes and programs across different architectures, which otherwise would require specific solutions.

Cloud infrastructures provide resources by means of virtualization, mainly VMs. They are managed by cloud management platforms (CMP) that integrate software tools to monitor and control the cloud computing resources. The most renowned ones are OpenNebula and OpenStack, and they can set-up dynamic

and flexible pools of computing and storage resources via simple web-based UI or programmatically.

In recent years, new ways of lighter software virtualization have emerged, together with the growing popularity of containers. Docker (Merkel, 2014) is probably the most well-known example of such light packaging. Containerization is focused on the creation, implementation, and execution of applications, easing their development and usage throughout their life cycle. The applications are bundled together with all their dependencies, making their distribution effortless. They can be easily shareable and portable, and their outputs can be reproducible. In contrast to VMs, containers do not have an OS, they share the host OS, which makes them less heavy than VMs (Figure 11). Containers can give more agility to both developers and operators and can be quickly deployed. However, VMs have a stronger separation from the host kernel, which makes them more secure.

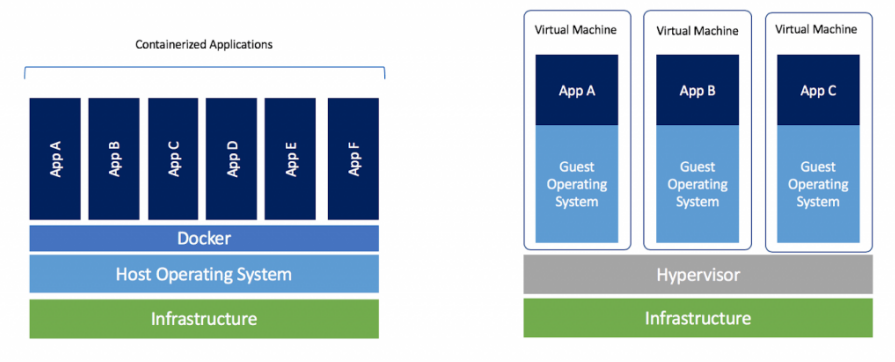


Figure 11. Components of docker containers and virtual machines.

Easy usage, portability, and reproducibility are some of the main advantages of docker containers, but their security risks and vulnerabilities make them unsuitable for most, if not all, HPC infrastructures. Docker images maintain root access to the host and can thus provide means to gain root access to the system

they are running on. Typically, HPC sites do not allow users to run Docker containers. Fortunately, most of them do allow Singularity containers (Kurtzer et al., 2017), which address these security issues. Unlike Docker, Singularity inherits permissions of the user who is running the container. Hence, unprivileged users outside the container will remain as such, and escalating privileges are prevented. Porting from Docker to Singularity is straightforward, as Singularity images can be easily created from prior docker images, and offer the same benefits of shareability, reproducibility, and portability.

Genome analysis can be as easy as executing a simple tool or reach higher levels of complexity where a myriad of different tools, with potential dependencies among them, is used. The latter usually requires integration of the results, often in different formats, burdening their unification. To help the development, deployment, portability, and reproducibility of such complex pipelines, workflow management systems have been employed (Ahmed et al., 2021). They present a solution to define and orchestrate computational pipelines across heterogeneous computational environments.

These shareable workflows can be uploaded to public repositories such as Docker Hub, a service provided by Docker for sharing and finding container images, or Dockstore, a free open-source platform for sharing analytical workflows developed by the Cancer Genome Collaboratory and used by the GA4GH. As part of the GA4GH, it takes part of promoting standards by defining best practices for describing tools in Docker containers with workflow language descriptors, such as Common Workflow Language (CWL) (Amstutz et al., 2016), Workflow Description Language (WDL) (Voss et al., 2017), and Nextflow (DI Tommaso et al., 2017). CWL is a standard for describing computational workflows that are portable and can be run in different environments preserving reproducibility. In the same way, WDL is a way to specify data processing

workflows, with human-readable syntax that allows easy definition of tasks and dependencies and supports their parallel execution. Nextflow has gained great popularity over the years and is currently being used in many local and large-scale projects, such as the ICGC-ARGO (see Introduction - section 1.4.2). It addresses reproducibility, efficient parallel execution, error tolerance, execution provenance, and traceability.

Despite the supposedly portability and extended usability of these systems, their use in pure HPC infrastructures is not exempt from breaking some regulations, and exceptions have to be made to allow their execution in those systems. For example, the Slurm executor of Nextflow can efficiently parallelize the tasks of a workflow, but they might be submitted as sequential jobs, which is not efficient within HPC systems, and requires a master process to monitor and schedule the whole workflow, often exceeding the maximum wall time allowed.

1.2.5 Translation of genomic knowledge into the clinics

Genomic research has generated an extensive hoard of data, yielding biologically meaningful findings that scientists and clinicians can use to decipher the role that genetic factors might play in the development of complex diseases, such as cancer. What is more, these discoveries can also be used to implement more accurate diagnostics, more effective treatment strategies, and, altogether, better decision-making in the clinics that will improve the life of patients. In oncology, it has also increased the options for molecular targeted therapies, which act on specific molecular targets to block tumor cell growth and proliferation.

On this ground, the role of genetics in health care is starting to become increasingly important. NGS plays a promising role in the era of precision medicine, where treatments are tailored to the patient's genetic make-up. It is

well known that genetic variation in tumors can explain the variability of treatment effectiveness between individuals and can guide therapy decisions.

Traditionally, patients would be treated according to their specific disease, or, in oncology, according to their cancer type and stage. However, it has been largely seen that this "one-size fits all" approach (Figure 12) does not always work well for all patients, who might have different responses to treatment (Figure 13). Research studies have found associations between genetic alterations and treatment resistance (Furman, Cheng, et al., 2014; Wagle et al., 2011; Woyach et al., 2014), asserting the urgency to integrate genomic analysis into clinical decision-making.

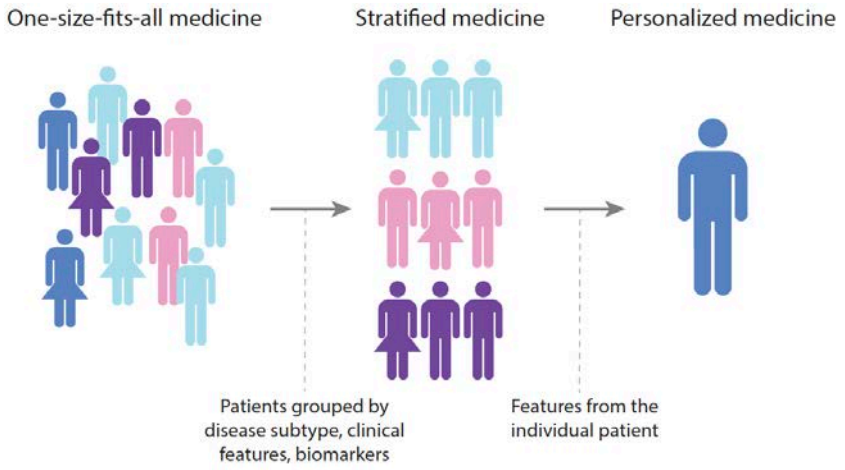


Figure 12. From one-size-fits-all to precision medicine. Traditionally, patients were grouped by disease type and treated equally. The next approach is more specific and stratifies patients into groups according to disease subtypes, clinical features, and available biomarkers. Finally, the personalized medicine approach proposes tailored treatments to each individual patient.

Personalized medicine (PM) considers the patient's molecular profile, together with their clinical history, and environmental and lifestyle factors. The awaited benefits of joining together all this information are earlier diagnosis and

preventive approaches, more precise prognostics, better treatment selection, and an overall improved patient management. The potential of PM does not stop at healthcare quality, it can also have an impact on economics, reducing costs of expensive treatments that might not work on individual patients.

The identification of targetable alterations and the coinciding development of small molecule-targeted and antibody-based therapies in cancer has encouraged the transition of genomic assays into clinical use (Berger & Mardis, 2018). Somatic mutations within a tumor can be incredibly helpful and it is increasingly being used to guide the selection of the most appropriate treatment for each patient according to their cancer's genome (Figure 13).

Cancer genomics has the power to remodel traditional medicine by identifying specific alterations that can guide clinical decision-making for each individual patient. But before incorporating research findings into patient care, the significance of potentially actionable alterations has to be evaluated, and biological and clinical interpretation of genomic events identified by computational predictions still remains a challenge (Good et al., 2014).

Bringing genomic testing into clinical practice is not straightforward, as genomic analysis, widely applied in research, has to be adjusted for its proper translation into the clinics (Morganti et al., 2020; Xu et al., 2019). Despite numerous efforts to define guidelines and best practices for variant calling, annotation, and variant interpretation, its application to the clinics is still problematic, due to the diversity of available tools, discrepancies of their results, and lack of real benchmarking datasets for variant detection, especially in the field of somatic variants.

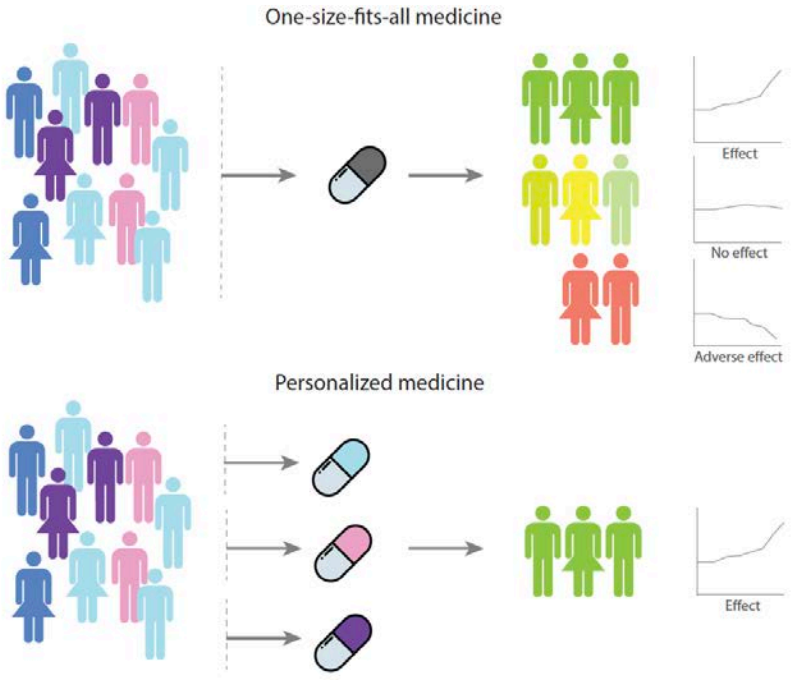


Figure 13. One-size-fits-all vs personalized medicine treatment strategies. Within the traditional management of patients, the same treatment is used for all patients with the same disease. This can lead to benefit for some patients, no effect for others, and, unfortunately, to adverse effects for some. Personalized medicine aims to give the best treatment option to each individual patient, achieving treatment success for all cases.

1.3 Cancer: a disease of the genome

The biological processes within our cells are tightly regulated. This intricate coordination preserves our health, but it is not infallible. Intrinsic and extrinsic factors can interfere and deviate the normal functioning of cells, leading to the formation of masses of cells that can grow uncontrollably, called neoplasms or tumors. Neoplastic cells can be classified as benign, when they have no capacity to invade other tissues, or malign, when they can spread to other parts of the body. Malign tumors, commonly referred as cancer, are a major threat to our health, and encompass more than 100 different diseases, which can originate

from most cell types or organs of the body, and have the capacity to reach, or metastasize, to tissues beyond their boundaries (Stratton et al., 2009).

Cancer has a high impact on human health and is one of the major causes of mortality. According to the World Health Organization (WHO), nearly 10 million people are estimated to have died of cancer in 2020. Despite research and clinical advances, the cancer burden continues to grow, and it is expected to reach 28.4 million cases in 2040 (Sung et al., 2021). The most common types of cancer vary among men and women, and incidence and mortality are higher in less developed countries. Strong health systems integrating new strategies for earlier diagnosis and better tailored treatments are needed to improve global cancer control, especially in transitioning countries.

1.3.1 Molecular basis of cancer

Cancer is a disease of the genome, caused by genomic aberrations that deregulate the normal functioning of the cells. While some genetic factors are hereditary, the root of most cancers lies on somatic variants that are accumulated throughout our life. These alterations, which can be as simple as a single nucleotide change or large events involving one or more chromosomes, can confer the cells advantageous capabilities that can lead to the formation of a tumor. Besides the genetic counterpart, epigenomic changes, comprising the DNA modifications that do not affect the sequence per se, can also affect the activity of genes, leading to uncontrolled monitoring of key biological processes, such as cell growth and proliferation.

The biology of cancer can be quite complex, but it has been largely seen that there are commonalities among different cancer types, mainly affecting the molecular machinery regulating cell proliferation, differentiation, and death.

These disrupted processes were summarized into six acquired capabilities or hallmarks of cancer (Hanahan & Weinberg, 2000):

1. Self-sufficiency in growth signals. Tumors have different mechanisms to promote cell growth and proliferation. They can overexpress membrane receptors or other proliferative signals or constitutively activate downstream molecules that activate and maintain chronic proliferative signaling.
2. Evasion of growth-inhibitory signals. Tumors can become insensitive to growth control signals, hence maintaining their ability to grow. Tumor suppressor genes that regulate cell growth and proliferation have been found inactivated in many animal or human cancers. The two major examples are *TP53* and *RB1*, which have a central role in pathways that determine the course of cells, activating senescence and apoptosis or stopping cell-cycle progression.
3. Resistance to programmed cell death. Apoptosis is a programmed cell death that is triggered when cells are damaged. Tumor cells have different ways to evade it, allowing them to continue to grow and proliferate. The most common strategy is *TP53* disruption, which is seen in more than 50% of human cancers and impedes its proapoptotic function. *BCL2* is another example that through overexpression can induce its antiapoptotic activity.
4. Limitless replicative potential. The length of the telomeres, which is shortened each time a cell divides, is used to determine when a cell should die, and thus controls the number of times a cell can divide. Cancer cells can have the capacity to extend telomeres avoiding their erosion by overexpressing telomerases that ensure telomere maintenance. Through this process, regardless of the times they have replicated, they can continue to replicate permanently.
5. Sustained angiogenesis. All cells need nutrients and oxygen to survive and obtain them from blood, which travels through vessels that irrigate tissues.

As tumor masses grow, they require the creation of new blood vessels (angiogenesis) to supply enough resources to their continuously dividing and growing cells. Angiogenic activation can be triggered by signaling proteins that are upregulated in tumors.

6. Activation of tissue invasion and metastasis. Cancer cells have the capability to invade surrounding and distal sites by altering the signaling between them and stromal cells and degrading the cellular matrix to gain motility. Loss of E-cadherin, a cell-to-cell interaction molecule, is one of the most common mechanisms to confer tumor cells their invasive phenotype.

These core principles, known as the hallmarks of cancer, are shared by most and probably all types of cancer, are illustrated in Figure 14.

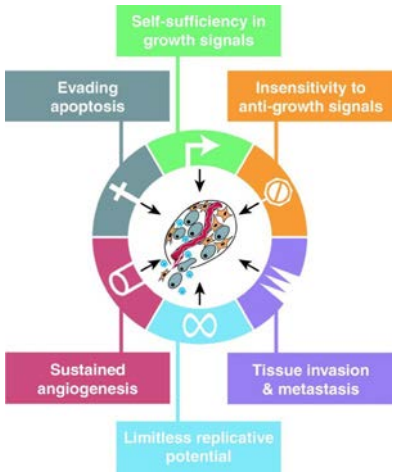


Figure 14. Hallmarks of cancer. Regulatory circuits disrupted in tumor cells. Image from Hanahan & Weinberg, 2000.

The six hallmarks of cancer represent a conceptual framework for describing the principles that govern the transformation of normal cells into neoplastic tissues. This pathogenesis, though, cannot only be seen as an isolated tumor mass. Malignant cells form complex tissues that are composed of different cell

types that can interact with one another, which constitute the tumor microenvironment. Normal cells forming the tumor-associated stroma are not passive participants but can be actively contributing to tumorigenesis. To incorporate this new layer, as well as new insights in cancer biology, a reexamination of the previous hallmarks was done, announcing two new enabling characteristics and two new hallmarks (Hanahan & Weinberg, 2011), depicted in Figure 15.

The two enabling characteristics are:

1. Genomic instability and mutation. The cancer genome successively accumulates genomic alterations that might confer advantageous properties and lead to deregulation of key cell regulation programs.
2. Tumor-promoting inflammation. The immune system can respond to tumors and, contrary to their intention, infiltrating immune cells can promote tumor progression by supplying molecules and signals that can enable multiple cancer hallmarks.

The two additional emerging hallmarks are:

1. Deregulating cellular energetics. The substantial activity of cancer cells to grow and proliferate entails some adjustments in their metabolism. Instead of using energy from the mitochondrial oxidative phosphorylation they turn to another system, the “aerobic glycolysis”, and possibly use glycolysis intermediates in biosynthetic pathways to aid the generation of new cells.
2. Evading immune destruction. Immune surveillance, our ever-alert immune system, can recognize and eradicate most initiating tumor cells. However, progressing tumors can bypass exposure to the immune system, avoiding their destruction.

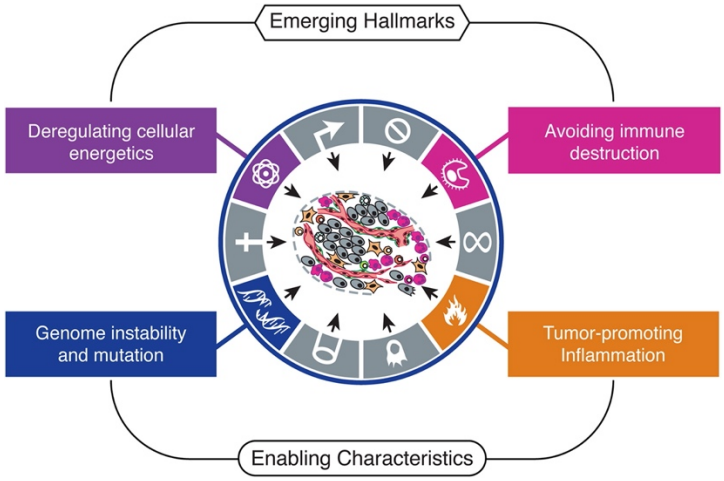


Figure 15. Enabling characteristics and new emerging hallmarks from the revision of the original 6 hallmarks of cancer. Image from Douglas Hanahan & Weinberg, 2011.

These molecular discoveries uncover not only the intricate processes of tumorigenesis, but, more importantly, the potential of being translated into the clinics, where they can be used for better diagnosis and treatment options.

The introduction of targeted therapies as new treatment strategies for cancer relies on previous knowledge of the molecular basis of cancer. In this line, insights into the pathogenesis of cancer and its underlying principles, previously described, are of utmost importance for the development of new therapeutic strategies. These promising treatments target specific proteins that control processes that promote cancer cell capabilities. In principle, the precision of these drugs can reduce their side effects, because they have less off-target activity, but the downfall is that initial clinical responses are usually followed by relapses. An explanation to this is the growing evidence that hallmark capabilities can be sustained through multiple pathways. To bypass the inhibition of one single circuit, tumor cells can use this redundancy to sustain a particular capacity, or they can switch to other hallmark capabilities to maintain their pathogenic abilities.

As we learn the bases or hallmarks of cancer, we might wonder how they are gained. What triggers malignancy in normal cells? These malignant cells derive from normal cells that go through a progressive transformation acquiring genetic and epigenetic changes that can target multiple sites of the genome and affect different regulatory pathways that, in the end, disrupt the normal behavior of the cell and activate one or more hallmarks of cancer.

1.3.2 Bioinformatics analysis of cancer genomes

In the era of NGS, cancer genomes can be analyzed by sequencing techniques followed by bioinformatics analysis. To decipher the mutational landscape of tumors, somatic alterations must be identified and discerned from germline mutations. Thus, a normal-matched sample from the same patient is commonly used.

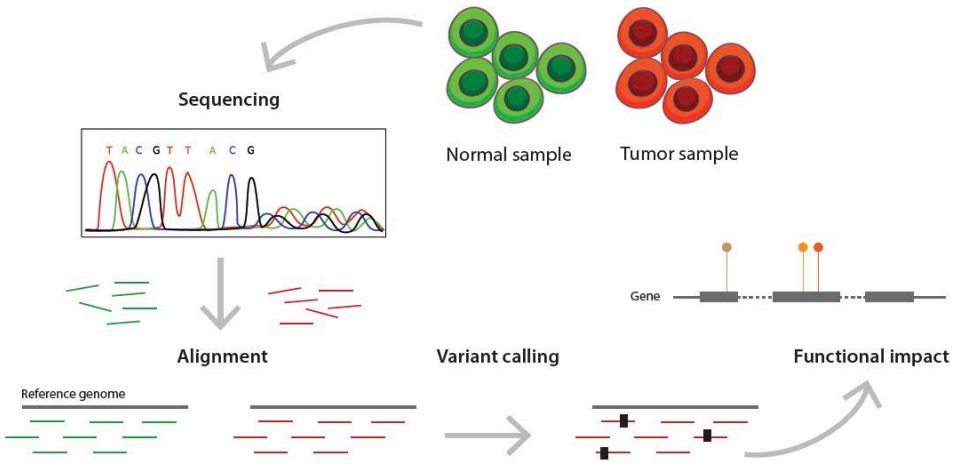


Figure 16. General strategy for the genomic analysis of tumor genomes.

The general strategy (Figure 16) starts from the comparison of normal and matched tumor samples, whose DNA is prepared through experimental protocols, and sequenced as previously explained (see Introduction - section 1.2.2.2). In the

case of cancer, both the normal and the tumor sample are aligned separately, and they are both inputs of the variant calling step, where somatic variants can be inferred from the comparison of tumor and normal variants. Finally, the functional impact of the detected variants can be evaluated, and other downstream analysis can be performed.

There are many factors that can complicate somatic variant calling, which reduce its accuracy due to the introduction of false positives or the miss of true variants. Starting with sequencing errors and alignment artifacts, the inputs of this core step are already obscured. And even more so if tumor samples come from FFPE material, which is known to have DNA fragmentations and alterations. In addition to these technical determinants, tumor samples themselves can defy variant calling methods, veiling the identification of low frequency variants due to tumor heterogeneity, or low purity of the sample. Matched normal samples can also be contaminated with tumor cells in some cancers, such as chronic lymphocytic leukemia. Automated workflows try to provide high accuracy, filtering out potential artifacts while keeping true variants, but, typically, a manual review of the results is also needed.

In the typical strategy, where tumor and normal matched samples are available, candidate variants are called on genomic positions where an alternate allele is supported by tumor reads and is not present in the normal sample (as long as there is no tumor-in-normal contamination). The frequency of the variant is defined by the variant allele fraction (VAF) that counts the percentage of supporting reads (total number of alternate reads divided by the total read depth, or coverage, at that position). The VAF is affected by the purity of the tumor sample and the copy number at that region, and many posterior analyses (e.g., tumor evolution) require its correction to indicate the actual fraction of tumor reads carrying the variant, the cancer cell fraction (CCF). Subclonal variants,

present only in a small number of tumor cells, present a low VAF and CCF and are the most difficult to detect, as they are only present in a minority of alternate reads (Figure 17).

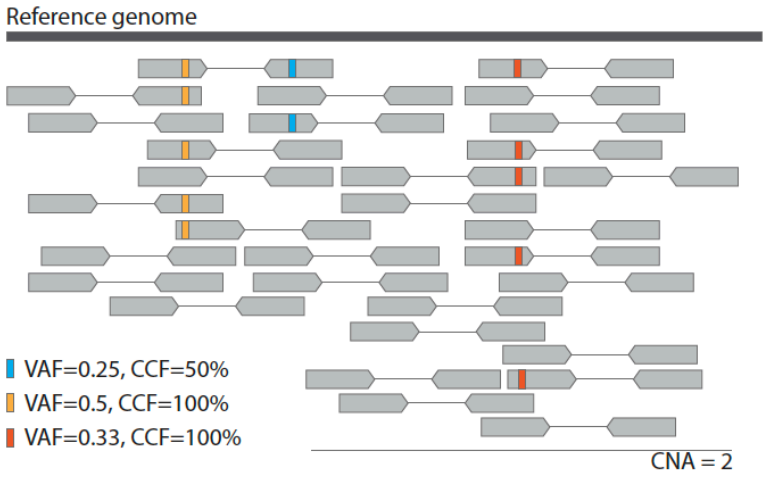


Figure 17. Schematic view of reads with mutations. The variants are marked as colored squares, the number of supporting reads is used to calculate the variant allele frequency (VAF), and the cancer cell fraction (CCF) that corrects for purity and CNA. The example shows a tumor with 100% purity. A subclonal variant is depicted in blue.

Variant calling can identify different types of variants, namely: SNVs, indels, CNAs and SVs (see Introduction - section 1.2.3.2). Some tools are dedicated to one single class, while others can detect different kinds of variants (e.g., sometimes SNVs and indels might be called jointly). The degree of difficulty in finding each type of variant is different, SNVs are the easiest and SVs the most complex ones. Point mutations and short insertions and deletions are detected as mismatches between the aligned reads and the reference genome, CNAs require more complex algorithms, where normalized coverage and/or B-allele frequencies must be considered, SVs can be detected by split reads, where part of the read maps to another region, and pair discrepancies in orientation, mapping chromosome, and/or insert size (Figure 18).

Figure 18. Examples of discordant reads to detect SVs. A deletion is represented at the top, paired ends have opposed orientation and larger insert size. An inversion is illustrated at the bottom, supporting paired reads have the same direction, as one of them is within the inverted region, and larger insert size. Split reads are depicted in blue.

Even though the tumor-normal design is the preferred approach, sometimes a matched normal sample is not available, especially in clinical settings. Tumor-only analysis greatly obscures true detection of somatic variants, as germline variants cannot be truthfully discarded. The most common attempt to overcome this limitation relies on public germline resources (Karczewski et al., 2020; Sherry, 2001) and panels of normals, calculated from pools of non-matched normal samples, that are used to filter out potential germline variants.

There is a large repertoire of individual tools and complex pipelines to perform variant calling. Gold standard datasets for benchmarking of variant discovery pipelines are essential to evaluate their performance. The Genome In A Bottle Consortium (GIAB) and the National Institute of Standards (NIST) compiled a dataset for germline variation that includes high-confidence genotypes and confidence regions for a set of samples. The high-confidence variants can be used

to assess the precision and sensitivity of variant callers. Best practices for benchmarking variant calling have also been developed by the GA4GH, and include a reference implementation of their strategy on the evaluation of germline variants (Krusche et al., 2019). However, benchmarking datasets and strategies are not so well defined for somatic variation, where publicly available real datasets are very limited, and tumor features such as intratumor heterogeneity hinder the results. A common approach to improve variant discovery of individual tools, is to combine the results of different methods. This strategy has been adopted by many institutional pipelines and also the main large-scale cancer genomics projects (Campbell et al., 2020; Ellrott et al., 2018).

In addition to the proper selection of tools for genome analysis, an understanding of the quality of the raw data, as well as the intermediate results, is essential. Quality control metrics can inform about the level of veracity of the final results and/or their limitations for a particular analysis.

1.3.3 Driver and passenger mutations

Since the very first moment of our existence, the fertilized egg, our genome can accumulate mutations that can arise due to intrinsic or extrinsic factors. To protect our well-being, our cells have very stringent mechanisms to prevent DNA damage, either by repairing it or by initiating cell death (apoptosis). However, some mutations might escape these controls and are settled in our genome. Most of them will have no effect, but a few might hit a key cellular function. When a variant confers the cell some selective advantage, it enhances the possibilities of this cell to expand and proliferate through positive selection which can lead to cancer formation.

Starting with the discovery of a point mutation that leads to the activation and transforming capacities of a gene, *HRAS*, in human bladder carcinoma (Reddy

et al., 1982), the search for gene abnormalities that can contribute to cancer development has been one of the pillars of cancer research (Bailey et al., 2018; Martínez-Jiménez et al., 2020). Driver mutations are defined as those that are directly implicated in oncogenesis and have been positively selected at some point during tumor evolution, though they do not need to be present to maintain the final cancer. On the other hand, other somatic variants are called passenger mutations when they do not confer any selective advantage to the cells carrying them, hence not promoting cancer development (Stratton et al., 2009).

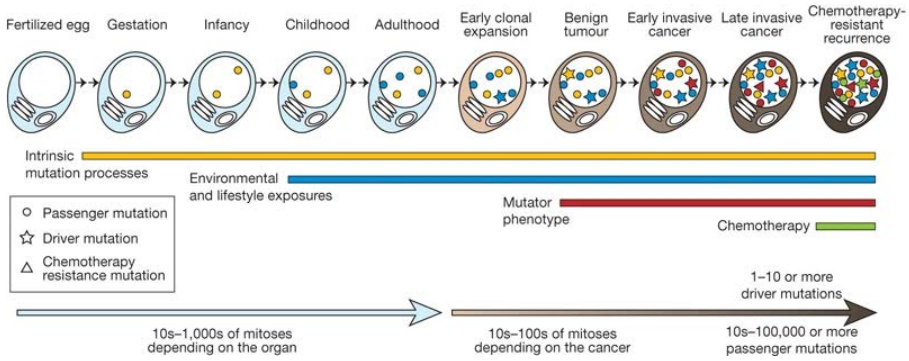


Figure 19. Lineage of mitotic cell divisions from the fertilized egg to a cancer cell. Mutations may be acquired due to intrinsic or extrinsic factors. Driver mutations, capable of malignancy transformation, can lead to clonal expansions that underlie tumor formation. Treatment forces can stimulate the emergence of resistant clones that often preexist before treatment.

Somatic mutations often occur during cell division, by exposure to mutagen agents, or due to failure of intrinsic mechanisms, such as defective DNA damage repair. Most of them will probably have no functional consequences by themselves. However, they will already be present in the genome when one or more driver events occur. Consequently, they will be carried by all cancer cells that originate from this initial malignant cell. Once the tumor is formed, it continues to evolve. It can follow a treatment-naive natural evolution or a constrained evolutionary trajectory due to the selective pressures of treatment.

During the latter, mutations in minor subclones can gain selective advantages within the new treatment-based environment, and can emerge as the dominant clone, which can be resistant to treatment (Figure 19).

The term driver is also extended to the genes that harbor such alterations. There are two major types of driver genes: (1) Genes that can promote cell proliferation and survival when activated by a mutation are known as oncogenes, whereas (2) driver genes that undergo loss-of-function are tumor suppressor genes, which in normal conditions should bring a mutated cell to apoptosis. In both cases, driver genes convey tumorigenic traits to the cells. Tumors carry an average of 4 driver mutations (ranging from <1/tumor to >10/tumor, depending on the cancer type), and it is estimated that half of the driver events occur in yet-to-be-discovered driver genes (Martincorena et al., 2017). Altogether, these mutations are only a small fraction of the overall mutational burden of the cancer genome. Passenger mutations, usually present in thousands, are commonly thought not to play a role in cancer development, although their contributing roles have also been described (Supek et al., 2014). These passenger mutations in fact represent an imprint of the mutagenic events that tumor cells have experienced and provide a valuable tool to reconstruct the history of the tumor and the mutagenic processes that have been active throughout the evolution from a normal to a malignant cell.

1.3.4 Mutational processes in cancer

The accumulation of somatic mutations in tumor genomes is the result of different processes that operate throughout our life and during the formation and evolution of neoplasms. These mutational processes generate a specific and characteristic mutational pattern and can be triggered by both endogenous and extrinsic factors, such as DNA repair mechanisms or mutagens like tobacco or

ultraviolet light. For example, tobacco smoking induces C>A transversions, while ultraviolet rays introduce C>T transitions in specific contexts (i.e., the change mainly occurs when it is preceded by a thymine and goes before a cytidine (T[C>T]C)). Each one of these mechanisms can leave an imprint in our genome, as they might have a preference in specific mutation types and in particular contexts. Statistic and mathematical models can recognize and deconvolute these signatures, which can be linked to specific etiologies or exposures through statistical association or experimental validation (Koh et al., 2020). Insights into these mutational processes have been nourished by the increasing availability of whole genome and exome sequencing and the gathering of large-scale genomic projects used to define catalogs of signatures of known and unknown etiologies. The COSMIC mutational signatures catalog is established as the largest compendium of mutational processes known to date. The current version compiles the latest work by the PCAWG Network (Alexandrov et al., 2020; Campbell et al., 2020), including data from more than 23,000 tumor exomes and genomes.

The imprint of mutational processes can be extracted from the mutational landscape of tumor genomes, which comprises a complex high dimensional dataset that can be dissected into individual signatures. Next, the contribution of each signature to each cancer genome can be quantified and linked to the exposure of each mutational process (Alexandrov et al., 2013). The non-negative matrix factorization (NMF) and model selection is one of most used approaches to resolve this deconvolution and can be used to blindly separate multiple patterns from a multidimensional dataset, as well as to estimate the relative contribution of each signature to the mutational catalog of individual tumors (Figure 20).

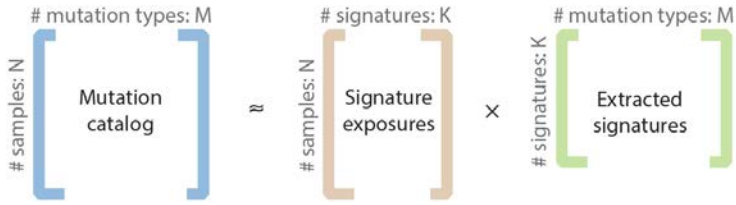


Figure 20. Deconvolution of mutational signatures. Decomposition of a mutational catalog of M mutation types from N samples into a set of K mutational signatures and their exposure to each sample.

Mutational signatures are commonly studied based on SNVs, but can also be characterized from dinucleotide substitutions, indels, CNAs, or SVs (Y. Li et al., 2020). The SNV 96-mutational profile is the most widely used and consists of 96 single nucleotide mutation types that consider the mutation itself, the single base substitution (SBS), together with the flanking bases at the 3' and 5' sides (Figure 21).

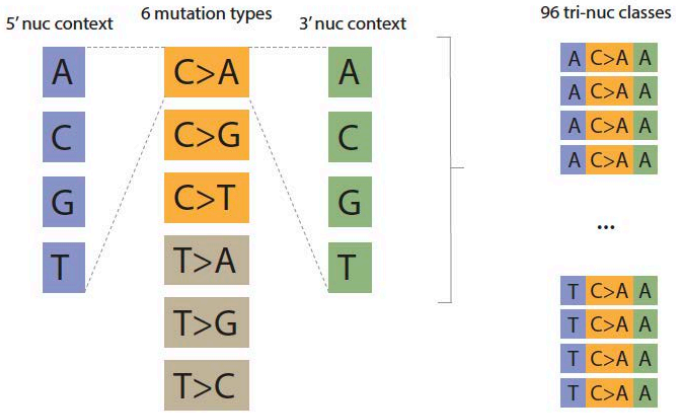


Figure 21. 96 tri-nucleotide classes of SNVs. The 96 classes take into account the variant, together with the 3' and 5' flanking bases, which provide the context.

Other approaches including SNVs are also used, where the context can be expanded to four flanking bases instead of two, leading to 1536 classes. This wider

approach can provide a better characterization and/or confirmation of novel mutational signatures (Haradhvala et al., 2018; Rustad et al., 2020).

As an illustrative example, the pattern of 4 single base substitution mutational signatures (SBS1, SBS4, SBS2 and SBS13) based on the 96 mutation types previously described can be seen in Figure 22. The relative amount of each class is indicated by the bar plots highlighting the predilection of specific mutation types. SBS1, together with SBS5, is present in virtually all cancer types and is considered a clock-like signature, as it correlates with the age of individuals and the rates of stem cell division, and may therefore serve as a mitotic clock (Alexandrov et al., 2015). It is characterized by the most common deamination reaction of 5-methylcytosine to thymine that generates G:T mismatches that can be fixed as C>T mutations when they are not previously repaired. Other known signatures with well-defined etiologies include those of external mutagens, such as tobacco smoking (SBS4), mainly defined by C>A transversions, or internal mechanisms such as the activity of the AID/APOBEC family of cytidine deaminases (SBS2 and SBS13).

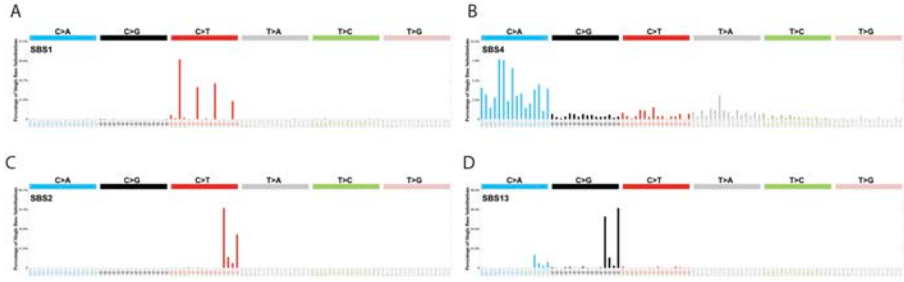


Figure 22. Examples of the classic 96-mutation profile of mutational signatures. A. Clock-like signature SBS1. B. DNA damage by tobacco smoke imprint, SBS4. C-D. Imprint of the activity of AID/APOBEC enzymes, SBS2 and SBS13.

There are many mutational signatures that can be attributed to known processes, but others remain elusive. In addition, mutational processes can be the

union of both the DNA damage and the consequent DNA repair. Cells have intricate apparatus to counteract DNA alterations, which trigger various DNA repair pathways. When there is a deficiency in those mechanisms, their attempts to correct mutations might lead to new mutational patterns. Altogether, the interplay between mutagens and awry DNA repair mechanisms can jointly shape mutagenesis, giving place to distinct mutational imprints. Hence, mutational signatures might not have a one-to-one relationship to mutagenic processes, but they can be variable and molded by DNA repair or replicative defects (Volkova et al., 2020).

The analysis of mutational signatures is hampered by the fact that purely mathematical algorithms might not be biologically accurate. The fitting of potential signatures in tumor samples might lead to misleading results, where signatures that are not biologically present in the sample might be theoretically identified. This can specially happen if the signatures share dominant peaks. In the same way, the cosine similarity function that is used to calculate signature similarities might not capture genuine comparisons, as it works best for signatures with hilly peaks, but it is less effective for flatter signatures. Additionally, mutational signatures might appear slightly different in different tissue types. Taking all this together, rather than blind confidence in mere mathematical algorithms, a final assessment of the biological validity of the results is an essential practice. Although there are no standard protocols to conduct these analyses, efforts towards best practices and consensus strategies have been made (Alexandrov et al., 2020; Maura, Degasperi, et al., 2019).

Mutational signatures do not only provide a view into the evolutionary history of tumors, but they can also have a clinical value. They can be used as biomarkers for endogenous DNA repair/replication defective mechanisms or exogenous carcinogen exposures, and can be indicative of prognosis and therapy

efficacy (Brady et al., 2021). Finally, treatment-induced mutations, such as those induced by chemotherapies, can also guide the investigation of the long-term effects of those exposures (Pich et al., 2021) and the study of evolution in tumors, pinpointing the expansion of seeding cells already present before therapy (Pich et al., 2019; Rustad et al., 2020).

1.3.5 Tumor heterogeneity

Tumors are traditionally classified by their primary site of origin or by tissue type. This initial categorization does not consider the variability that exists among patients. It has been largely seen that tumors have a unique combination of genetic alterations. Each person's cancer harbors a set of DNA changes that make it distinct from other patients, even if they have the same cancer type. This variation is called intertumoral heterogeneity, and accounts for different prognosis or treatment responses among individuals. However, this variability among patients is not the only source of heterogeneity in tumors. Even within a tumor, different cancer cell populations can coexist, each having particular mutational profiles, features, and capabilities that generate its intra-tumor heterogeneity (ITH) (Figure 23).

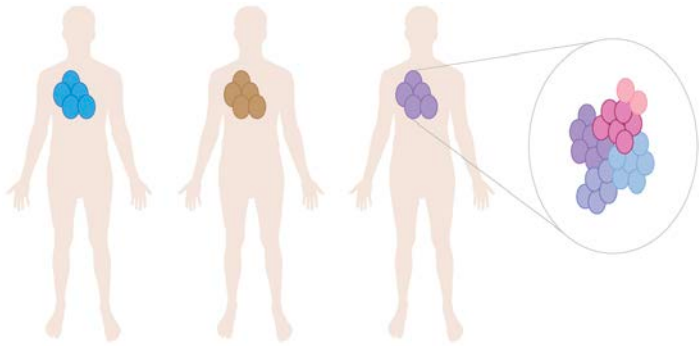


Figure 23. Intertumor and intratumor heterogeneity. Variation among patients with the same cancer time comprises the intertumor heterogeneity, while variation within a tumor of a single patient constitutes the intratumor heterogeneity.

ITH is an urgent clinical challenge, as it is implicated in therapy resistance and cancer evolution (Dentro et al., 2021). Early studies considered tumors as a homogeneous mass and mainly focused on clonal alterations. Subclonal mutations are more difficult to detect because they are only present in a small fraction of reads, often below the limits of detection of variant callers. However, thanks to our growing understanding of how tumors evolve, and the reducing costs of sequencing which provides affordable higher coverage sequencing, we can now infer the composition of tumors and unveil their underlying ITH.

Cancer evolution is marked by genetic diversification, clonal selection within the possibly changing microenvironment, and posterior expansion of the most advantageous subclones (Greaves & Maley, 2012). Evolutionary changes at the level of DNA changes can be traced and used to assess the subclonal architecture of tumors. Subclonal reconstruction based on genome sequencing of bulk tumor samples relies on the frequency of somatic mutations to identify cancer '(sub)clones', which are entities that share a number of mutations that were present in a common ancestor. The most used approaches perform an unsupervised clustering of variant allele frequencies, adjusted for copy-number and tumor purity, to identify clusters of mutations that have similar cellularity. The resulting clusters represent the different subclones that might be present in the studied sample. Finally, ITH can reveal a tumor's life history, elucidating the temporal order of the acquired somatic events. The tumor's phylogenetic tree can be inferred from the different subclones and their feasible relationships (Nik-Zainal, Van Loo, et al., 2012). The trunk of the evolutionary tree represents the mutations identified in all cancer cells, while subclonal alterations, present only in a subfraction of tumor cells, make up the branches. However, sequencing of a tumor sample only provides a static snapshot of its genetic landscape and will likely provide an underestimation of the actual variety of tumor cell subpopulations (Gerstung et al., 2020). Hence, using multiple samples from the

same patient, multi-regional or longitudinal, can add valuable information that will aid the decomposition of more subclones and their dynamics.

A tumor's ITH and its capacity to evolve and adapt to changing environments, such as therapy aggressive constraints, are key contributors to therapeutic failure and dismal outcome of cancer (Greaves, 2015). Cancer medicine should not only consider the clonal population at a time, but also the underlying ITH, which can foster tumor evolution and clonal shifts where the tumor's composition can be completely changed. ITH may also be used as a prognostic or predictive biomarker (Venkatesan & Swanton, 2016). However, despite the known role of clonal evolution in treatment failure, ITH and clonal dynamics are infrequently considered to inform clinical decisions.

1.3.6 Tumor evolution

Cancer is a dynamic disease, it evolves over time, specially under selective pressure of treatments, making patient management more complex. This evolution is fueled by the underlying ITH, where an admixture of cell subpopulations, or subclones, interact and compete for resources, leading to the expansion of the fitter clones. This evolutionary process was described by Nowell (Nowell, 1976) as a stepwise series of events driven by the acquisition of successive somatic mutations and selection of advantageous subclones. This model resembles Darwin's theory of natural selection, which has been adopted to explain the basis of tumor evolutionary trajectories (Figure 24). Viewed in this way, cancer evolution is based on two essential processes, the acquisition of genetic variation in individual tumor cells and natural selection, where cells with selective advantages with respect to their neighboring cells will outcompete them and will act on the resultant phenotype (Stratton et al., 2009). This phenomenon

is called clonal expansion and can be seen from the formation of a tumor to their final stages, especially when tumor cells need to escape treatment constraints.

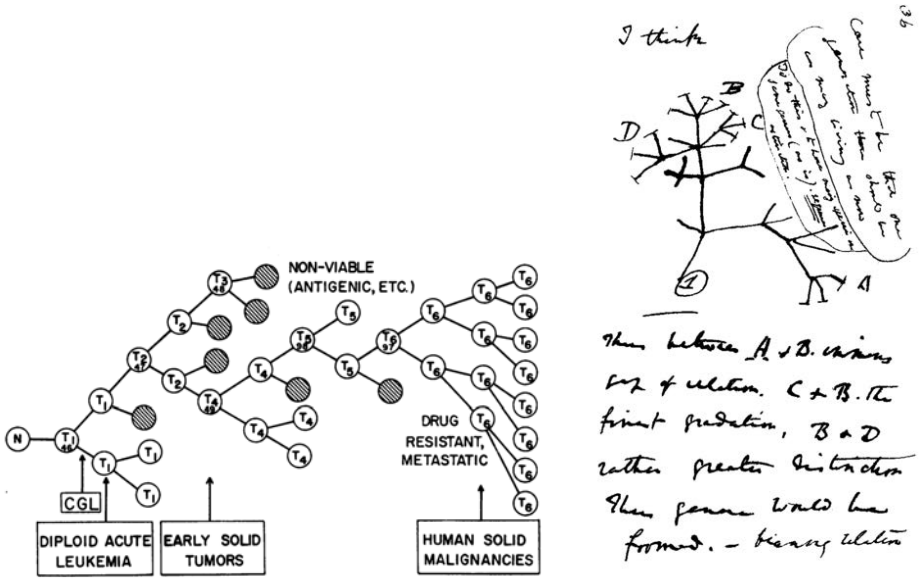


Figure 24. Tumor evolution by Nowell in 1976 and Darwin's theory of natural selection. Left, Nowell proposed a model of clonal evolution where a neoplastic cell gains selective growth advantages and proliferates (T1). Within the expanding tumor cells new genetic alterations can be acquired (T2 to T6) and can, in turn, confer selective fitness advantages (T6), or disadvantages leading to their extinction (hatched circles). Image from (Nowell, 1976)*. Right, Darwin's iconic tree drawing of 1837 showing the evolutionary tree of speciation. Image from Charles Darwin's Notebook, 1837.*

The first models to describe tumorigenesis were based on Darwin's principles, but emerging evidence from new technological developments in genomic analysis posit alternative modes of evolution that cannot be explained by the conventional stepwise processes. Altogether, tumor dynamics can be explained with Darwin's theory and beyond (Figure 25).

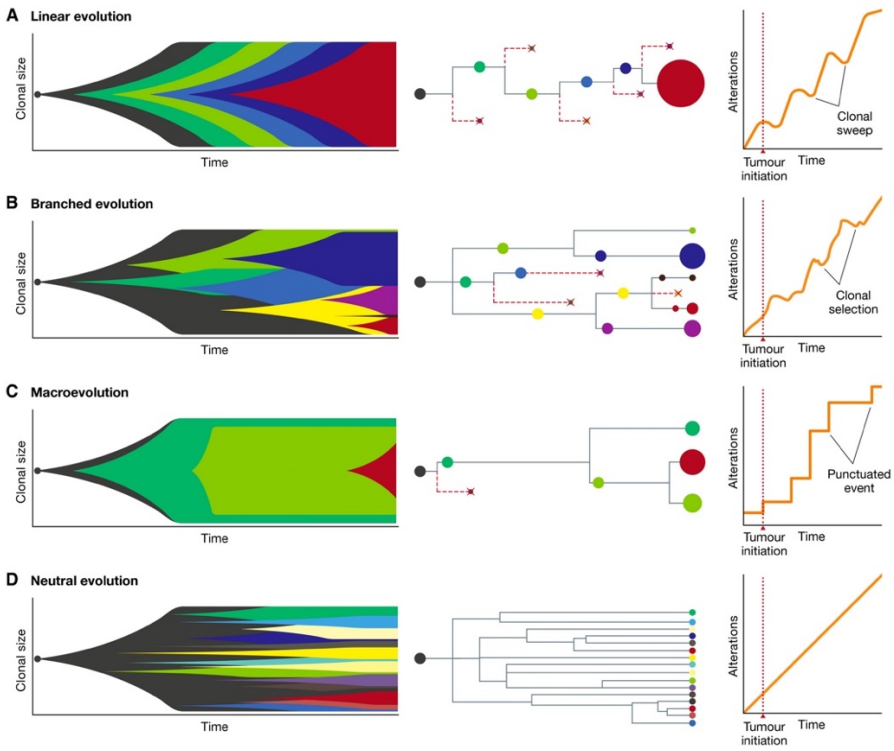


Figure 25. Models of tumor evolution. Linear evolution (A), branched evolution (B), macroevolution (C) and neutral evolution (D). Muller plots (left) represent the clonal dynamics and their clonal size over time, phylogenetic trees show the lineages of clones (center) and linear plots (right) indicate the number of alterations over time. Image from Vendramin et al., 2021.

The Darwinian view of cancer models tumors as a large population of cells with diverse genetic alterations that can give rise to distinctive subpopulations. These subpopulations, or subclones, compete with each other and face changing pressures from the microenvironment or imposed treatments. As in Darwin's selection, the fitter clones to the current specific conditions survive, while less advantageous subclones can diminish or even disappear. The ever-changing environment of tumors underlies their ever-changing dynamics, where clones that were dominant at a time, may reach a bottleneck and be depleted, while other minor subpopulations in the past might achieve a favorable position later

on and become the dominant population. Under this model, the main evolutionary patterns that have been recognized in tumors are linear evolution and branched evolution (A. Davis et al., 2017). In the former, mutations are acquired linearly in a step-by-step process, and new driver mutations have such strong selective advantages that result in clonal sweeps where they outcompete all other subclones and become dominant. In the latter, there is a coexistence of distinct subclones that diverge from a common ancestor and evolve independently. These different clonal lineages can harbor different driver mutations that can promote their expansion. Finally, evidence of parallel evolution has been observed for some tumor suppressor genes, suggesting that inactivation of the same gene can occur multiple times, likely driven by selective pressures.

Even though this Darwinian process can explain the history of tumors to some extent, this model is not sufficient to encompass the full spectrum of cancer evolutionary trajectories, and other non-Darwinian mechanisms have been described to explain tumor evolution. Darwin's gradualism does not consider evolution by one-hit catastrophic events (see Introduction - section 1.2.3.3) that bring about multiple genetic alterations at the same time. In cancer, such macroevolutionary events, including chromoplexy or chromothripsis, have been described and can drive tumor initiation and progression (Baca et al., 2013; Stephens et al., 2011). Another feature of tumor evolution that is not supported by Darwin's theory is neutral evolution. Cancers emerge from a single cell, and neutral mutations within this cell or the first cell divisions are present at high frequencies in the final population, regardless of the action of selection. In the same way, each subclone originates from a single cell, and early neutral mutations are found in a large proportion of the subclone cells. In this mode of evolution there is no selection during the lifetime of the tumor, and random mutations accumulate over time leading to its extensive ITH. Evidence for neutral evolution

has been reported for some tumor types (Caravagna et al., 2020; Williams et al., 2016), however this hypothesis has been countered by others (Tarabichi et al., 2018).

The mutational landscape of the disease and its subclonal composition can define the evolutionary forces that will drive progression and/or treatment resistance. Tremendous levels of heterogeneity can be the culprit that fuels this evolutionary process. When strong selection pressures are applied upon this ITH, the disease can evolve to more aggressive forms, which sets one of the central obstacles to curative therapy. Tumors are a composition of an admixture of cell subpopulations, which challenges therapeutic approaches and calls for procedures that consider tumors not as one disease entity but rather thousands of variations of this disease. Moreover, therapy can induce novel mutagenesis (Pich et al., 2019, 2021) and/or accelerate the clonal expansion of more aggressive and resistant clones. As an example, subclones carrying lesions that hamper DNA repair are resistant to chemoimmunotherapy, and they are thus selected by it. It has also been seen that clonal evolution is more frequent in tumors receiving chemoimmunotherapy than in treatment-naïve tumors, where the clonal architecture can be in equilibrium (Landau et al., 2015).

Tumor evolution has important clinical and therapeutic implications. Hence, identifying the underlying mechanisms of evolution to therapeutic response and resistance is of the utmost relevance to inform and better design cancer treatments and clinical trials.

1.4 Fostering large-scale cancer research and its translation into the clinics

Many bioinformatics applications have been implemented to handle small size projects. As data grows, analysis becomes computationally expensive, both in terms of cost and time, and methodologies to investigate these expanding datasets must be adapted to enable genomic analysis at scale. To pave the road to large-scale genomic projects, cloud computing is emerging as a key infrastructure to handle distributed or federated datasets and analysis that would otherwise be very time and resource consuming. Genomics researchers are also increasingly relying on academic and commercial clouds to accomplish cost-effective large-scale analyses and examination of big datasets bypassing the need of local infrastructures and data transfer burden.

Large-scale genomic data analysis has to deal with numerous bottlenecks arising at different levels, from ever growing data sizes, computationally demanding algorithms, up to the need of data sharing among research and clinical communities. Many of these challenges have been addressed by the leading cancer genomic initiatives described in the following sections.

Cancer genomic studies, included within large-scale initiatives or from independent projects, have identified potential cancer drivers and its role in the formation and progression of tumors. Side by side, there have been works on centralizing all this scientific knowledge to curate and present it in an organized manner. In this direction, several databases and public resources have been developed.

1.4.1 Catalogs of sequence variants

Numerous databases collecting genomic variation data have been developed throughout the years, some include all kinds of variant annotations while others can be cancer specific. There are population-based databases, such as dbSNP (Sherry, 2001), the Exome Aggregation Consortium (ExAC) (Lek et al., 2016), or the Genome Aggregation Database (gnomAD) that include population frequency data of the variants identified within each resource. Other databases report the pathogenicity of variants, such as the UniProt Humsavar database (https://www.uniprot.org/docs/humsavar), or dbNSFP (Liu et al., 2016), a compendium of non-synonymous SNPs and their functional predictions.

Following with the functional annotations, some resources include additional information found in the literature and curated annotations. ClinVar (Landrum et al., 2018) is a well-known archive at the National Center for Biotechnology Information (NCBI). It is freely available and provides information for interpretation and clinical significance of both germline and somatic variants. CIViC (Clinical Interpretation of Variants in Cancer) (Griffith et al., 2017) is a community-driven resource for the clinical interpretation of variants in cancer. It is open source, offers open access, via a web portal or public application programming interfaces (APIs), and provides accurate annotations with provenance of supporting evidence.

Probably one the most well-known databases of somatic mutations is The Catalog of Somatic Mutations in Cancer (COSMIC), which started with data from four genes (Bamford et al., 2004), and has grown to include almost 6 millions of coding mutations across 1.4 million cancer samples (Tate et al., 2019), together with non-coding mutations, copy-number alterations, gene-fusions, and the largest catalog of mutational signatures. In parallel, a curated catalog of genes

driving cancer, the Cancer Gene Census (CGC), is also available. The database can be interrogated through web pages that support graphical and tabular views of the results.

Besides these laborious catalogues of variants and their annotations, other resources are also focused on the visualization of the data. The cBioPortal (https://www.cbioportal.org/) was designed to integrate data from different platforms to explore it and perform data analytics in an easy manner (Gao et al., 2013). It provides a web resource for searching, visualizing, and analyzing multidimensional cancer genomics datasets. Researchers can interactively explore genetic alterations, with multiple graphical summaries, and link them to clinical outcomes. The portal includes data of already existing projects, curated scientific results, and can also be installed locally to work on regional data.

While some resources are mainly focused on assembling and handling already generated and published data, some can also be used to provide information on new variants. The Cancer Genome Interpreter (CGI) (Tamborero et al., 2018) is a platform that systematizes the interpretation of tumor genomes. It uses current knowledge and evidence to interpret newly identified variants, annotating potential driver alterations and their possible association to treatment responses.

Some initiatives are specifically focused on personalized medicine in oncology, including diagnostic and prognostic information, clinical trials, and therapy response, like the Personalized Cancer Therapy (PCT) (https://pct.mdanderson.org) at the MD Anderson Cancer Center that compiles and integrates scientific knowledge on cancer alterations and their implication for cancer therapy, the Jackson Laboratory Clinical Knowledgebase (CKB) (S. E. Patterson et al., 2016), or OncoKB (Chakravarty et al., 2017). The Molecular Tumor

Board Portal (MTBP) is a clinical decision support system that unifies genomic analyses from European cancer centers within the Cancer Core Europe (CCE) network (Eggermont et al., 2019). The portal automates capture, interpretation, and reporting of data across the CCE sites, and is used to select candidates for on-going clinical trials. Additionally, they offer a public resource for investigators outside the network that provides a framework for classifying the functional and predictive relevance of a set of variants. Variants are categorized according to up-to-date evidence from the integration of expert-curated knowledge bases, bona fide biological assumptions, and bioinformatics predictions. Next, functionally relevant variants are associated with biomarkers of disease diagnosis, prognosis, and therapy response as reported by some of the aforementioned databases (Chakravarty et al., 2017; Griffith et al., 2017; Tamborero et al., 2018).

Altogether, these repositories aim to facilitate the interpretation of variants in clinical and research settings. Besides these knowledge-based resources, dedicated to gather and curate scientific knowledge, other initiatives also promote and coordinate research projects from the raw data generation up to the compilation of their results, making them easily and homogeneously accessible. These projects themselves come with data portals to access the comprehensive catalogs of genomic alterations in cancer that they generate.

1.4.2 Large consortia and international and national initiatives

Large-scale tumor sequencing efforts have been led by big consortia, such as the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) or The Cancer Genome Atlas (TCGA) (Ellrott et al., 2018). They have been promoting cancer research, organizing diverse cancer projects, and collecting their results into comprehensive catalogs of genomic alterations. Their shared mission is to launch and coordinate numerous cancer research projects organized in a

collaborative framework, including tens of cancer types. The ICGC is a multinational consortium involving many countries around the world, where each project specifically analyzes one cancer type (Figure 26), whereas the TCGA is based and coordinated within the United States. Projects within their umbrella adhere to agreed requirements in terms of ethical approval, sample quality, minimal clinical annotation, and data sharing. The full inventory of somatic mutations, clinical information, additional analysis, and raw data is homogenized and organized into databases that can be examined via user-friendly web portals.

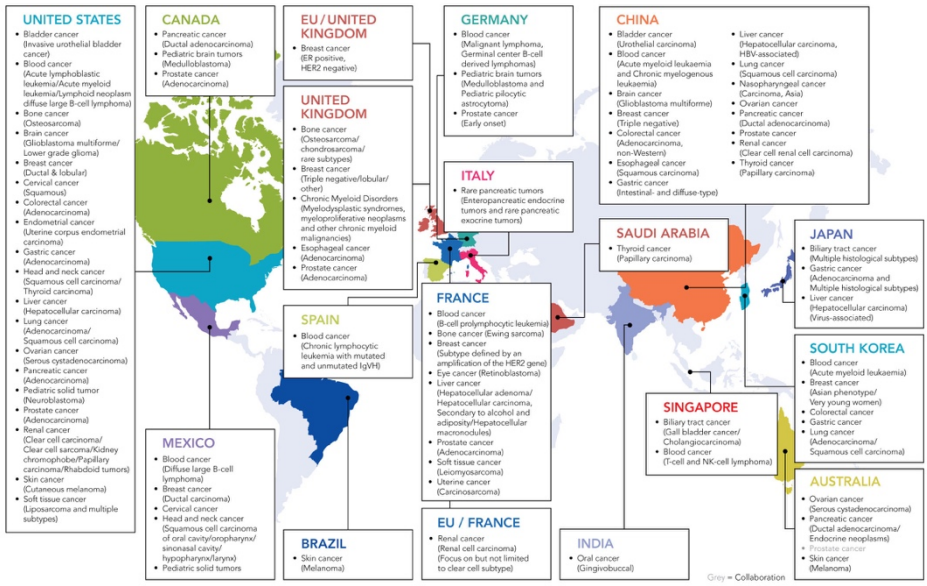


Figure 26. The international cancer genome consortium. Countries and projects that contributed to the final release in 2018.

As a natural evolution of these large-scale tumor sequencing projects, which generate sequence data from thousands of tumors, the ICGC launched a new worldwide initiative, the Pan-Cancer Analysis of Whole Genomes (PCAWG) (Campbell et al., 2020). Within this new phase, more than 2,600 normal-tumor genome pairs across 38 cancer types and from 14 jurisdictions were to be jointly

analyzed to answer questions related to the causes, formation, and evolution of tumors, together with the prevention, diagnosis, and treatment of cancer. This project was set to be the most comprehensive analysis of cancer whole genomes up to date and required the setting up of infrastructures capable of performing large-scale analyses, supporting storage of vast amounts of data, and providing computational and data access to the researchers. The design, implementation, and execution of the project is described in the corresponding Results section (Results - Chapter 1: Study 1).

Other initiatives, more specific to the analysis part and their translation into the clinics, are also trying to standardize and inform best-practice protocols to analyze genome data. The Sequencing Quality Control 2 (SEQC2) project (Mercer et al., 2021) sought to develop reference materials to assess quality-control metrics for NGS analysis, benchmark the impact that experimental and bioinformatic factors can have, and evaluate the inter- and intra-laboratory reproducibility. Herewith, participating researchers and clinicians worked together to create consensus standards for best practices in clinical settings.

Genomic analysis is growing to be part of health care systems. The aim of any biomedical scientific discovery should be its impact to ultimately benefit patients: its translation into the clinics. The vast amount of generated data and subsequent scientific results must be pushed towards this end. Within this scope, the European infrastructure for translational medicine (EATRIS) brings together resources, expert services, research tools, and education and training programs to make the translation of scientific advances into medical products that can improve our health and life quality.

Key consortia within cancer research are also moving towards the applicability into health care. The ICGC-ARGO (Acceleration Research in Genomic

Oncology) is a new effort from the ICGC to strengthen cancer research and its translation into the clinics. Over the next ten years, ICGC-ARGO aims to coordinate the integration of homogenic genomic analysis and phenotypic data on 200,000 cancer patients. This detailed and curated clinical and genomic dataset will be used to address key clinical and biological questions regarding cancer origin, progression, and resistance to treatments. The project will gather high quality data from clinical trials and well annotated cohorts, from hospitals and data centers distributed around the world, and will make it available to the entire research community, using mechanisms for efficient and responsible data sharing that will enable collaborative and combined analysis to accelerate research into the causes and management of cancer. Overall, the project aims at fostering scientific impact and translating it into health impact. The addressed questions will be relevant to the patients, tackle unmet clinical needs on how cancer can change with time and treatment, and investigate informative molecular data for precision oncology and prognostic markers.

Following the PCAWG strategy, genomic data will be harmonized, comprehensively annotated, and homogeneously analyzed in regional data processing centers (RDPC). The RDPCs will be the foundational units where genomic data will be submitted and processed through a series of standardized and containerized pipelines. The results will be sent to the Data Coordination Center (DCC) for integration with all ICGC-ARGO data sets and distribution to the community. ICGC-ARGO comes with a full stack of software products to cover all stages of a cancer genomics project (Figure 27). The system can be used to build a genomics platform from scratch, where users can collaborate and share their results. Data lies at the center of the organization; submitters can securely upload genomic and associated clinical metadata, access components authorize users to view and download controlled data, supplementary products provide interactive

visualizations and code-based analysis environments, and researchers can share their searches and results across the scientific community.

Figure 27. Overture software stack. A collection of open-source and extendable solutions for big-data genomic science that can be used to support cancer genomic research. Image from https://www.overture.bio.

At the national level, initiatives such as MedPerCan (Medicina Personalitzada Catalunya Cancer) are also pushing translational research to address clinical needs and improve health outcomes. The MedPerCan project, from the Pla Estratègic de Recerca i Innovació en Salut (PERIS), proposed to establish a multidisciplinary circuit among different hospitals, data centers, and sequencing facilities for the implementation of personalized medicine in oncology in Catalonia (Figure 28). This project was developed in the context of research, but aimed to explore its feasibility in clinical settings, as one of the major goals was to evaluate the impact that genomic analysis can have in clinical decisions.

This pilot project evaluated the feasibility of the use genomic data for more precise diagnostics and treatment recommendations, and implemented a

prospective strategy that could be used in the public health care system. In particular, three use cases with particular unmet clinical needs were addressed: risk of hereditary cancer, first-line treatment response, and treatment selection at advanced stages of the disease. The project included different institutions with expertise in all ares of the circuit, from hospitals and clinical research institutions to sequencing and supercomputing centers, and from doctors and researchers to bioinformaticians and computer engineers. These multidisciplinary teams worked together to integrate genomic analysis with clinical information to support clinical decision-making.

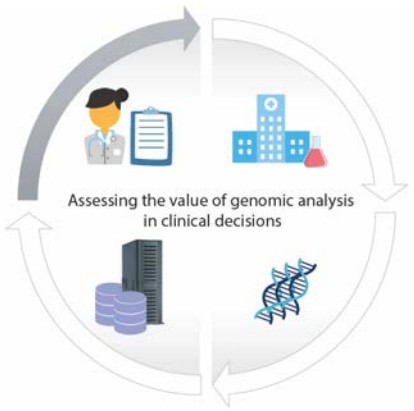


Figure 28. Circuit of the MedPerCan project. The cycle starts at the hospitals, where samples from patients are collected and sent to the sequencing center. The sequencing data is sent to the data center where it is analyzed, and the results are populated into a database accessible via a web interface. The genetic variants can then be evaluated by a panel of experts that will come up with the best clinical decisions based on the genetic make-up of the patients.

1.4.3 Infrastructures to facilitate data sharing and large-scale analyses

Together with these actions, an endeavor towards standardization and best practices for genomic analysis and data sharing is a must. In this direction, the Global Alliance For Genomics and Health (GA4GH) is a policy-framing and technical standard-setting organization for genomic analysis and data sharing

(Figure 29). This global alliance is meant to enable rapid progress in biomedicine, creating and maintaining interoperability of technical standards and harmonizing procedures for data sharing (Rehm et al., 2021). Open standards are designed to enable storage and data access, as well as to homogenize data processing tools, which will allow results from different projects to be comparable and integrable.

To carry this out, real world initiatives (so-called driver projects) present the inputs and requirements of the community and interact with the technical and foundation working streams that provide mechanisms for their implementation.

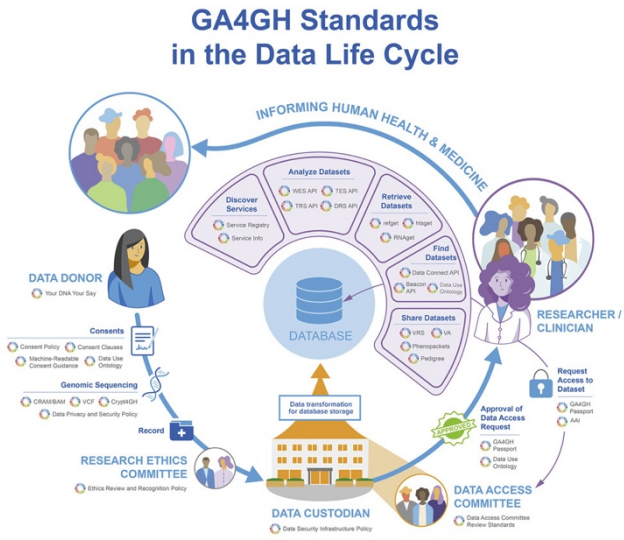


Figure 29. Standards by the GA4GH throughout the cycle of Life sciences data generation. Image from Rehm et al., 2021.

Other proposals also bring out the challenges of cross-border data sharing and their integration. The EUropean-CANadian Cancer network (EUCANCan) proposes a federated infrastructure whose mission is to enable Personalized Medicine in oncology by promoting the generation and sharing of harmonized genomic and phenotypic data. Here, the analysis of the data is handled from a

different angle, instead of reanalyzing everything from scratch, they aim to design a strategy for the evaluation of the used methods and pipelines and set a minimum quality threshold upon which the results could be safely combined. The federated infrastructure guarantees that regional data can be stored locally, while it is identifiable, searchable, and findable within the federation. The software stack used to build this model is based on the ICGC-ARGO project previously described.

To respond to the numerous datasets and resources generated by the community there are initiatives trying to promote their shareability and easy usage. The iPC (Individualized Paediatric Cure) project has the aim of providing clinicians with the tools and knowledge to create individualized treatment strategies for children with cancer. The project will collect, standardize, and harmonize existing clinical knowledge and medical data that will be used to create artificial intelligence treatment models for each patient. Researchers will then apply these models on virtual patients to evaluate treatment toxicity and efficacy to assess if they can improve patient survival and life quality. EOSC4Cancer is a European project that aims to accelerate research and innovation by providing smooth access, management, interoperability, and reuse of digital information. It will connect a set of interoperable nodes (e.g., European Cancer Centres, Research Infrastructures, and Medical Centres) that provide access to FAIRified cancer-related data within a trusted users environment(s). Their ambition is to put in the hands of clinicians and researchers the necessary means to address the different steps during the individual cancer patient journey, from prevention and diagnosis to advanced stages and treatment. Different data sources relevant in cancer research will be mobilized and interconnected. The usage of tools, data analytics, and machine learning methods will be leveraged by their integration into virtual research environments and cancer analysis portals.

It is also worth mentioning that there are other initiatives which are not cancer specific but share the overarching goal to promote data sharing as the basis to enable scientists and clinicians to better understand disease and give the best possible personalized treatments.

The European '1+Million Genomes' Initiative (1+MG) aims to make genome information of at least 1 million European citizens accessible in the EU by 2022, with both genotypic and phenotypic data available and properly linked. This promising cohort has the power to provide new striking research that might translate into improved patient management, allowing for more personalized treatments. To pursue its goal, the project will involve stakeholders with different backgrounds, from health care professionals and patient organizations, to researchers, engineers, and more. They will help national and regional authorities build a federated infrastructure, making sure ethical and legal aspects are covered to allow for genomic data sharing across borders.

To facilitate the cooperation and coordinate the signatory countries, the EU Commission granted the Horizon 2020 project Beyond 1 Million Genomes (B1MG). It will uphold the creation of a network for genetic and clinical data sharing across Europe, providing legal and technical guidance, as well as defining standards and best practices. The infrastructure will be set up for the long-term, and will go beyond 1+MG.

Efforts on standards and recommendations to guide the life cycle of health data and to accomplish responsible data sharing are also being made. The goal of HealthyCloud is to define specifications, standards, and best practices to enable health research across Europe. Stakeholders from the EU member states, including academia, industry, healthcare providers, patients' organizations and policy makers will join together to assist the generation of a Strategic Agenda for

ethical and legal good practices for use of health data, and sustainable computational resources.

1.4.4 Current challenges in cancer research

Despite many years of NGS, continuously emerging new tools, and large international initiatives and consortia, such as the ICGC, the TCGA or the GA4GH, unified or standard protocols for variant calling, the core of genomic analysis, still remain a challenge.

Cancer research and its translation into the clinics faces many challenges at many different levels. At the global level, data sharing and harmonization is a major concern, including not only legal but also methodological aspects. Both have been addressed, and are continuously dealt with, in large-scale genomic projects. These initiatives gather sizable genomic data and have identified the main obstacles in the field. They all have seen that discrepancies among variant calling tools and lack of homogeneous results complicate their integration. While some decide to re-analyze everything from scratch (Campbell et al., 2020; Ellrott et al., 2018), where extra resources and time are spent, others come up with a smoother solution where they try to harmonize different variant calls as long as they are above set thresholds of good quality data. The shortage of benchmarking datasets for somatic variant calling (Alioto et al., 2015; Griffith et al., 2015) aggravates this situation, because assessing the best variant calling pipeline is not straightforward, and the evaluation of the methods can differ greatly depending on the input data (e.g., low vs high tumor purity, low vs high coverage, low frequency variants, or FFPE samples).

Small-scale studies also deal with these problematic procedures. The lack of benchmarking datasets, together with the demanding characteristics of some tumor samples, such as low purity, FFPE archival material, lack of matched-normal

sample for tumor analysis, impedes the selection of one single strategy that can be blindly trusted, and brings out the need of bioinformatics expertise to correctly interpret the results. Bioinformatics knowledge is sometimes missing in clinical environments, where easy-to-use methods, as well as highly accurate results, are essential for clinical application.

1.5 Chronic lymphocytic leukemia (CLL)

Chronic lymphocytic leukemia (CLL) is the most common leukemia in adults in western countries, though it is less prevalent in Asian countries. The average age at diagnosis is around 70, and it is more common in men than women. People with first-degree relatives with CLL have more than twice the risk of developing this disease.

CLL is a type of cancer characterized by the accumulation of B-cells (Figure 30), a type of white blood cells, in bone marrow, lymph nodes and peripheral blood. It is thought to be preceded by monoclonal B-cell lymphocytosis (MBL), a typically asymptomatic state, in which an increased number of monoclonal B-cells is already present in blood.

CLL has been at the forefront of cancer research thanks to the accessibility of tumor samples, taken from the bloodstream, and its usually slow growing nature which provides an ideal setting for longitudinal studies. The biology of CLL is a complex that integrates factors from the cell-of-origin, the microenvironment, and DNA alterations. Several signaling pathways, and notably the B-cell receptor (BCR) pathway, have a central role in CLL development and clonal expansion of malignant clones.

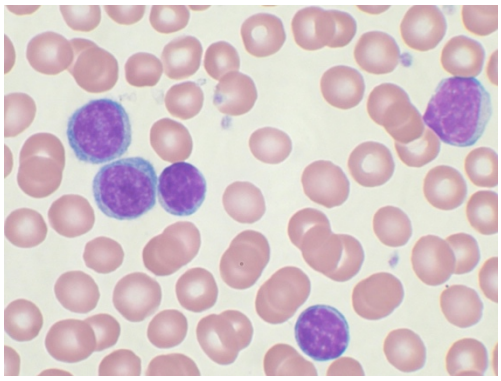


Figure 30. High-power magnification (1000 X) of a Wright's stained peripheral blood smear showing chronic lymphocytic leukemia. The lymphocytes with the darkly staining nuclei and scant cytoplasm are the CLL cells.

During the last decade fruitful findings have shed some light on the genetic susceptibility, the molecular mechanisms driving the disease, the genomic and epigenetic dysregulations, and the patterns of clonal evolution leading to progression, treatment resistance, and adverse transformation into more aggressive lymphomas. These new discoveries have the potential to be translated into the clinics by exploring new therapies and management strategies. Nonetheless, the TP53 disruption and the immunoglobulin status remain the only two biomarkers that are routinely being used in the clinics (Hallek et al., 2018).

1.5.1 Normal B-cell differentiation

Foreign agents, including viruses, bacteria, or fungi, express specific antigens that can be identified by our immune system that triggers a response to neutralize or destroy them and any other cell that has been infected. As part of our adaptive immune system, B cells go through different stages to mature and acquire full specificity to fight foreign antigens. The process starts in the bone marrow, where hematopoietic stem cells (HSCs) start differentiating into multipotent progenitors that eventually give rise to myeloid and lymphoid lineages, including B cells.

During their development, B cells present different immunophenotypes characterized by different gene expression profiles and immunoglobulin rearrangements (Figure 31). In the end, B-cell production is committed to find the most effective antigen-binding. This diversity of B-cell receptors (BCRs) is made possible by molecular and cellular mechanisms capable of generating a broad repertoire of receptor molecules and the selection of the most efficient ones for further expansion.

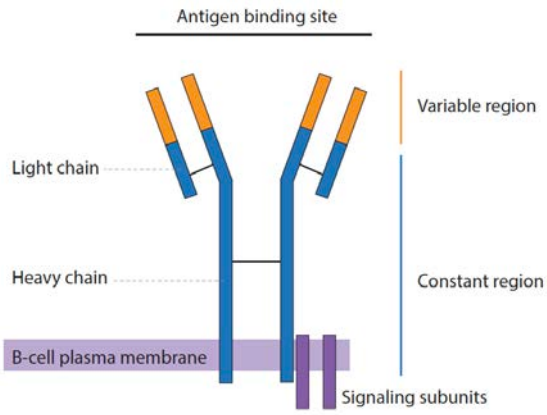


Figure 31. Simplified schema of the B-cell receptor (BCR). The immunoglobulin molecule contains two identical heavy chains and two identical light chains that are produced through genomic rearrangements during B-cell development. The antigen binding site has a variable region that can acquire higher affinity during maturation.

The first step to generate the BCR takes place in the bone marrow, where different DNA segments of the immunoglobulin (Ig) genes are joined together. These segments include the variable (V), diversity (D), and join (J) regions, which are randomly selected and rearranged from the numerous V, D, and J sequences that are available in the genome in a process called V (D) J recombination. The BCR structure is composed of two identical heavy chains (H-chain) and two identical light chains (L-chain) bonded together. Both types of chains are formed by a variable region and a constant region. The variable region, which contains the

V, D, and J segments or the V and J segments for the H- and L-chains, respectively, forms the Ig antigen-binding site, whereas the constant domains determine the Ig isotype and its functions. The two H- and L-chains compose the antigen-recognition structure of the BCR (Pieper et al., 2013). Those B cells with a functional Ig that does not bind to self-antigens leave from the bone marrow to the bloodstream.

At that point, naïve B cells can go to the secondary lymphoid organs, where they can be activated upon antigen recognition through the BCR. These stimuli induce proliferation and other processes to gain higher antigen-binding affinity. The development process will continue upon activation by an antigen, with or without the help of T cells. The T-cell dependent activation starts with the creation of the germinal centers (GCs), which are transient structures for B-cell proliferation and BCR affinity maturation. Proliferative B cells in GCs, called centroblasts, undergo somatic hypermutation (SHM) in the so-called dark zone of the GCs to diversify and fine-tune their antigen-binding capabilities. Activation-induced cytidine deaminase (AID) is the enzyme responsible for this mutagenic process and acts upon the Ig H- and L-chain genes. B cells with enhanced BCR go to the light zone where positive selection is imposed, and only those with higher ability to bind to antigens proceed to the next steps.

At this stage, B cells are called centrocytes and experience the class switch recombination (CSR), also induced by AID, which can modify their constant region to other isotypes with particular effector functions (Klein & Dalla-Favera, 2008). In the end, those B cells with a high-affinity receptor that are positively selected will differentiate into long-lived memory B cells, which can be rapidly activated upon reinfection of known antigens, or plasma cells, specialized in the production and secretion of large quantities of antibodies (Figure 32). T-cell independent activation occurs in the marginal zone of the lymphoid follicle, where B cells can

also be activated and undergo SHM and CSR, but they are short-lived, and their resulting antibodies have lower affinity (Bortnick & Allman, 2013).

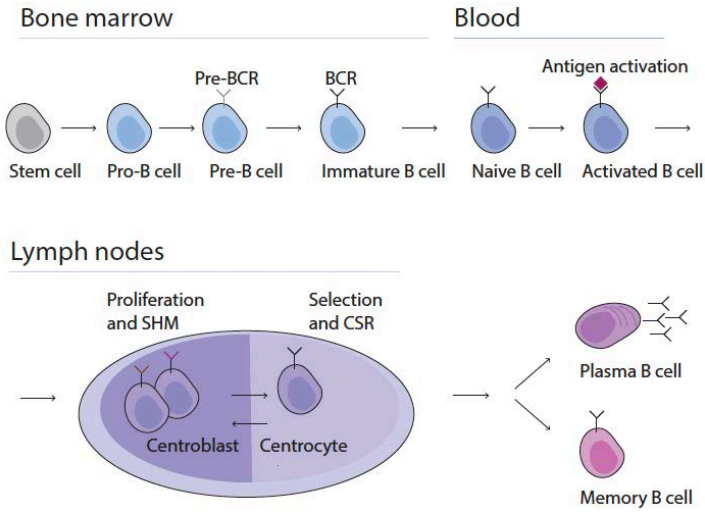


Figure 32. B-cell development stages. Starting from stem cells in the bone marrow, B cells go through different states until they mature and are differentiated into plasma or memory B cells. The compartments are indicated at the top. The mechanisms underlying the main steps are shown.

1.5.2 Genetic predisposition to CLL

Population and family based studies have reported a 7.5- or 8.5-fold increased risk for first-degree relatives of CLL patients (Cerhan & Slager, 2015; Goldin et al., 2004). Possible models of genetic inheritance have been described for sporadic and familial CLL: if there is only one member of the family with CLL, it is likely sporadic or associated with a low risk allele; if there is a second family member it is less likely to be sporadic and might be associated with low risk alleles; when there is a higher number of affected family members it is likely that there is a high risk allele (Sellick et al., 2006).

High risk alleles are very rare, linkage or family-based, and probably involve a gene that is involved in the disease (Brown et al., 2012). Low risk variants can be

identified with GWAS, and up to 45 risk loci have been associated with CLL (Law et al., 2017). Most of these SNPs are in the non-coding regions of the genome, and their mechanisms are being elucidated by integrating genome-wide sequencing, transcriptomics, and epigenomics (Speedy et al., 2019). A lot of them are associated with active regulatory elements, suggesting that they might have a role in gene expression. In CLL, risk loci do not only influence the risk of developing CLL, but also the outcome of the disease. More recently, risk loci for progressive CLL have been described (Lin et al., 2021), illustrating the impact that germline variants can have not only on getting the disease, but also on the clinical phenotype.

NGS WES studies have also explored germline variation in CLL. Family-based studies have identified genes commonly altered in families: *POT1* and other shelterin complexes (Speedy et al., 2016), *ITGB2* (Goldin et al., 2016), and *NFATC4* (Itchaki et al., 2017). Rare variants in sporadic CLL cases versus controls were found to be enriched in *ATM* and *CDK1*, with frequent loss of the normal allele of *ATM* (Tiao et al., 2017).

1.5.3 Cell-of-origin and molecular subtypes

The cell(s) of origin of CLL, i.e., the non-malignant cell from which malignancy develops, have not yet been fully characterized. Hematopoietic stem cells (HSCs) might already acquire some of the earliest changes that can lead to clonal expansions of CLL-like cells (Kikushige et al., 2011), and common genetic alterations of CLL, such as trisomy of chromosome 12 (tri12), deletion of chromosome 13q [del(13q)], and mutations in driver genes *SF3B1*, *NOTCH1*, and *XPO1* have been found in the hematopoietic progenitors of some patients (Damm et al., 2014; Gahn et al., 1997). These findings suggest that even though CLL is a neoplasm of mature B cells, malignant transformation starts at earlier stages of B

lymphocytes differentiation. The mechanisms leading these initial steps to fully oncogenic mature B cells are not well known, but immunogenetic analyses of the BCR found antigen selection to have a major role in driving tumor cells' clonal selection (Agathangelidis et al., 2012; Stamatopoulos et al., 2016). This observation is supported by the striking bias in the use of some immunoglobulin heavy-chain variable region (IGHV) genes, the essentially identical complementary-determining region 3 (CDR3) in some cases, and the highly homologous Ig rearrangements, named stereotypes, which account for 30% of the patients.

Traditionally, CLL is classified into two main molecular subtypes that are based on the mutational status of the IGHV and define two different entities (Figure 33), with particular genomic and epigenomic alterations and distinct clinical course (Damle et al., 1999; Hamblin et al., 1999). CLL expressing mutated IGHV (M-CLL) have a more indolent behavior and originate from B cells that have gone through the germinal center, whereas CLL presenting unmutated IGHV (U-CLL) are more aggressive and derive from pre-germinal center B cells (Seifert et al., 2012). The assessment of the IGHV mutational status is routinely performed in both research and clinical settings, where Sanger or NGS protocols with a cutoff of >= 98% identity between tumor and germline sequences are used. Its prognostic and predictive value underlies its clinical importance, where evaluation of the IGHV status is recommended at diagnosis or before treatment initiation (Rosenquist et al., 2017).

Concordantly, epigenetic studies have reported that M-CLL has a methylation signature of a normal post-germinal center cell (i.e., memory-like), and U-CLL maintains a naïve-like methylation signature, corresponding to a cell that maturated outside of the germinal center (Kulis et al., 2012; Oakes et al., 2016). Interestingly, a third intermediate epigenetic group was identified with a

methylation profile between naïve-like and memory-like CLLs, suggesting that it could derive from a not-yet-identified normal B cell. The 3 subtypes present different mutational profiles, usage of IGHV genes, and clinical outcomes (Kulis et al., 2012; Puente et al., 2015; Queirós et al., 2014).

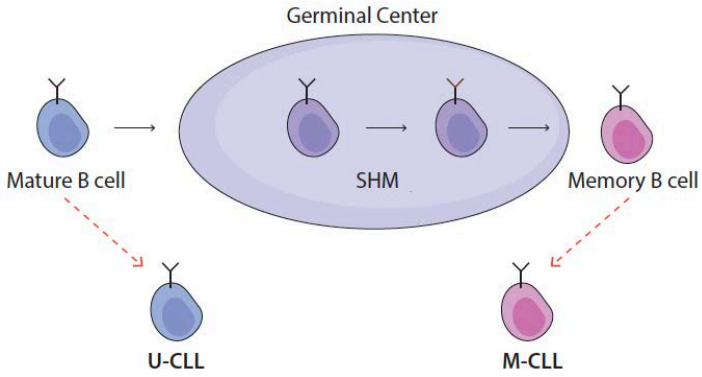


Figure 33. Schematic view of the two main molecular subtypes of CLL. M-CLL is derived from antigen-experienced B cells that have gone through the germinal center, where immunoglobulin somatic hypermutation takes place. Contrarily, U-CLL originates from a pre-GC mature B cell.

All these studies support that the cell of origin, determined by its immunogenetic or epigenetic profile, can be an important determinant of the clinical and biological behavior of the tumor.

1.5.4 Landscape of genomic alterations in CLL

Genomic sequencing of the last ten years has unraveled the mutational landscape of CLL. Even though recurrent mutations have been identified, CLL remains a very heterogeneous disease from the genomic point of view, and the tail of low frequency recurrent variants and affected genes is notably long. Furthermore, the genetic profile can change and evolve from early to relapsed refractory patients, increasing the complexity and heterogeneity of the disease.

Up to date, WGS- and WES-based studies include more than 1,000 cases, and identified an average of 2,500 mutations per tumor (Puente et al., 2015), which correlates with the mutational status of IGHV (M-CLL, despite their better outcome, have a higher number of somatic mutations). The mutational processes contributing these mutations are mainly related to aging and to the activity of AID (Alexandrov et al., 2020; Kasar et al., 2015; Puente et al., 2015). The noncanonical AID (nc-AID) signature is responsible for mutations in M-CLL and has also been seen in other lymphoid neoplasms derived from cells that have germinal center experience.

CLL's low mutational burden is accompanied by few chromosomal translocations. The most common cytogenetic alterations are 13q deletions (including miR15 and miR16 genes within the minimal deleted region), 11q deletions (encompassing *ATM*), 17p deletions (including *TP53*), and trisomy 12. Their strong correlation with patients' prognosis brought them into clinical use. Patients carrying del(11q) or del(17p) have a significantly worse outcome than those harboring del(13q) (Döhner et al., 2000). The most common translocations involve IGH and different oncogenes and are present in a very small percentage of cases (Figure 34).

The coding landscape of CLL is characterized by few recurrently mutated genes, present at most in 10-15% of cases (Figure 35). The most recurrently mutated genes at CLL diagnosis are *NOTCH1* (8–12%), *SF3B1* (9–11%), *TP53* (5–8%), and *ATM* (5–7%), but the frequencies can vary in CLL progression, reflecting their impact on the clinical evolution of the disease (Landau et al., 2015; Nadeu et al., 2018).

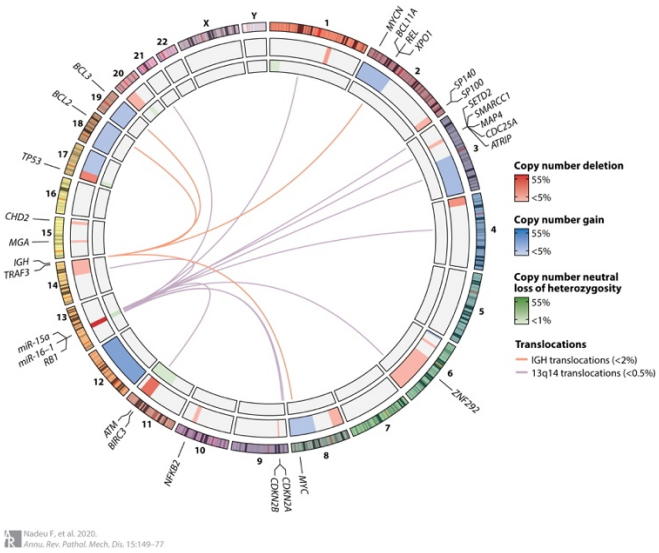


Figure 34. Recurrent chromosomal alterations in CLL. Summary of most recurrent chromosomal rearrangements and target genes. The outer circle shows the chromosomes, followed by copy number alterations (deletions in red, gains in blue, and loss of heterozygosity in green). The intensity of the color is proportional to the fraction of patients carrying each alteration. Translocations are drawn in the inner circle linking together different regions of the genome. Image from Nadeu, Diaz-Navarro, et al., 2020.

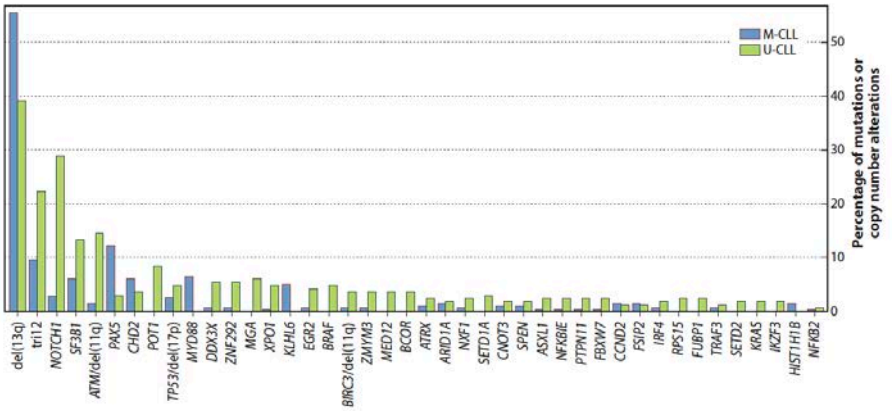


Figure 35. CLL driver alterations. Frequency of the most common driver alterations at CLL diagnosis separated by IGHV mutational status. Image adapted from Nadeu, Diaz-Navarro, et al., 2020.

These driver genes converge into key deregulated pathways (Figure 36), including NOTCH1 signaling (*NOTCH1* and *FBXW7*), DNA damage response and genomic stability (*ATM*, *TP53*, and *POT1*), RNA splicing and metabolism (*SF3B1*, *U1*, *XPO1*, *DDX3X*, and *RPS15*), NF-κB signaling (*BIRC3*, *NFKB2*, *NFKBIE*, *TRAF2*, and *TRAF3*), B-cell receptor and Toll-like receptor signaling (*EGR2*, *BCOR*, *MYD88*, *TLR2*, *IKZF3*, and *KRAS* or *NRAS*), and chromatin modifiers (*CHD2*, *SETD2*, *KMT2D*, *ASXL1*).

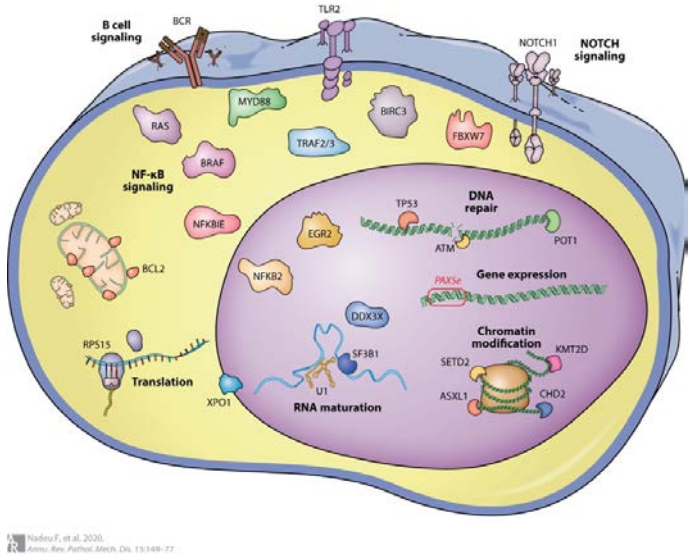


Figure 36. Main disrupted molecular pathways in CLL. Illustrative view of the main cellular pathways affected by CLL mutations in driver genes. Image from Nadeu, Diaz-Navarro, et al., 2020.

NOTCH1 is a known oncogene that encodes a transmembrane protein whose signaling cascade produces *NOTCH1* intracellular domain (NICD) that, once in the nucleus, forms a transcription complex that can switch on the expression of target genes involved in processes related to cell fate, differentiation, proliferation, and survival (Andersson et al., 2011). *NOTCH1* emerged as one of the most mutated genes in CLL, and its frequency rises with disease progression (Rosati et al., 2018). As such, it is associated with worse prognosis (Sportoletti et al., 2010). Recurrent

mutations disrupt or result in the loss of its PEST domain required for the degradation of the NICD. Moreover, even CLLs without *NOTCH1* mutations can accumulate NICD within the nucleus and show similar *NOTCH1* expression signatures (Fabbri et al., 2017). The mechanisms behind this phenomena are not fully understood, but *FBXW7*, a negative regulator of *NOTCH1*, is mutated in a subfraction of patients where it can stabilize NICD and is associated with increasing levels of downstream genes (Close et al., 2019).

TP53 and *ATM* are tumor suppressor genes and key elements in the DNA damage response pathway. Deletions of their loci (17p and 11q, respectively) often co-occur with mutations. They are associated with higher genomic complexity and confer poor prognosis (Campo et al., 2018; Ouillette et al., 2010; Stankovic & Skowronska, 2014). They are also implicated in resistance to chemotherapy, but not to novel agents (Brown et al., 2017).

POT1 encodes a component of the shelterin complex of the telomeres, and its mutations often occur in the domains that bind to telomeric DNA, leaving CLL-mutated cells with lots of telomeric and chromosomal abnormalities, and conferring adverse prognosis (Ramsay et al., 2013).

Some mutations in genes participating in RNA splicing and metabolism can trigger a cascade of events that result in altered mRNA transcripts and proteins that may conduct the pathogenesis of the disease. *SF3B1* encodes a subunit of splicing factor 3B and, when it is mutated, it can lead to mis-splicing near the 3' splicing sites in numerous genes involved in DNA damage response, telomere maintenance, and NOTCH1 signaling (Mansouri et al., 2013; L. Wang et al., 2016). A murine model with *SF3B1* mutated and *ATM* deletion showed that both alterations are necessary to overcome cellular senescence, induced by *SF3B1* mutated, and generate a CLL-like disease in elderly mice. These CLL-like cells show

genomic instability and dysregulation of multiple cellular processes associated with CLL, such as BCR signaling, which is decreased in *SF3B1* mutated cells, and they are more sensitive to BTK inhibitors (Yin et al., 2019). CLL also carries other mutated genes involved in RNA transport and splicing, *XPO1* and *DDX3X*, but their functions are not fully understood. A recently discovered recurrent mutation in the U1 small nuclear RNA (a noncoding component of the spliceosome) has been identified in CLL. The mutation creates novel splice junctions and alters the splicing pattern of multiple genes, including known cancer driver genes (Shuai et al., 2019). *RPS15* encodes a protein of the 40S ribosomal subunit, which acts as a nuclear export factor of this ribosomal component. Alterations in this protein induce changes in global protein synthesis and translational fidelity, affecting translational machinery and cell metabolism (Bretones et al., 2018; Ljungström et al., 2016).

The pathogenesis of CLL is also modulated by constitutive activation of nuclear factor-κB (NF-κB) signaling, which regulates important cellular processes linked to cancer progression, cell survival, and proliferation. Only a few recurrently mutated genes have been identified, including *NFKB2*, *TRAF3*, and genes encoding BIRC3 and NFKBIE, inhibitors of the noncanonical and the canonical NF-κB pathway, respectively. However, NF-κB can also be mediated by upstream cell surface receptors like BCR or toll-like receptors (Mansouri et al., 2016).

BCR signaling has a pivotal role in CLL pathogenesis. It is required for the survival of mature B cells and of most neoplastic mature B cells. Upon activation, downstream pathways lead to cell survival and proliferation (Kipps et al., 2017). Contrary to other lymphoid neoplasms, activation mutations are uncommon in CLL. *EGR2* encodes a transcription factor (TF) downstream of the BCR pathway. Its activating mutations are associated with aggressive forms of the disease and

confer poor outcome (Young et al., 2016). Toll-like receptor signaling activation in CLL is mediated by *MYD88* and *TLR2* mutations that increase interleukin (IL)-6 and IL1RA levels, suggesting that they may promote a favorable microenvironment for tumor cell survival (Beà et al., 2013; Puente et al., 2011).

Mutations in chromatin remodeler genes, capable of modifying the epigenomic landscape of tumors, are less common in CLL than other lymphoid neoplasms. *CHD2* encodes a protein that binds to histone marks involved in transcription. Its mutations confer defective association with active chromatin and change the transcriptomic profile of the tumor (Rodríguez et al., 2015). *SETD2* encodes a histone methyltransferase, and it is postulated as a tumor suppressor gene. Loss-of-function mutations are associated with other poor prognosis alterations and poor outcome for the patients (Parker et al., 2016).

These discoveries have not only identified recurrent common alterations in genes and pathways associated with the pathogenesis of CLL, but have also shown how specific genetic alterations can serve as biomarkers for prognostication and prediction of response to therapies. Moreover, the identified disrupted cellular pathways can be potential targets for new therapies. Lastly, all these genetic insights can be used to guide novel treatment algorithms for the clinical management of patients. Nonetheless, the vast heterogeneity among CLL tumors poses the need of large cohorts where the impact of genes mutated at low frequencies can be further characterized.

1.5.5 Clonal dynamics

Early studies based on NGS allowed us to come forward in terms of genetic alterations, transcriptomic dysregulations, and epigenetic changes of CLL. Our understanding of CLL heterogeneity has also evolved over time, from differences between affected individuals to the genetic variability found within individual

samples (Gruber & Wu, 2014). As in any other cancer type, CLL comprises not a monolithic population, but rather an admixture of subpopulations. The significance of what is happening during the evolution of the disease relies mainly on concerns about progression, relapse, and resistance to treatment. The presence of subclonal driver alterations has been identified as an independent risk factor for disease progression in WES-based studies (Landau et al., 2013, 2015). The deconvolution of genetic alterations according to their clonality and their categorization into clonal and subclonal events, allows the temporal reconstruction of the acquisition of mutations into early or late events, respectively.

In CLL, the initiating events are generally CNAs, including del(13q), tri12, and del(11q), and mutations in driver genes such as *SF3B1*, *POT1*, *TP53*, *ATM*, *NOTCH1*, or *BIRC3*, occur later in the course of the disease (Landau et al., 2015; Nadeu et al., 2018). Longitudinal studies based on WES or WGS have confirmed these findings and have identified three different patterns of CLL evolution (Figure 37): stable equilibrium (subpopulations are maintained over time), linear evolution (alterations are sequentially acquired in a single clone), and branched evolution (different subclones coexist and evolve). ITH entails that multiple tumor subpopulations can coexist together and, in the presence of selective pressures such as treatment, this composition can change, leading to clonal sweeps where resistant subclones can become dominant. Nonetheless, the presence of mutations with competitive advantages, such as *SF3B1,* can promote clonal changes and progression of the disease even before any treatment (Nadeu et al., 2016; Schwaederlé et al., 2013). After chemotherapy, which has been in the frontline of CLL treatment for many years, resistant subclones (i.e., harboring *TP53* mutations) can become dominant. Following treatment lines can again

impose selection barriers, leading to more clonal changes and expansion of the fitter clones.

Several works have identified genetic alterations that lead to therapy resistance (Herling et al., 2018; Woyach et al., 2014), and have also found that often the predominant population at the time of relapse can be tracked back as a small or minute subpopulation before treatment initiation with very sensitive techniques (Burger et al., 2016; Landau et al., 2015, 2017). This points to the idea that the capacity for evolution is already present at the time of treatment initiation and is certainly a conundrum for CLL therapy management.

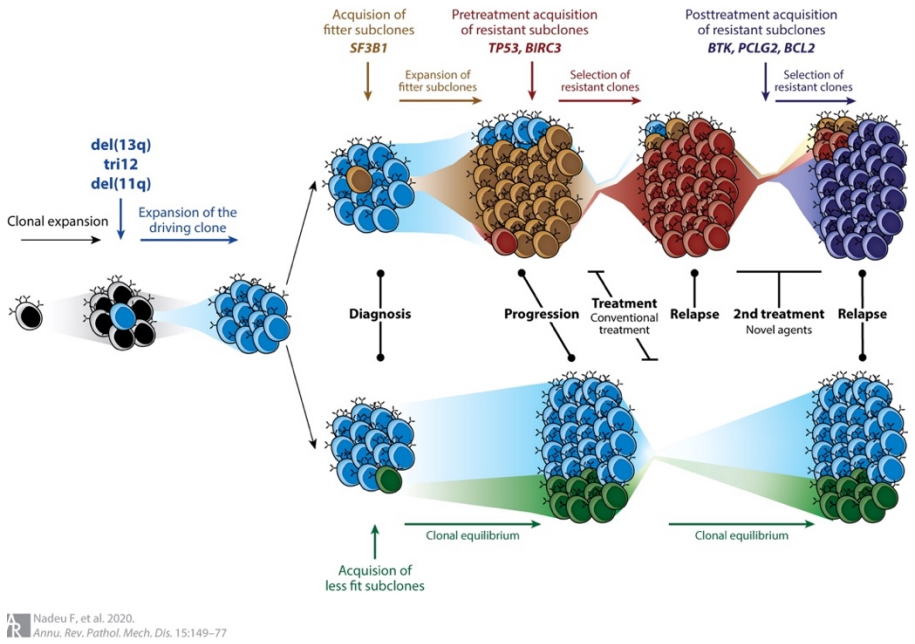


Figure 37. Patterns of evolution of CLL and sequential acquisition of driver alterations through the course of the disease. CLL starts with the expansion of a subclone harboring early alterations. The presence of mutations with proliferative advantages can lead to clonal shifts before treatment, in the same way that resistant subclones can become dominant after treatment regiments (Top). Tumors with less advantageous mutations may be more stable and maintain subclonal equilibrium before and after treatment (Bottom). Image from Nadeu, Diaz-Navarro, et al., 2020.

New capabilities of single-cell sequencing provide great opportunities to further dissect CLL heterogeneity and evolution. They provide an unprecedented opportunity to better understand how mutations, activation states, and protein expression have an impact on disease. Getting down to single cells can inform us about the composition and functional aspects of the cells that are contributing to the set phenotype and can be exploited to predict patient-specific dynamics. Multiple studies have begun to explore CLL at the single-cell level (Gohil & Wu, 2019), and give hope for a better characterization of leukemic cells in aid of better prognostication, earlier diagnosis, and treatment optimization.

1.5.6 Treatment advances and clinical challenges

CLL has a very heterogeneous clinical course, ranging from patients that have an indolent disease and might not need any treatment for years, to others that suffer more aggressive forms of the disease or whose CLL even transforms into a deadly neoplasm. The watch-and-wait approach is the current standard of care for patients without symptoms, and some people can be managed solely with this surveillance for years before the disease progresses.

CLL treatment decisions are based on the symptoms and the status of few genetic aberrations and the IGHV mutational status, associated with risk of progression and response to traditional treatments (Hallek et al., 2018). In the clinics, the prognosis of CLL is mainly assessed based on IGHV mutations and *TP53* disruption (Patel & Pagel, 2021), which predicts a more aggressive disease course, is a less favorable prognostic marker for chemoimmunotherapy (Döhner et al., 2000; Zenz et al., 2010), and is best treated with novel therapies (Hallek, 2019).

B-cell receptor signaling has a pivotal role in B cells survival and proliferation, and in many B-cell malignancies (Stevenson et al., 2011). Hence, it is not surprising that it has become a target for therapeutic intervention. Novel agents that inhibit

Bruton's tyrosine kinase (BTK), the apoptosis regulator B-cell leukemia/lymphoma 2 (BCL-2), and phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta (PI3Kδ) have been approved for CLL in the recent years (Byrd et al., 2013; Furman, Sharman, et al., 2014; Roberts et al., 2016) (Figure 38). BCR stimulation triggers the activation of BTK, which is involved in the regulation of cell migration, adhesion, survival, and proliferation. PI3Kδ is another key kinase downstream of BCR. BTK inhibitors have come to the frontline of CLL treatment thanks to their higher efficacy with high-risk patients and the more favorable toxicity compared with chemoimmunotherapy (Scheffold & Stilgenbauer, 2020).

Ibrutinib is an irreversible BTK inhibitor that forms a covalent bond to the target Cys-481 in the active site of BTK (Burger & Buggy, 2013). It is approved for both first line and relapsed/refractory disease stages.

Approved PI3K inhibitors include duvelisib and idelalisib, approved in 2014 for relapsed/refractory CLL patients. Idelalisib showed efficacy in patients with *TP53* disruption, but it was associated with worse adverse effects than ibrutinib or venetoclax (Lampson et al., 2016) and it is not the first option to treat patients. Duvelisib also showed discontinuation of treatment due to adverse effects in a fraction of cases (Flinn et al., 2018).

BCL-2 is an apoptosis regulator that has anti-apoptotic properties and is overexpressed in CLL cells, promoting their survival. Venetoclax is a BCL-2 antagonist that can disrupt this survival mechanism. It was first approved for CLL with 17p deletion but is now also approved as a chemotherapy-free combination regimen for previously untreated CLL patients by the FDA.

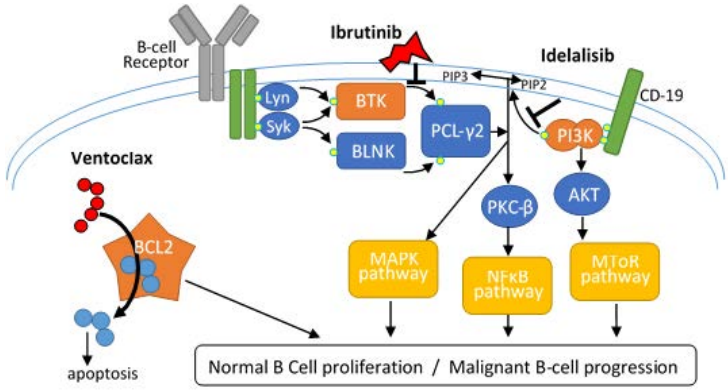


Figure 38. Representation of the B-cell receptor signaling pathway and mechanisms of action of Bruton's tyrosine kinase (BTK) inhibitor (ibrutinib), phosphatidylinositol 3-kinase (PI3K) inhibitor (idelalisib) and BCL-2 inhibitor (venetoclax). All three novel agents inhibit key molecules of BCR signaling and B-cell proliferation. Image from E. Ezekwudo et al., 2019.

Relevant advances in our understanding of CLL biology together with the approval of new targeted therapies have greatly benefited the outcome of high-risk patients. But despite these advances in treatment options and their improvement on the management of high-risk patients, CLL still remains an incurable disease with unmet clinical needs, especially in the case of an extreme form of evolution into an aggressive lymphoma, the Richter transformation.

1.5.7 Richter transformation

CLL is commonly an indolent neoplasia of mature B-cells but, in some cases, it can not only progress more rapidly and confer worse prognosis, but also transform into a high-grade B-cell lymphoma known as Richter transformation (RT), which is associated with a dismal clinical outcome, with an overall survival of less than one year. Back in history, this histopathological phenomenon was first described by Maurice Richter in his article in the *American Journal of Pathology* in 1928 (Richter, 1928) as a transformation of CLL into a more aggressive lymphoma,

and has thus been termed Richter syndrome (RS). The incidence of RT in treatment-naïve patients is rare, but it is found in up to 10% and 20% of cases after chemoimmunotherapy (CIT) and targeted therapies, respectively (W. Ding, 2018). Even though these novel small molecule inhibitors have significantly improved the outcome for CLL patients, the prognosis of RT is extremely poor, and it remains an incurable and deadly disease with urgent clinical needs.

Although different types of transformation, including Hodgkin lymphoma (HL), plasmablastic lymphoma, and other rare lymphomas have been reported, the majority of RT shows the same histological characteristics as diffuse large B-cell lymphoma (DLBCL) (Figure 39), but the molecular profile and clinical course of RT is distinct from de novo DLBCL and shows an intermediate genomic complexity between CLL and DLBCL (Fabbri et al., 2013). Transformation of CLL to a clonally related DLBCL, as assessed by the identity of the IgH rearrangement, accounts for the majority of cases and has very poor prognosis, whereas the development of DLBCL unrelated to the prior CLL clone has an outcome similar to de novo DLBLC (Rossi et al., 2011).

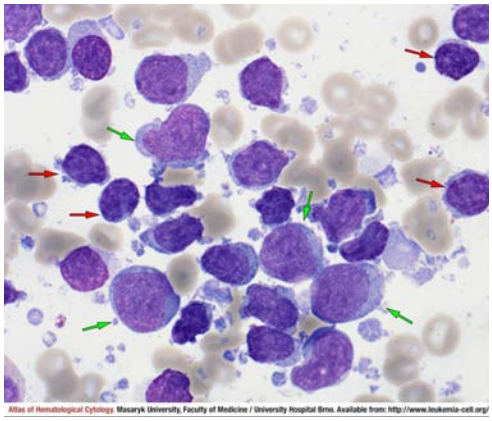


Figure 39. Chronic lymphocytic leukemia transformation to diffuse large B-cell lymphoma (DLBCL), Richter's transformation. BM MGG (1000×). Image from https://www.leukemia-cell.org/atlas.

The genomic mechanisms underlying RT are explored mainly in cases after CIT and mostly using WES or targeted approaches (Beà et al., 2002; Chakraborty et al., 2021; Chigrinova et al., 2013; Fabbri et al., 2013; Klintman et al., 2021; Rossi et al., 2011; Scandurra et al., 2010). Recurrent alterations at transformation are the deletion of *CDKN2A*, which has not been described in CLL, *TP53* disruption, *NOTCH1* mutations, *MYC* translocations or amplifications, and other less recurrent cytogenetic alterations (Figure 40). Risk factors for the development of RT have been studied and include clinical characteristics and molecular and genetic changes. Associated high-risk genomic aberrations have been identified, including CLL carrying subset #8, *TP53* disruption, *MYC* activation, trisomy 12 (particularly in the absence of del13q14) and *NOTCH1* mutations (Rossi et al., 2009, 2012).

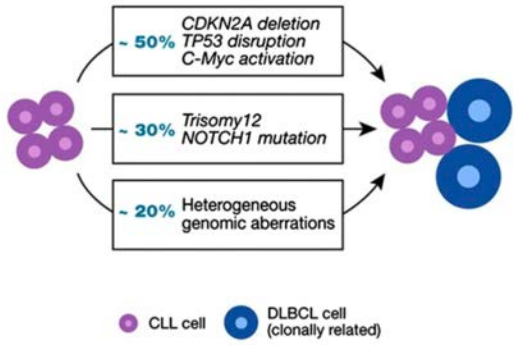


Figure 40. Somatic genetic characteristics associated with Richter transformation. Image from Parikh & Shanafelt, 2014.

Two genetic mechanisms leading to RT have been described. The main one is found in about half of the patients and affects their outcome. It is related to the inactivation of *TP53* (by loss or by somatic mutations) and loss of *CDKN2A/B*, a known tumor suppressor gene involved in cell-cycle regulation, inducing high proliferation rates and deregulation of cell cycle, apoptosis, senescence, and cellular metabolism. These alterations were mutually exclusive with a second

group characterized by the presence of trisomy 12, followed by the acquisition of *NOTCH1* mutations. This second genetic pathway was observed in about one third of the patients (Chigrinova et al., 2013). Another mechanism that apparently orchestrates RT has been seen in murine models that have shown that constitutively active AKT transforms CLL towards aggressive lymphoma, via overactivation of *NOTCH1* (Kohlhaas et al., 2021). In another murine in vivo CLL model, biallelic inactivation of *TP53* and *CDKN2A*/*CDKN2B* lead to more aggressive disease, allowing BCR-dependent/costimulatory signal-independent proliferation of CLL cells (Chakraborty et al., 2021). DNA damage response pathways have also been identified as a potential mechanism driving RT, as numerous mutations in involved genes have been identified and pathway-based clonal deconvolution analysis demonstrates high clonal-expansion probability (Klintman et al., 2021).

Despite improvements on the outcome of high-risk CLL patients under novel targeted therapies, RT also develops in patients under novel agents and it usually occurs as an early event, within the first 4-16 months (Anderson et al., 2017; Innocenti et al., 2018). Most genetic studies of RT in the era of novel agents include patients treated with ibrutinib. The occurrence of resistance-associated mutations in *BTK* and *PLCG2* is higher in CLL progression than RT, where they are either not reported, identified in a smaller fraction of cases, or they are different from those of CLL (Innocenti et al., 2018; Kadri et al., 2017). Even though there are fewer studies exploring the genetic aberrations under novel agents, the recurrent genetic alterations in RT identified are similar to those under CIT (Figure 41).

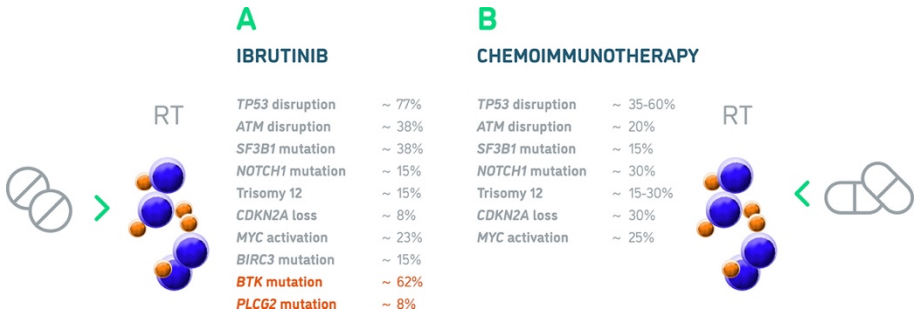


Figure 41. Recurrent genetic aberrations in RT developed on ibrutinib (A) and CIT (B). Image from Petrackova et al., 2021.

Although previous studies have identified risk factors and recurrent alterations of RT, the mechanisms underlying this transformation are not thoroughly understood. The evolutionary history of RT and its driving genomic/epigenomic determinants remain largely unknown. An in-depth understanding of this process might help to anticipate RT and guide novel treatment strategies, improving the outcome of these high-risk patients.

2 Objectives

The goal of this thesis is to design, set-up, and apply computational methodologies to answer specific biomedical questions within genome oncology. The three main objectives correspond to the three conceptual blocks that constitute the thesis:

1) To adapt, implement, and execute the large-scale cancer genomics project Pan-cancer Analysis of Whole Genomes (PCAWG) infrastructure into supercomputing premises.

2) To assess, develop and implement methods and strategies for genome analysis in oncology, with specific focus on variant identification, characterization, and interpretation.

3) To answer specific biomedical questions in the context of the genomics of CLL using the methodology generated within this thesis:
   a) To understand the molecular mechanisms and the clonal evolution of Richter transformation in CLL to ultimately identify early clinical markers.
   b) To identify the role of *ATM* germline variants and somatic mutations in the evolution of CLL in a young adult.

3 Methods

3.1 PCAWG computational infrastructures and workflow frameworks

For the generation of infrastructures related to the PCAWG initiative (Objective 1), our group faced several challenges. Our center, the BSC, is mainly based on high-performance computing (HPC). HPC provides advantages in terms of computational power and efficiency, but also imposes restrictions, and the cloud-based solutions adopted by project could not be directly implemented in our premises.

Part of the PCAWG infrastructure was dedicated to the execution of workflows, a group of tools and scripts that need to be run in a specific order. Due to the large size of the data and its distributed nature, these workflows or pipelines were set to be executed in different computing centers. Portability and reproducibility among them were key points that needed to be addressed and virtualization techniques were used to ensure them.

Initially, the PCAWG project used the SeqWare workflow execution engine (O'Connor et al., 2010), an open source portable software infrastructure designed to analyze massive genomic datasets, to bundle software, and to execute pipelines in virtual machines (VMs). Later, the project adopted docker as a key enabling technology for running workflows across different platforms. Dockstore was used to place and share PCAWG workflows.

Both types of executions (i.e., workflows either in VMs or packaged into docker containers) were performed in MN3. The BSC's computational resources committed to the project are summarized in Table 1. Note that the runs in the VMs system report the actual running time of the executions. However, this setting required the separation of a whole rack of MN3 (see Results - Chapter 1:

Study 1) which was reserved the whole time for the project, thus having dedicated CPU core-hours even when it was not utilized. Besides this computational power, the BSC also stored between 500GB and 1PB of data throughout the whole project.

Table 1. BSC's computational resources dedicated to PCAWG.

Pipeline	*Infrastructure*	*# Samples*	*CPU core-hours*
Alignment	VMs	550	200,000
Sanger Variant Calling	Docker	500	600,000
DKFZ/EMBL Variant Calling	Docker	850	900,000

3.2 Data collection

3.2.1 Benchmarking datasets

Within the second objective of the thesis, we evaluated several strategies for variant identification in tumor genomes. The results of this work are detailed in Results - Chapter 2: Study 2 and Study 3.

To evaluate the accuracy of variant calling (VC) methods, benchmarking datasets with validated "truth" variants are crucial. These datasets are usually generated within particular studies, covering a limited number of variants, and less than a handful of comprehensively characterized samples are available and findable for somatic VC (Alioto et al., 2015; Griffith et al., 2015). Besides these attempts to fully characterize whole-genome or whole-exome samples, orthogonal validation with other experiments, such as gene panels or RNA-seq, can also be used to evaluate variant calling performance (Ellrott et al., 2018), although it might be biased towards specific regions of the genome (e.g., coding regions in the case of gene panels). Finally, one can restrict the validation approach to specific mutations validated by Sanger or deep sequencing, which is

expensive and makes it practically impossible to cover all mutations in a whole genome.

Synthetic data can also be used for benchmarking. This simulated approach is very convenient to reach the points that real data cannot catch and to have complete control over the "true" variants as well as the false positives. However, it lacks the artifacts and other intricate features only present in real data. Therefore, the results presented within the thesis will be mainly based on real samples.

In order to evaluate our variant calling strategies (see Results - Chapter 2: Study 2 and Study 3) we first used validated variants coming from WES and WGS that have been entirely characterized.

For WES, we used the exome capture of a primary acute myeloid leukemia (AML), sequenced at a coverage of ∼433x, and a matched normal skin sample (Griffith et al., 2015). This data was produced in a study where they performed both WES and WGS of a primary and relapsed AML tumor (AML31). After extensive filtering, validation of ∼200,000 putative SNVs by deep sequencing (coverage of ∼1,000x), and manual review, they produced a list of "platinum" or "validated" variants that contained 1,343 high-quality SNVs. We used the subset of these variants that corresponded to the exomes and downsampled the original WES to 140x and 90x, for the tumor and normal samples, respectively, to match the sequencing coverage that was being used within the MedPerCan project (see Results - Chapter 2: Study 2).

For WGS, we used the medulloblastoma sample (MB99) that was prepared for a benchmarking exercise within the context of the ICGC (Alioto et al., 2015). They sequenced the tumor sample at ∼300 × in five different sequencing centers and used this high coverage to curate a gold set of somatic mutations. They

categorized these variants into different tiers, according to the difficulty or easiness of being detected. The benchmarks presented in the Results section (Results - Chapter 2: Study 3) focus on Tier 1, which includes 962 SNVs and 337 indels with a VAF≥10%, and Tier 4, that contains additional mutations with low VAF and ambiguous local alignments (1,263 SNVs and 347 indels), to calculate the recall and precision, respectively. We used the downsampled WGS at 30x that were used within the PCAWG project to match the coverage of the ongoing projects.

As there were not any other published datasets with such a wide characterization of somatic variants, we complemented our benchmarks with other real data approaches including orthogonal validations using high coverage gene panels. We used the data from two previously published studies that performed gene panel sequencing on a subset of diffuse large B-cell lymphoma (DLBCL) WES samples that were part of the MedPerCan project and a number of chronic lymphocytic leukemia (CLL) WES and WGS samples that had previously been sequenced, respectively (Karube et al., 2018; Nadeu et al., 2018; Puente et al., 2015).

For our evaluations, we used the variants (SNVs and indels) identified in the high coverage gene panels as the “true” variants and considered “false” positives those mutations seen in the WES/WGS samples within the gene panel territory, but not detected in the gene panel results.

For the CLL series, we used the reported variants directly, while for the DLBCL set we re-analyzed the gene panels using the same pipeline that was used for CLL as the original publication only reported non-synonymous variants. We could evaluate 29 WGS and 64 WES CLL samples on 28 CLL driver genes (Results - Chapter 2: Study 2 and Study 3), and 13 WES DLBCL samples on 106 DLBCL driver

genes (Results - Chapter 2: Study 2). We used variants with VAF above 10%, as gene panels are able to detect low frequency variants that cannot be identified by WES or WGS at lower coverages. More details follow in the results' corresponding section (Results - Chapter 2: Study 2 and Study 3).

The orthogonally validated dataset of DLBCL was also used to assess the accuracy of tumor-only analyses. In this case, further filtering based on the functional impact was applied. We considered the impact annotation from snpEff (Cingolani, Patel, et al., 2012) and selected variants with a "HIGH" or "MODERATE" impact, which include disrupting mutations such as in frame or frameshift indels, missense SNVs, stop gained, or start loss variants.

3.2.2 Richter transformation study cohort

As part of the third objective of the thesis, we applied and interpreted the results of cancer genomics methodologies. The first study was about Richter transformation in CLL (Results - Chapter 3: Study 4).

A total of 19 patients fulfilling the criteria of RT after pathological revision and with good quality samples were included in this study. All of them were subjected to whole-genome sequencing. Low purity tumor samples and normal samples with high tumor contamination hampering the detection of somatic variants were discarded. Integration of multiple omics (genome, epigenome, and transcriptome) and resolution levels (bulk and single cell) was available for a subset of cases. Besides this CLL-RT cohort, described in more detail below, we also included a dataset of 147 previously published WGS of CLL at diagnosis (ICGC-CLL cohort) and 27 WGS of post-treatment CLL samples (CLL post-treatment cohort) for the mutational signature analysis (Figure 42).

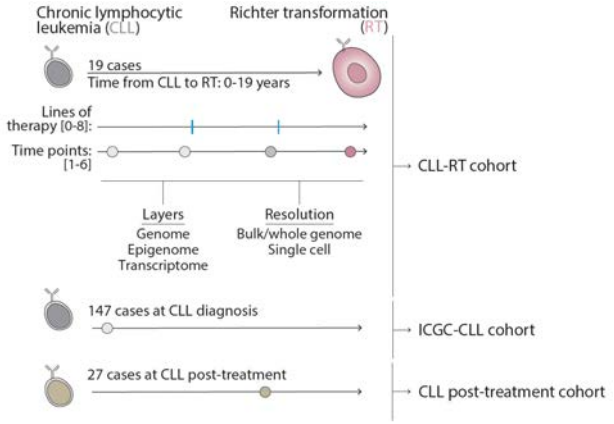


Figure 42. Summary of the CLL-RT cohort and additional datasets. 19 cases were included in the RT study, with multiple samples collected along the course of the disease and analyzed with different technologies. Two additional cohorts (bottom) were included in specific analyses (i.e., mutational signatures).

Out of the 19 selected cases of the CLL-RT cohort, three developed RT before therapy, while in the other cases the transformation occurred after chemoimmunotherapy or after multiple lines of treatment including targeted therapies (ibrutinib, duvelisib, idelalisib, and venetoclax) (Table 2). Note that the majority of the latter patients received several lines of treatment before the transformation.

Table 2. Treatments that the patients received prior to RT

Summary of the last treatments prior to RT

Treatment	None	CIT	Ibrutinib	Duvelisib	Idelalisib	Venetoclax
# Cases	3	6	6	2	1	1

In all but one case we collected and analyzed multiple synchronous and/or longitudinal samples (range 2-8 samples/case), which were obtained at different time points of the disease from CLL diagnosis to RT. For 12 cases we had a

complete WGS data set (germline, CLL, and RT samples were analyzed), while the previous CLL sample or germline material was not available for 1 and 6 cases, respectively. Bulk RNA-seq, methylation arrays, and ATAC-seq/H3K27ac were available for 6 non-overlapping cases. Single-cell DNA and single-cell RNA sequencing were available for 4 and 5 cases, respectively (Figure 43 and Figure 44).

Regarding the types of transformation, 17 cases had a diffuse large B-cell lymphoma-type (RT-DLBCL), 1 case had a plasmablastic lymphoma transformation (RT-PBL, case 1669), and 1 developed a prolymphocytic leukemia transformation (RT-PLL, case 3299). For simplicity, all cases were analyzed together as RT (Figure 44).

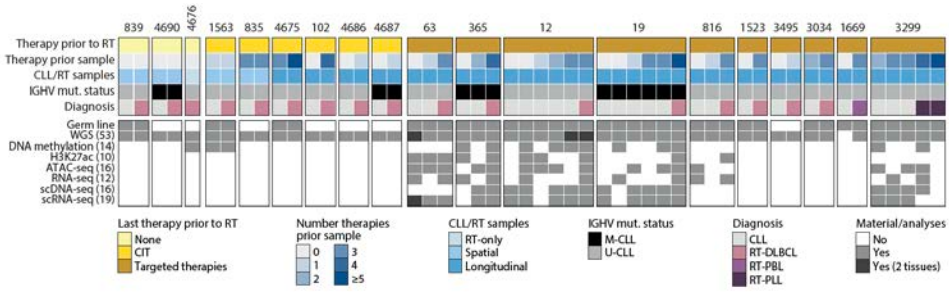


Figure 43. Details of the cases included in the CLL-RT cohort and available data types. Each case is represented by the upper numbers, while each column represents a time point. The first 5 rows indicate the last therapy prior to RT, the number of therapies the patient received before the corresponding sample, whether RT and other samples are longitudinal or spatial, the IGHV mutational status, and the diagnosis at the time of the corresponding sample. The 8 lower rows indicate whether the material and analyses are available or not at each time point.

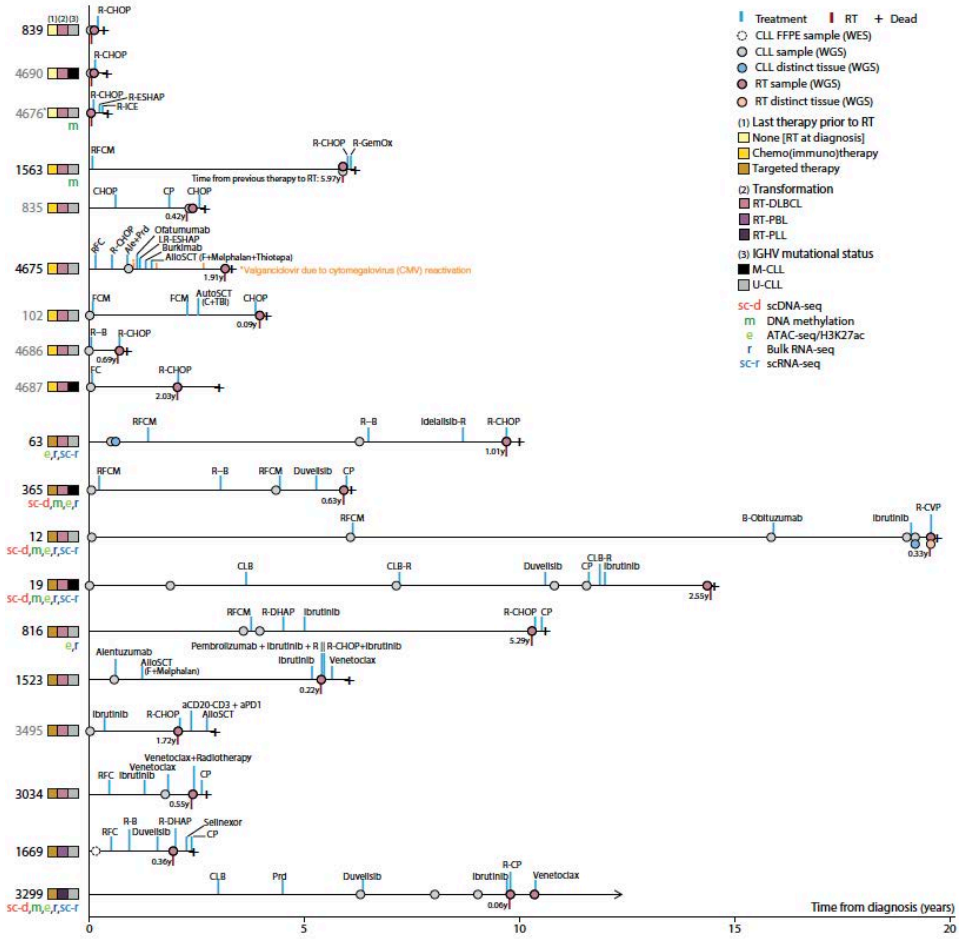


Figure 44. Cohort studied and types of Richter transformation. Representation of the disease course of the patients included in the study. Each sample analyzed, treatment, and date of Richter transformation (RT) are depicted. Cases labeled in gray lacked germline DNA. Case 4676 also lacked DNA from the previous CLL sample. Cases are grouped based on the therapy received prior to RT in three groups: cases developing RT before any treatment, after chemo(immuno)therapy, and after targeted therapy. The type of transformation [RT-DLBCL, diffuse large B-cell lymphoma type; RT-PLL, prolymphocytic transformation; RT-PBL, plasmablastic transformation] and IGHV mutational status are also shown. Additional molecular studies conducted in each case are also depicted. Abbreviations: Ale: alemtuzumab; AlloSCT: allogeneic stem cell transplantation; AutoSCT: autologous stem cell transplantation; B: bendamustine; Burkimab: rituximab, methotrexate, dexametasone, ifosfamide, vincristine, etoposide, cytarabine, doxorubicin and vindesine; C: cyclophosphamide; CHOP: cyclophosphamide, doxorubicin, vincristine and prednisone; CLB:

chlorambucil; CLB-R: chlorambucil and rituximab; CP: cyclophosphamide and prednisone; F: fludarabine; FCM: fludarabine, cyclophosphamide and mitoxantrone; G-GemOx: rituximab, gemcitabine, and oxaliplatin; LR-ESHAP: lenalidomide, rituximab, etoposide, methyl-prednisolone, cytarabine and cisplatin; M: mitoxantrone; Prd: prednisone; R: rituximab; R-B: rituximab and bendamustine; R-CHOP: rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone; R-CVP: rituximab, cyclophosphamide, vincristine and prednisone; R-DHAP: rituximab, dexamethasone, cytarabine and cisplatin; R-ESHAP: rituximab, etoposide, methyl-prednisolone, cytarabine and cisplatin; RFC: fludarabine, cyclophosphamide and rituximab; RFCM: rituximab, fludarabine, cyclophosphamide and mitoxantrone; R-ICE: rituximab, ifosfamide, carboplatin and etoposide; TBI: total body irradiation.

3.2.3 Case report of CLL carrying *ATM* germline variants

The second study within the third objective of the thesis was a case report of a young adult with CLL harboring *ATM* germline variants (Results - Chapter 3: Study 5). To characterize this case, we covered 8 years of genomic evolution from CLL diagnosis.

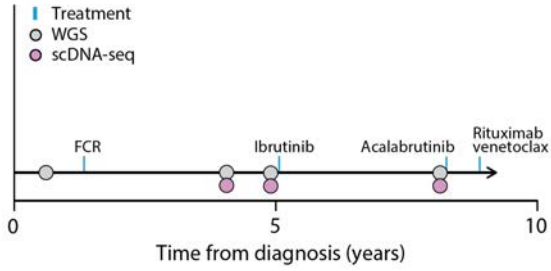


Figure 45. Clinical course and samples analyzed. Samples are represented by small circles. Sequencing with different technologies (i.e., WGS and scDNA-seq) are indicated by colors. The treatments that the patient received along the course of the disease are annotated. FCR: fludarabine, cyclophosphamide and rituximab.

Samples were available at 4 time points, and whole-genome sequencing was performed for all of them at 30x. The germline and the tumor sample at diagnosis (T1) were included in our previous ICGC-CLL study (Puente et al., 2015), and the

coverage of T1 was increased to 60x during the current study to improve the detection of mutations. Adequate samples for scDNA-seq were available for the last three time points (Figure 45).

3.3 Bioinformatics analysis

NGS analysis starts with the sample preparation, followed by the sequencing procedures, the primary data analysis, the downstream analysis, and the final interpretation of the results. Bioinformatics analysis covers all data processing, from the FASTQ files obtained after sequencing, their alignment, and the detection of mutations, to the downstream analysis, which is meant to provide for the biological interpretation of the results (Figure 46).

The next subsections describe each of the steps of the primary and downstream analyses together with a brief description of the methods that have been used during this thesis.

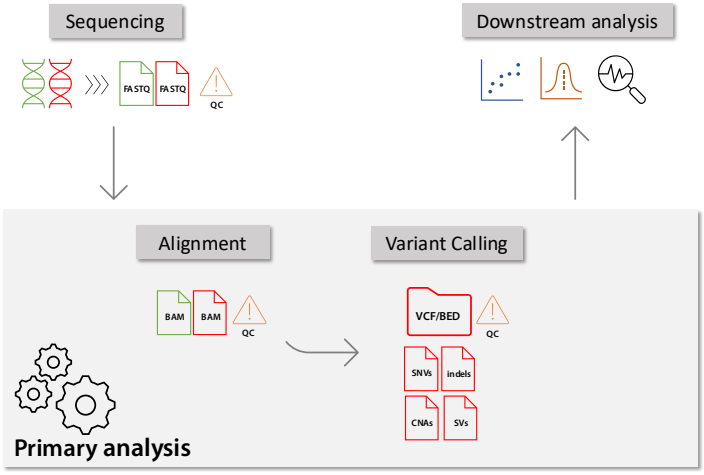


Figure 46. Bioinformatics analysis for cancer genomics. Main steps to analyze tumor genomes: alignment of read files and identification of somatic variants are the main steps of the primary analysis, which outputs are used for the downstream analysis.

3.3.1 Alignment

Sequencing yields a collection of reads, encoded with the 4-letter alphabet (A, C, G, and T) and stored in FASTQ files. These DNA sequences are mapped to a reference genome to find out their locus. The alignment process is not straightforward as it is shadowed by uncertainties concerning repetitive regions, multi-mapping, and alterations inherent in tumor samples. BWA (H. Li & Durbin, 2009) is the most widely used aligner, specially its BWA-MEM algorithm. This program was used to map the reads to the human reference genome (GRCh37) in all the studies of the thesis. The output was converted to BAM and sorted using biobambam2, a suite for processing BAM files. Optical or PCR duplicates were flagged by the same package. Finally, SAMtools (Danecek et al., 2021), a single executable that offers multiple commands for working on alignment data, was used to index the final BAM file (Table 3). BAM files were subjected to base quality score recalibration (BQSR) upon recommendation of some variant callers (i.e., Mutect2, MuSE, and lancet). GATK's BQSR was used to perform this step. Program versions and repositories are detailed in Table 3.

An additional tool was included for PDX samples in order to filter out reads coming from mouse DNA. Disambiguate is an algorithm that separates sequencing reads of two species derived from grafted samples. It operates on previous alignments of all reads against the reference genomes of the two species and disambiguates their source. Applying this process prior to variant calling, or other analyses, allows more accurate results as they are not confounded by cross-species contamination.

Table 3. Program versions used for alignment.

Program	*Version*	*Reference*
BWA (BWA-MEM)	v0.7.15	https://github.com/lh3/bwa

Biobambam2	v2.0.65	https://gitlab.com/german.tischler/biobambam2
SAMtools	v1.9	https://github.com/samtools/samtools
GATK	v4.0.2.0	https://github.com/broadinstitute/gatk
Disambiguate	v1.0.0	https://github.com/AstraZeneca-NGS/disambiguate

3.3.2 Variant calling

Variant calling (VC) is the process that identifies or "calls" variants in the genome. In essence, it determines the differences between the analyzed sample and the reference genome. Variant discovery can be specialized in germline or somatic variants, or in particular types of variants, including SNVs, indels, SVs, or CNAs (see Introduction - section 1.2.3.2 and Figure 7). Strategies to tackle each kind of alteration differ, and tools are sometimes focused on one or two single classes. Somatic variant calling requires additional considerations and is more difficult than germline variant calling, where all variants are expected at a high frequency (50% or 100% of the reads). Ideally, the analysis of somatic variants in tumors uses an additional matched normal sample to distinguish somatic variants from germline variants, that is, selecting the mutations that are present in the tumor but not in the normal sample. Furthermore, tumor purity and heterogeneity have a direct impact on the frequency of the variants, which become more difficult to detect as the number of supporting reads decreases (see Introduction - section 1.3.2 and Figure 17).

Variant calling tools utilize heuristic, Bayesian, and other sophisticated probabilistic models to infer statistical correctness of candidate variants. Confidence of the calls can be based on mapping quality scores, base quality scores, strand bias, and GC content, among others. In the next section, there is a brief description of the methods used within the thesis, and their scope is summarized in Table 4.

3.3.2.1 Tools

Due to the great variety of variant callers generated by the community, the selection of VC tools becomes the first challenge for the analysis of genomes. As many variant callers are designed for a specific range of variants, the complete analysis of genomes requires the combination of several callers. In addition, since each caller has its strong and weak points, the combination of callers also helps remove false calls when filtering by a minimum number of callers that identified a certain variant. Among the available programs, we have chosen a selection of those that are most widely used, and those that have been selected by large-scale initiatives like the PCAWG or the TCGA (Campbell et al., 2020; Ellrott et al., 2018). Overall, they cover all types of variants and support WES and WGS analyses (Table 4).

Table 4. Variant calling tools and types of variants they detect.

Program	*Variants*	*Sequencing*
HaplotypeCaller (v4.0.2.0)	SNV/indel (germline)	WES/WGS
cgpCaVEMan (v1.12.0)	SNV	WES/WGS
Mutect2 (v4.0.2.0)	SNV/indel	WES/WGS
MuSE (vv1.0rc)	SNV	WES/WGS
Sidrón	SNV/indel	WES/WGS
Platypus (v0.8.1)	SNV/indel	WES/WGS
Lancet (v1.0.5)	SNV/indel	WES/WGS
Strelka2 (v2.8.2)	SNV/indel	WES/WGS
cgpPindel (v2.2.3)	indel	WES/WGS
SvABA (v7.0.2)	indel/SV	WES/WGS
SMuFin	SNV/indel/SV	WES/WGS
cgpBattenberg (v3.2.2)	CNA	WGS
ascatNgs (v4.1.0)	CNA	WGS
Facets (v0.5.14)	CNA	WES/WGS
BRASS (v6.0.5)	SV	WGS
Delly2 (v0.8.1)	SV	WGS
CNVkit (v0.9.3)	CNA	WES/WGS

HaplotypeCaller (McKenna et al., 2010) is a germline variant caller of short variants (SNVs and indels) from GATK. It performs local assembly of all reads in each region of the genome with potential variation to identify all possible haplotypes. Next, it realigns each haplotype against the reference haplotype to determine potentially variant sites. To determine the likelihood of the haplotypes given the read data, it performs a pairwise alignment of each read against each haplotype with the PairHMM algorithm. These likelihoods are input to a probabilistic model to assess the likelihood of variants implied by the assembled haplotypes. For each potentially variant site, the Bayes' rule is applied to calculate the posterior likelihoods of each genotype given the read data and the most likely is selected.

cgpCaVEMan (Jones et al., 2016) is a wrapper of the CaVEMan algorithm, an expectation maximization-based somatic substitution-detection method. It generates SNV calls, considering copy number segments, purity, and ploidy information. It can be combined with a set of validated post-hoc filters, including problematic regions based on the UCSC High Seq Depth track, a panel of normals, and germline indels called by cgpPindel, among others, to improve recall and positive predictive value.

MuTect2 (McKenna et al., 2010) is one of the most popular tools and is part of GATK. It performs local assembly of haplotypes and uses a Bayesian somatic genotyping model, based on the HaplotypeCaller machinery, to call somatic short mutations (SNVs and indels). MuTect2 filters variants based on mapping quality, strand bias, read position bias, and panels of normals, among others, to eliminate artifacts due to library preparation, sequencing, and mapping.

MuSE (Fan et al., 2016) is a statistical approach for SNV calling based on a Markov substitution model for molecular evolution that models the evolution of

the allelic composition of the normal and tumor samples at each reference base. It uses a sample-specific error model to identify tier-based cutoffs (PASS, Tiers 1 to 5) and improve overall accuracy. The input BAM files are recommended to be processed following the GATK Best Practices, including the recalibration of best quality scores.

Sidrón is a proprietary software to detect SNVs that implements a probabilistic binomial model that uses genotyping data to calibrate sequencing error per sample (Puente et al., 2011). This program was run at the developer's premises by collaborators.

Platypus (Rimmer et al., 2014) uses local realignment of reads and local assembly to detect SNPs, MNPs, and short indels (deletions up to several kb, using the assembly option). Identification of somatic variants requires custom filtering, we used the somaticMutationDetector.py script to identify somatic indels called by Platypus with a minimum posterior of 1.

Lancet (Narzisi et al., 2018) is another somatic variant caller for short read data that was recently published at the time. It can detect SNVs and indels, including deletions up to 400bp and insertions shorter than 200bp. It uses a localized micro-assembly strategy based on the colored de Bruijn graph assembly paradigm that jointly analyzes tumor and normal reads within the same graph. Due to its pure local-assembly strategy, it currently has longer runtimes than other methods based on alignment.

Strelka2 (S. Kim et al., 2018) is a method for germline and somatic small variant calling, including SNVs and indels. It uses a novel mixture-model-based estimation, complemented by a final empirical variant scoring (EVS) step, which is based on machine-learning variant classification approaches.

cgpPindel (Raine et al., 2015) is a customized version of pindel, updated to support split read mappings by BWA-MEM and to provide additional post calling filtering, including a panel of aberrant sites and UCSC High Seq Depth regions.

SvABA (structural variation analysis by assembly) (Wala et al., 2018) is a tool to efficiently detect SVs and indels genome-wide. It performs local assembly to create consensus contigs from sequence reads that diverge from the reference genome, i.e., gapped, clipped, unmapped, and discordant read pairs, and compares them to the reference to annotate the variants.

SMuFin (Somatic MUtation FINder) (Moncunill et al., 2014) is a reference-free method capable of identifying multiple types of somatic variants from the direct comparison of tumor samples with their matched normal sample. SMuFin can detect SNVs, indels, and SVs in a single run. This tool was run by Montserrat Puiggròs from the Computational Genomics group at BSC.

cgpBattenberg is a wrapper of the Battenberg algorithm (Nik-Zainal, Van Loo, et al., 2012), which identifies subclonal copy number from whole-genome sequencing based on allele ratios by haplotype rather than individual SNPs.

ascatNgs (Raine et al., 2016) is a wrapper of the ASCAT (Allele-Specific Copy number Analysis of Tumors) method, which derives copy number profiles of tumor samples, considering normal cell admixture as well as tumor aneuploidy.

Facets (Shen & Seshan, 2016) is an algorithm to estimate the fraction and allele specific copy number states corrected for tumor purity, ploidy, and clonal heterogeneity from matched tumor and normal sequencing, including WGS, WES, and targeted sequencing.

BRASS (Nik-Zainal, Van Loo, et al., 2012) determines potential rearrangement breakpoints from pair-end sequencing. It considers read pairs

where both ends map but are not marked as properly paired, groups them based on mapped locations, and performs an assembly.

Delly2 (Rausch et al., 2012) is a SV prediction method to discover, genotype, and visualize all kinds of structural variants, including deletions, tandem duplications, inversions, and translocations, based on read-depth, paired-end, and split-read information. It can use a panel of normal samples to filter out artifactual false positives and germline SVs.

CNVkit (Talevich et al., 2016) is a Python library and command-line software toolkit to infer and visualize copy number alterations and it was designed for use with hybrid capture, including both WES and target panels. It uses both the targeted reads and the off-target reads to infer copy number events evenly across the genome. It can use a pooled reference to determine somatic copy number alterations when the matched normal sample is unavailable. Filtering and merging strategies, developed by Dr. Ferran Nadeu from IDIBAPS, were applied to the results of the Richter transformation study (Results – Chapter 3: Study 4): CNAs smaller than 500kb or with an absolute log2 copy ratio (log2CR) <0.3 were removed. CNAs within any of the immunoglobulin loci were removed. CNAs were classified as gains if log2CR >0.3, deletions if log2CR <(-0.3), high-copy gains if log2CR >1.1, and homozygous deletions if log2CR <(-1.1). Note that the log2CR cutoff was set to 0.15 for two samples with low tumor cell content (102-01-01TD and 4690-03-01BD). To avoid the high segmentation of the CNA profile obtained using CNVkit, we merged CNAs belonging to the same class if they were separated by <1Mb and with an absolute log2CR difference <0.25.

An increasingly popular approach for variant calling is the combination of different tools, each one with its own strengths and weaknesses, to provide a consensus result that outperforms all integrating methods. The final calls can be

the union, intersection, or selection of variants detected by a minimum number of programs or machine learning strategies. Prior to combining the results, individual calls can be filtered according to custom criteria (e.g., allele frequencies, high-confidence regions, or other quality-related metrics) to improve performance and, next, they should be normalized in such a way that they can be compared to the results of the other methods. Finally, equivalent variants from different programs are merged together and those fulfilling the merging strategy criteria will form the final list of variant calling results.

3.3.2.2 Filtering of variant calling results

Each variant caller has its own statistics and conventions to distinguish true variants from sequencing errors or other artifacts, and usually provides additional annotations on which they based their findings, together with variant-related features, such as the variant allele frequency. All this information can be used to fine-tune the original results from the program, for instance to achieve higher precision at the expense of losing some sensitivity, or to apply more stringent criteria in problematic samples. An illustrative exploration of this kind of filtering can be seen in Results - Chapter 2: Study 3.

In Study 3, we applied the following filters for the benchmarking:

For SNVs:

- CaVEMan: CLPM [number of soft clipped bases in variant allele reads] >0 and ASMD [median alignment score of variant allele reads] <90, <120, or <140 for sequencing read lengths of 100, 125, or 150 bp, respectively
- Mutect2: MMQ [median mapping quality of supporting reads] <60
- Strelka2: SomaticEVS [somatic Empirical Variant Score] <17, MQ [RMS mapping quality] <60, and MQ0 [total mapping quality zero reads] >0

- Lancet: FETS [phred-scaled p-value of the Fisher's exact test for tumor-normal allele counts (right-sided)] <10 and SB [phred-scaled strand bias of the Fisher's exact test (two-sided)] <4

For indels:

- Mutect2: MMQ [median mapping quality of supporting reads] <55
- Strelka2: SomaticEVS [somatic Empirical Variant Score] <0, MQ [RMS mapping quality] <55, and MQ0 [total mapping quality zero reads] >0
- Lancet: FETS [phred-scaled p-value of the Fisher's exact test for tumor-normal allele counts (right-sided)] <10 and SB [phred-scaled strand bias of the Fisher's exact test (two-sided)] <4
- Platypus: MQ [root mean square of mapping qualities of reads at the variant position] < 55
- SvABA: MAPQ [mapping quality from BWA-MEM of the assembled contig] <55

For SVs:

- SvABA: MAPQ [mapping quality from BWA-MEM of the assembled contig] <60
- Delly2: MAPQ [median mapping quality of paired ends] < 60
- Brass: BAS [brass assembly score: a maximum score of 100 indicates a perfect pattern of 5 vertices in Velvet's de Bruijn graph] < 90

In Studies 4 and 5, we applied very similar filters:

SNVs detected by CaVEMan with CLPM>0 and ASMD values <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.

Indels variants with MMQ<60, MQ<60, and MAPQ<60 identified by Mutect2, Platypus, and SvABA, respectively, were removed.

SVs were filtered out if they had BAS<90 for BRASS or MAPQ<60 for SvABA and Delly2.

3.3.2.3 Merging and consensus variant calling results

To achieve higher agreement among variant callers, we first normalized their results to obtain a unified representation of the genetic variants. SNVs were normalized and indels were left-aligned and normalized using bcftools (Danecek et al., 2021) and intersected using custom scripts. For SVs, a custom script, developed by Ana Dueso from the Computational Genomics group at the BSC, was used to merge the calls considering a window of 300bp around the breakpoints.

In the MedPerCan project (see Results - Chapter 2: Study 2), we applied the following programs and merging strategies:

- Somatic SNVs were identified using CaVEMan, Mutect2, MuSE, Strelka2, and Lancet. No additional filters were applied, and consensus results retained variants called by at least three programs.
- Somatic indels were called using Pindel, Platypus, Mutect2, Strelka2, Lancet, and SvABA. No additional filters were applied, and variants detected by at least 3 programs were kept.
- Germline variants were determined running HaplotypeCaller and applying the following filters: read depth <8, fisher strand >25.0, quality by depth <6.0, and RMS mapping quality <50.0.

In Studies 4 and 5 (Results - Chapter 3), we applied the following programs and merging strategies for somatic variant calling:

- SNVs were called using CaVEMan, Mutect2, and MuSE. Filtering was applied as previously described, and variants detected by at least two programs passing custom filters were kept.

- Indels were identified using SMuFin, Pindel, SvABA, and Mutect2. Filtering was applied as previously described, and variants detected by at least two programs passing custom filters were retained.
- CNAs were obtained from visual inspection and manual consensus of the results of ASCAT, Battenberg, and Genome-wide Human SNP Array 6.0, when available. Manual curation of CNAs was done by Ferran Nadeu. CNAs within immunoglobulin loci were not considered.
- SVs were extracted using SMuFin, BRASS, SvABA, and Delly2. Filtering was applied to each program as previously described. Variants detected by at least two programs and at least one of them passing custom filters were kept. IgCaller was used to call SVs within the immunoglobulin regions. All SVs were visually inspected using the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

In the Richter transformation cohort (Results - Chapter 3: Study 4), there were special considerations to adapt to the particularities of some cases:

- For the two cases that underwent allogeneic stem-cell transplant before RT (cases 1523 and 4675). We performed variant calling on tumor vs patient's germline and tumor vs donor's germline in parallel. Next, we intersected both results and only kept the variants identified in both analyses in order to filter out both the patient- and the donor-specific germline variation.
- For the cases lacking a normal sample, a restricted tumor-only variant calling was applied. SNVs and indels were identified within the coding regions of the considered driver genes (see Methods - section 3.3.5). First, mini-BAM files, covering only the driver gene regions, were obtained using Picard tools. Next, we applied a multi-caller approach using Mutect2 (GATK, v.4.0.4.0) (McKenna et al., 2010), VarScan2 (v2.4.3) (Koboldt et al., 2012), VarDictJava (v1.4) (Lai et al., 2016), LoFreq (v2.1.3.1) (Wilm et al., 2012), outLyzer (v1.0) (Muller et

al., 2016), and freebayes (v1.1.0, https://github.com/freebayes/freebayes). Variants were normalized using bcftools (v1.9) (Danecek et al., 2021) and annotated using snpEff/snpSift (v4.3t) (Cingolani, Platts, et al., 2012). Only non-synonymous variants that were identified as PASS by ≥2 algorithms were considered. Variants reported in 1000 Genome Project, ExAC, or gnomAD with a population frequency >1% were removed. This pipeline was developed and run by Ferran Nadeu. Finally, we investigated if the resulting variants were also present in normal samples of our ICGC database of 506 WES/WGS (Puente et al., 2015) and removed them accordingly.

3.3.2.4 Benchmarking

Systematic errors and biases can arise from experimental processing of the samples in the laboratory, as well as from the application of computational methods. Errors can be introduced at all stages: library preparation, sequencing, mapping, and variant calling. Sequencing technologies are prone to errors that can result in incorrect base calling, subsequently leading to incorrect alignment and/or wrong identification of variants (Dohm et al., 2008). Thus, variant calling tools implement sophisticated algorithms that try to disentangle true variants from methodological artifacts.

The evaluation of the performance of variant calling individual tools and consensus approaches requires benchmarking datasets able to assess their accuracy. As mentioned above, these datasets are challenging to identify and to incorporate into the studies. Ideally, they should be composed of the usual normal and tumor sample pairs together with a list of "truth" variants that have been previously validated using orthogonal approaches. Based on this, we can compare the variant calling results with the "truth" variants, often referred to as "Golden", and use a collection of metrics to define their efficiency (Figure 47). The variants

that are detected by the algorithm under evaluation and are present in the list of "truth" or "Golden" variants correspond to the true positives (TP); the variants that are identified by the program but are not present in the "truth" set are considered false positives (FP); and the variants that are not predicted by the tool but are present in the "truth" list are defined as false negatives (FN). Using these three values we can determine the recall of the predicted results, which is defined as the fraction of all "truth" variants that the caller is able to detect, as well as the precision, which measures the fraction of "wrong" or false variants identified by the caller. In order to classify and rank variant callers, the community is using the F1-score, which corresponds to a harmonic mean of the two previous calculations.

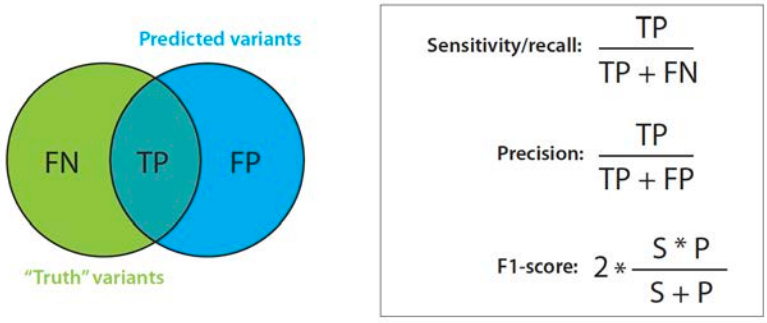


Figure 47. Benchmarking of variant calling results. The predicted variants from the evaluated pipeline are compared to the "Truth" variants of the benchmarking dataset (left), and a set of metrics (right) are used to define their performance. TP: true positives, FN: false negatives, FP: false positives, S: sensitivity, P: precision.

In the same manner as the merging step, variants must be normalized before they can be compared. Both the "truth" and the predicted variants should be represented in the same way to accurately verify their concordance. Thus, SNV and indel normalization was applied prior to benchmarking, and a window of 300bp was used for SVs comparison, as their detection can be less accurate and programs often cannot determine them with base pair resolution.

3.3.3 Quality control

The quality of the biopsies before the sequencing step also has a great impact on the bioinformatics results and seemingly on their biological interpretation and downstream decisions. Quality control (QC) should be assessed all the way from the FASTQ files to variant calling results, and downstream analysis. QC metrics are mainly defined for the primary analysis, but cautious biological interpretation of the mathematical or statistical inferences of downstream methods should also be considered. There are many tools that are used for QC and return sets of metrics that can be grouped into the three main data types that are used and produced during the primary analysis: FASTQ files, aligned BAM files, and variant calling results, usually in VCF format (Figure 46). A brief description of the most popular metrics and programs is exposed below, and the versions used within this thesis is detailed in Table 5.

FastQC stands out as the most widely used tool for QC of FASTQ files. It provides a simple way to do some quality control checks on raw sequence data coming from high-throughput sequencing (HTS) pipelines, including per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and kmer content.

Picard is a suite of tools for manipulating HTS data and supports typical formats such as SAM/BAM/CRAM and VCF. It includes different programs to collect QC metrics from different sequencing data (PCR targeted sequencing, WES, WGS, RNA-seq, etc.). It provides one of the most used QC metrics, the mean coverage, which indicates the average number of reads covering each position, usually excluding low quality, duplicated, and overlapping paired reads,

depending on the tool and set parameters. We used the default settings (i.e., minimum mapping quality > 20, minimum base quality > 20, and excluding duplicates and overlapping bases from paired reads) so that the resulting counts include the informative reads for variant calling. The output of the program provides extensive information about the excluded reads and the fraction of bases that attained a defined set of minimum coverages.

Conpair (Bergmann et al., 2016) is a method to estimate the concordance and contamination for tumor-normal pairs. It provides a fast verification that both the normal and tumor samples come from the same patient, as well as cross-individual contamination level estimation.

MultiQC (Ewels et al., 2016) is a tool to create a single report from multiple tool outputs across multiple samples for easy visualization. It supports many QC programs through additional plugins.

Table 5. Program versions used for quality control.

Program	*Version*	*Reference*
FastQC	v0.11.5	www.bioinformatics.babraham.ac.uk/projects/fastqc
Picard	v2.10.2	https://broadinstitute.github.io/picard
Conpair	v1.0	https://github.com/nygenome/Conpair
MultiQC	v1.7	https://github.com/ewels/MultiQC

3.3.4 Variant annotation

In order to understand the functional impact of mutations, they are usually annotated against gene sets to find if they are in a coding region and to determine their potential effect when they are in a gene.

SNVs and indels were annotated using snpEff/snpSift (v4.3t) (Cingolani, Patel, et al., 2012; Cingolani, Platts, et al., 2012) using RefSeq as a reference

(GRCh37.p13.RefSeq). SnpEff is used for variant annotation and functional impact prediction on genes and proteins, while snpSift annotates variants using external databases. SVs were annotated using a custom script that annotates gene sets within the proximity of break points. A window of 200Kbp was considered followed by manual curation. Driver genes in CLL and DLBCL were considered for the CLL studies (Results - Chapter 3: Study 4 and Study 5).

3.3.5 Driver alterations

In Study 4, driver alterations previously described in CLL (Landau et al., 2015; Puente et al., 2015), DLBCL (Chapuy et al., 2018; Karube et al., 2018), or other B-cell neoplasms were considered as potential drivers in RT. All types of alterations (SNVs, indels, CNAs, and SVs) were considered in the definition of drivers. Alterations that were recurrent and/or had functional references in the literature and that had not been previously reported in RT (Chitalia et al., 2019; De Paoli et al., 2013; Fabbri et al., 2013; Klintman et al., 2021) were considered as novel drivers in RT.

3.3.6 Characterization of complex structural rearrangements

Structural variants are genomic rearrangements that can result in deletions, duplications, insertions, inversions, or translocations. Sometimes these alterations do not occur independently, but they are rather generated in a single-hit event. To better characterize the genomic complexity of our samples, we clustered simple SVs into complex events that, potentially, might have emerged from one punctuated event. This work was carried out in the Richter transformation study (Results - Chapter 3: Study 4).

First, we classified SVs into deletions, duplications, inversions, and translocations according to the variant callers that detected them. Next,

inversions and translocations were tagged as reciprocal when both ends of the double-stranded DNA break were identified. Finally, we further categorized SVs into simple or complex events, which included chromothripsis, chromoplexy, templated insertions, breakage-fusion bridge cycles, and kataegis, based on the following definitions:

- Chromothripsis was determined by the presence of seven or more SV breakpoints occurring in a single chromosome or seven or more oscillating changes between two or three copy number states (Nadeu, Martin-Garcia, et al., 2020; Puente et al., 2015), and supported by additional hallmarks (Cortés-Ciriano et al., 2020; Korbel & Campbell, 2013; Stephens et al., 2011). Chromothripsis events were subcategorized according to the number of chromosomes involved.
- Chromoplexy was defined by the presence of translocations reshuffling three or more chromosomes and leading to chains of rearrangements (Baca et al., 2013). A window of 50kb was used to link proximal SVs and identify potential chains of rearrangements.
- Cycles of templated insertions were determined by the presence of segments of copy number gains in three or more chromosomes interlinked through structural variants (Y. Li et al., 2020; Nadeu, Martin-Garcia, et al., 2020).
- Breakage-fusion bridge cycles were identified as a series of focal copy number increases and fold-back inversions, along with the presence of telomeric deletions (Nadeu, Martin-Garcia, et al., 2020).
- Other complex events defined chains of rearrangements having more than two chained SVs and not meeting any of the previous criteria. These events were further subclassified based on the number of involved chromosomes.
- Kataegis was defined as genomic segments having mutation clusters of six or more consecutive SNVs with an average inter-mutation distance ≤1 Kb (Mayakonda et al., 2018).

A preliminary classification was done according to the previous criteria, and was further refined by manual curation, including a visual comparison with the results of ChainFinder (Baca et al., 2013) and ShatterSeek (Cortés-Ciriano et al., 2020) used to identify chromoplexia and chromothripsis, respectively. Thanks to the multiple longitudinal samples of our study, the presence of the involved alterations in each time point of each case was also considered to finalize the categorization.

Driver genes within 25kb of a breakpoint or that fall in a deletion or an amplification were annotated to SVs.

3.3.7 Immunoglobulin gene rearrangements

IgCaller (Nadeu, Mas-de-les-Valls, et al., 2020) was used to analyze immunoglobulin gene rearrangements, including heavy and light chain rearrangements as well as class switch recombination from WGS. Based on current guidelines (Rosenquist et al., 2017), the sequences obtained from IgCaller were used as input of IMGT/V-QUEST (Brochet et al., 2008) to annotate the genes, functionality, and IGHV mutational status. The ARResT/AssignSubsets online tool (Bystry et al., 2015) was used to analyze stereotypy.

3.3.8 Mutational signatures

Mutational signatures analysis was performed on SNVs classified according to their trinucleotide context into 96 categories (Figure 21) and following three main steps previously described (Maura, Degasperi, et al., 2019): de novo extraction of mutational signatures on all the samples, assignment of the extracted signatures into one or more already know mutational signatures (from COSMIC or any other resource), and fitting of the identified signatures into each of the sample profiles (Figure 48).

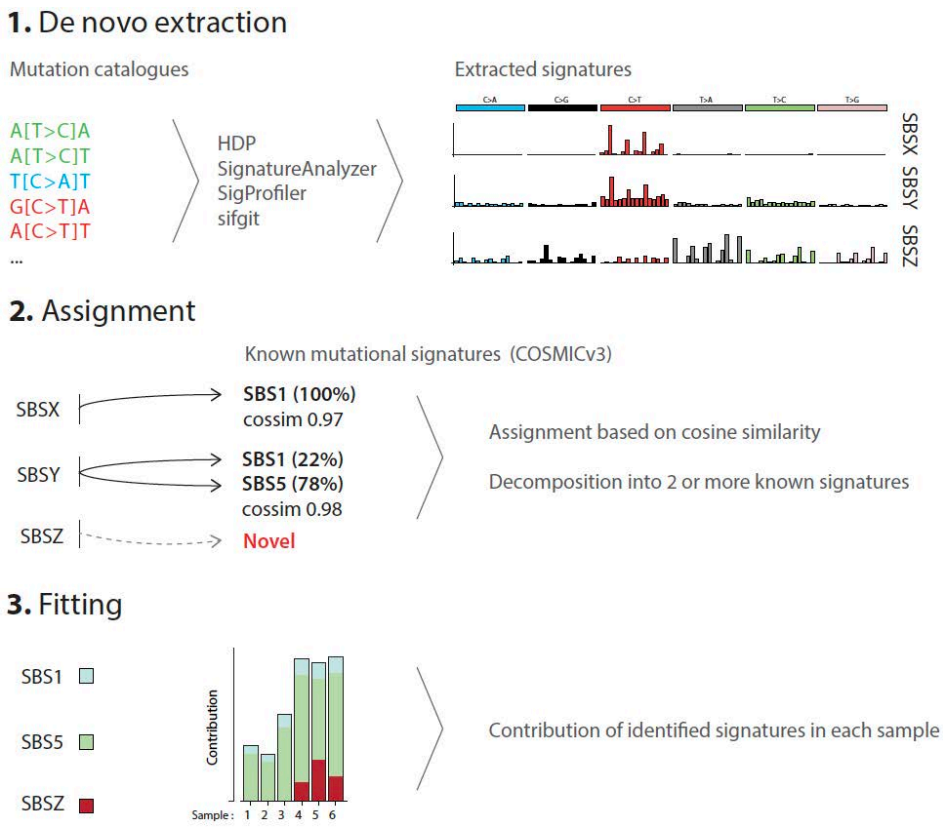


Figure 48. Strategy for the analysis of mutational signatures. The recommended procedure starts with the de novo extraction of mutational signatures, followed by their assignment to already known processes (i.e., signatures from COSMIC or literature) based on their cosine similarity (cossim), and lastly, the estimation of their contribution to each sample.

1. De novo extraction of mutational signatures

First, de novo extraction of mutational signatures is performed to extract patterns of variants (i.e., trinucleotide context of SNVs) from input samples. This step identifies both known and potentially novel mutational processes that might have been active during the tumor's life. There are multiple programs that implement this task, and most of them are based on a NMF approach (see Introduction - section 1.3.4 and Figure 20). To ensure that the results are more

robust and not method-dependent, we ran 4 different algorithms: HDP, SignatureAnalyzer, SigProfiler, and sigfit. The program versions that were used are detailed in Table 6.

HDP (Hierarchical Dirichlet Process) is a non-NMF-based signature extraction method that implements hierarchical Bayesian Dirichlet process. It models mutation classes counts across cancer samples and produces a set of components, the mutational signatures, with a characteristic distribution over the possible mutation categories. Four independent sampling chains were run, each initialized with default parameters, followed by 10,000 burn-in iterations, and the collection of 200 posterior samples off each chain with 200 iterations between each.

SignatureAnalyzer (Alexandrov et al., 2020) uses a Bayesian variant of NMF that allows automatic inference of the optimal number of signatures. It was run with default parameters.

SigProfiler's SigProfilerExtractor (Alexandrov et al., 2020) uses multiple NMF iterations to identify the optimal number of mutational signatures and their activities in each sample. It was run with 1,000 iterations and a maximum of 10 extracted signatures.

Sigfit implements a flexible Bayesian inference of mutational signatures. It enables simultaneous fitting and extraction of signatures and includes four model classes: multinomial, Poisson, normal, and negative binomial. It was run using the multinomial model (equivalent to the NMF approach) to extract five signatures with 10,000 burn-in iterations and 20,000 sampling iterations.

Table 6. Program versions used for mutational signatures analysis.

Program	*Version*	*Reference*
HDP	v0.1.5	https://github.com/nicolaroberts/hdp
SignatureAnalyzer	v0.0.7	https://software.broadinstitute.org/cancer/cga/msp
SigProfiler	v1.0.8	https://github.com/AlexandrovLab/SigProfilerExtractor
sigfit	v2.0.0	https://github.com/kgori/sigfit

The extracted signatures among programs, or different cohorts, can be slightly different even if they represent the same mutational processes. At the same time, the extracted signatures will not be identical to the reference ones, which might be inferred from larger cohorts (i.e., PCAWG) and, thus, more robust. To diminish these minor differences, the similarity between two signatures is used to assign their equivalence, and then the identified reference signatures are used in further analyses. To assess the similarity between two signatures *A* and *B* of *n* mutation types, the cosine similarity is calculated:

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

This equation returns values between 0 and 1. When two mutational profiles *A* and *B* are identical, their cosine similarity will be 1, whereas when they are completely independent, it will be 0. In practice, values close to 1 are used to consider equivalent mutational signatures.

2. Assignment to reference signatures

After the de novo extraction of mutational signatures, the resulting signatures are assigned to already known patterns (i.e., to the COSMIC catalog of mutational signatures and other published results). This assignment is done based on their cosine similarity, and can be done on a one-to-one bases, or require the

deconvolution of the extracted signature into two or more of the reference ones. Finally, it can also happen that an identified signature is a split of a reference one, which would be composed by multiple extracted signatures.

We considered that two signatures are the same if their cosine similarity is greater than 0.85. Extracted signatures that could not be assigned to any reference signature were decomposed into multiple reference signatures using a previously described expectation maximization (EM) algorithm (Lee-Six et al., 2019). The EM algorithm was first run on the set of reference signatures that had already been identified and, next, on the whole set of reference signatures only for the signatures that could not be properly deconvoluted on the first try.

Signatures that could not be recognized from COSMIC (v3.2) or other known mutational processes (de Kanter et al., 2021; Kucab et al., 2019) were considered novel.

3. Fitting of the mutation catalogs and the identified signatures

The fitting process estimates the contribution of the identified signatures, which can be novel processes or reference signatures, in each sample. This purely mathematical calculation tends to overfit the results, leading to the signature bleeding effect, where all signatures might be assigned to all samples. Therefore, some prior knowledge about the mutational processes operating in the studied cohort is crucial.

We have implemented a refined fitting algorithm (Figure 49) based on previous studies (Alexandrov et al., 2020; Maura, Degasperi, et al., 2019). Explicit guidance takes into account previous biological knowledge to minimize overfitting of signatures (false positives) as well as missing signatures (false negatives). The fitting function is applied to one sample at a time, which makes it reproducible

regardless of the other samples in the cohort. The contribution of each identified signature (reference or novel) in each sample was measured using MutationalPatterns (v3.0.1).

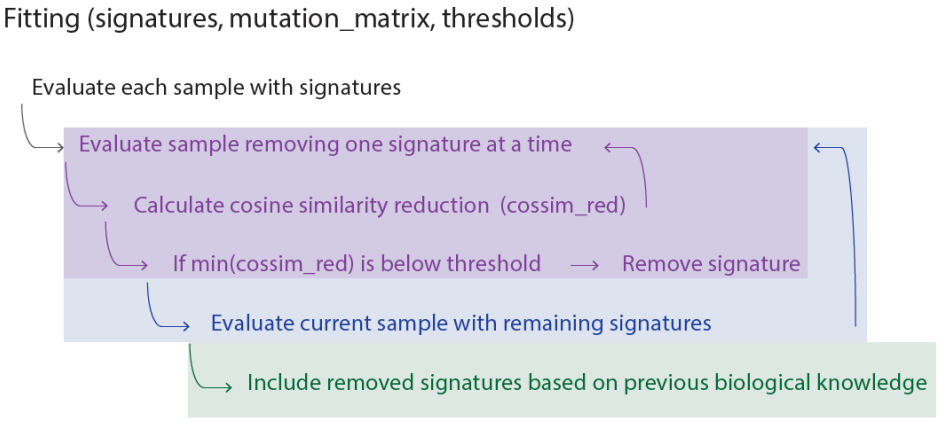


Figure 49. Fitting of mutational signatures. This strategy calculates the contribution of a set of signatures to mutation catalogs from a group of samples, represented in a mutation matrix. Next, it re-evaluates each sample, removing the less contributing signatures (below a set threshold) to avoid signature bleeding. Finally, it also accounts for missed signatures that are known to be present in the sample.

To avoid inter-sample bleeding, we iteratively removed the less contributing signature if it reduced the cosine similarity below a set threshold (0.01). To find the less contributing signature we iterated over all possible signatures, removed one signature at a time, and recalculated the cosine similarity with the new fitting including the remaining signatures. This process was repeated, removing one signature after another, until no signatures could be removed below the threshold. This method cleans up low-contribution signatures, which might represent false positives. SBS1 and SBS5 have been reported in all normal and tumor samples, as they represent clock-like processes that correlate with the age of individuals and can serve as a cell division/mitotic clock (SBS1). As such, they were always included, and were reassigned if they happened to be removed in

the previous iterations and if they improved the cosine similarity, as previously advised (Alexandrov et al., 2020; Maura, Degasperi, et al., 2019). Similarly, SBS9 was added in M-CLL samples if the addition improved the cosine similarity, as it is known to be active in mutated cases of CLL.

As a final step, mSigAct (v2.1.1) (Ng et al., 2017) was used to account for low-contribution mutational signatures. It can assess if the addition of a mutational signature is statistically significant. In particular, it was used in the Richter transformation study (Results - Chapter 3: Study 4) to assess the contribution of SBS-melphalan, a previously described signature related to Melphalan treatment (Maura et al., 2021).

The analysis of mutational signatures in the Richter transformation study (Results - Chapter 3: Study 4) included further work and considerations that are explained below.

To increase the sample size to analyze mutational signatures, we integrated the mutations identified in the CLL-RT cohort together with those of 147 CLL samples published before (Puente et al., 2015). Note that all these 147 additional CLL genomes have been re-analyzed using our current bioinformatic pipeline for harmonization purposes.

We used a principal components analysis (PCA) to simplify the mutational profiles of CLL and RT into two dimensions. Dimensionality reduction was performed on the 96 classes of point mutations integrating the ICGC-CLL WGS cohort, the SNVs detected in the first CLL (time point 1) and RT samples of the same patient from the RT-CLL cohort, and the CLL post-treatment cohort. The percentage of SNVs assigned to each category in each sample was used to obtain the PCA.

Careful review of the signatures extracted led us to refine the previous assignment procedure in two cases: 1) we combined two HDP signatures that together constituted SBS5 to avoid splitting of signatures, and 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. The reasoning behind the latter exception is that these signatures were only acting in one of the samples analyzed, which impaired a more precise extraction of the signatures.

Signatures extracted by only one program, that were present in ≤3 samples, and had striking one/few-peak profiles never been described before, were considered artifacts. This removed 7 signatures privately extracted by HDP.

The novel SBS-RT extracted by HDP was considered for downstream analyses since it had less background noise than the one extracted by SignatureAnalyzer, favoring a higher specificity during the fitting step.

The fitting of signatures to the mutational profile at the level of subclones followed additional considerations: only the signatures that were found to be present in the corresponding sample were used, and the final step of adding SBS9 in M-CLL samples was skipped to avoid its addition in multiple subclones with low evidence.

mSigAct was used to account for mutational signatures that might have been missed due to low number of mutations and/or samples. Two cases in our study received Melphalan, which has an associated mutational signature (Maura et al., 2021; Rustad et al., 2020). Although this signature could not be extracted from our cohort, probably due to low incidence, we included it in the fitting step, as we found it was significantly present in some of our samples using mSigAct.

Clustered mutational signatures were extracted from clustered mutations. The set of clustered mutations was built on mutations with an inter-distance below 1,000 bp, as previously described (Mayakonda et al., 2018).

Mutational signature transcriptional and replication strand bias analyses were performed using the MutationalPatterns R package, which performs a Poisson test for strand asymmetry in any of the 96 mutation types (Rustad et al., 2020). Replication strand was annotated based on the left or right replication direction of the timing transition regions previously described (Haradhvala et al., 2016). The transcriptional strand was annotated using the TxDb.Hsapiens.UCSC.hg19.knownGene R package. The main peaks of the SBS-RT signature were used to determine if there was any evidence of replication and/or transcriptional strand bias associated with the SBS-RT.

We assessed the contribution of SBS-RT to coding mutations (SNVs) in RT subclones (also including tumor-only cases in which the CLL sample was used as 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 profile, as previously described (Yang et al., 2021).

The reference epigenome of CLL (Beekman et al., 2018) was used to explore the contribution of the different mutational processes in the different regulatory regions of the genome. To that aim, we simplified the described chromatin states in four categories: heterochromatin [H3K9me3_Repressed (E10), Heterochromatin Low_Signal (E11)], polycomb [Posied_Promoter (E4), H3K27me3_Repressed (E12)), enhancer/promoter [Active_Promoter (E1), Strong_Enhancer1 (E2), Weak_Promoter (E3), Weak_Enhancer (E5), Strong_Enhancer (E6))], and transcription [Transcription_Transition (E7),

Weak_Transcription (E8), Transcription_Elongation (E9)]. Besides, we utilized the high-resolution genomic replication timing data from lymphoblastoid cell lines to map the activity of mutational processes in early/late replication regions of the genome (Koren et al., 2012). We lifted over from hg18 to hg19 the replication timing data from the original publication and determined peaks/valleys of early/late replication as those regions of ≥1 Kb with absolute replication timing >0.5. All SNVs of the CLL and RT subclones (the latter including also those mutations identified in the tumor-only cases in which the CLL sample was used as germline) were classified in any of the four defined chromatin states and early/late replication regions. A cutoff of 0.005 was used to remove the less contributing signature during the fitting step instead of the 0.01 applied during the analysis of mutational signatures per sample/subclone. Only the signatures that were identified in the CLL and RT subclones were considered when analyzing the processes active in CLL and RT, respectively. An enrichment of SBS-RT mutations in any of the defined categories was tested using a log2-fold change between the observed and expected number of SBS-RT mutations per region based on their length.

The mutational signatures analysis of the case report (Results - Chapter 3: Study 5) did not include the extraction step, as the sample size was small. Instead, COSMIC mutational signatures known to be found in CLL were considered (SBS1, SBS5, SBS8, and SBS9) (Alexandrov et al., 2020; Kasar et al., 2015; Puente et al., 2011). We measured their contribution into each identified subclone using the previously described fitting approach and iteratively removing the less contributing signature if removal of the signature decreased the cosine similarity between the original and reconstructed 96-profile less than 0.01.

3.3.9 Subclonal architecture and clonal evolution

Massively parallel sequencing data can be used to characterize tumors' intra-tumor heterogeneity, reconstruct their subclonal architecture, and dissect the evolutionary trajectories of the disease. Variant calling of somatic mutations in bulk samples can give information on the prevalence of each variant (see Introduction - section 1.3.2). These frequencies can be used to infer the events and mutational processes that came first, and those that are acquired later during the disease evolution (Landau et al., 2013; McGranahan et al., 2015; Nik-Zainal, Van Loo, et al., 2012).

Each tumor sample represents a "snapshot" taken along a temporal and spatial axis, which can be used to characterize the tumor's alterations at a specific moment. On top of that, the frequency at which these mutations are identified within this finite time point can be used to reveal the temporal order of acquisition of these events (Figure 50). This reasoning is mainly based on SNVs, which are easier to detect and, thus, more reliable. The VAF of a mutation i is calculated as the fraction of reads that support the mutated allele:

$$VAF_i = \frac{Mutated\ reads_i}{Total\ reads_i}$$

Next, the VAF is adjusted by the local copy number (N_T) and tumor purity (p), the fraction of tumor cells within the sample. This gives an estimate of the fraction of tumor cells carrying the mutation, the cancer cell fraction (CCF). The following formula can be used to calculate the CCF of a mutation i (Dentro et al., 2017):

$$CCF_i = \frac{VAF_i}{m_i p}[pN_{T,t,i} + (1-p)N_{T,n,i}]$$

Where $N_{T,t,i}$ is the number of chromosome copies in tumor cells at locus i, $N_{T,n,i}$ is the number of chromosome copies in normal cells at locus i, usually 2, and m_i is the mutation multiplicity.

These values can be used to classify variants as clonal, when they are present in all tumor cells, or subclonal, when they are only present in a portion of tumor cells. Clonal mutations represent early events, as they occurred at, or prior to, the most recent clonal expansion. Conversely, subclonal mutations represent later events.

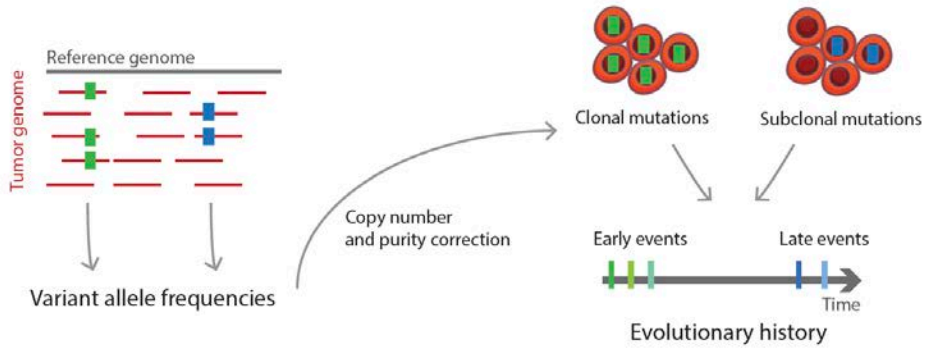


Figure 50. Strategy to infer the timing of somatic events from bulk sequencing. Frequencies of somatic mutations (represented by green and blue small rectangles) can be used to determine their clonality and their relative timing.

We can also go one step further, and use the CCF of all variants within a sample to determine its subclonal architecture (Dentro et al., 2017). Clustering of CCF can identify clusters of mutations with similar frequencies that estimate the distinct tumor cell subpopulations that are present at the time of sampling. Next, the phylogeny among subclones can be inferred from the CCF of each cluster, or subclone. The trunk constitutes the set of mutations that are shared by all cells, the founding clone, and the branches represent the variants that are

acquired later during tumor evolution and correspond to subclonal diversification from the parental clone (Figure 51).

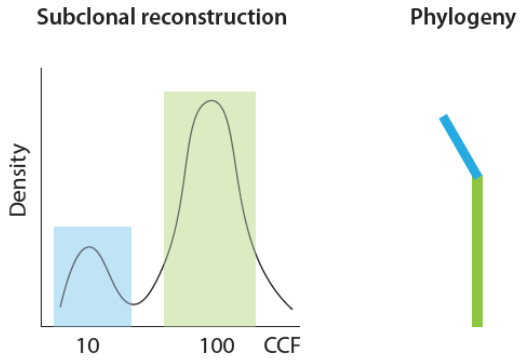


Figure 51. Subclonal reconstruction based on clustering of CCF. The subclonal architecture of tumors can be inferred from the clusters of mutations with similar CCF (left). The phylogeny of the clusters can be inferred from their CCF and it is represented as a tree (right), whose branches' length is proportional to the number of mutations within that cluster.

Subclonal reconstruction based on single-sample analyses has limitations and can underestimate ITH (A. Davis et al., 2017; Dentro et al., 2021). Variants identified as clonal in one sample might be subclonal in another spatial sample from the same patient. Temporarily, multiple samples can also dissect clusters of mutations that might have similar frequencies at one time point but diverge during tumor evolution (Figure 52).

Comprehensive spatial and/or longitudinal studies can yield more accurate ITH assessment and, thus, more reliable insights into tumor composition and evolution. However, as in any other NGS analysis, previous steps can introduce errors that can influence posterior analyses. Sequencing, alignment, and variant calling can drag uncertainties into the clustering of subclones and the inference of their phylogenetic relationship. Moreover, the VAF of SNVs can be noisy and influenced by local sequencing depth, especially in low coverage samples, where

the difference of one mutated read can significantly vary its VAF and, therefore, its CCF. To account for this variability, a binomial distribution can be used to model this noise and capture the effect of copy number state, read depth, and variant frequency.

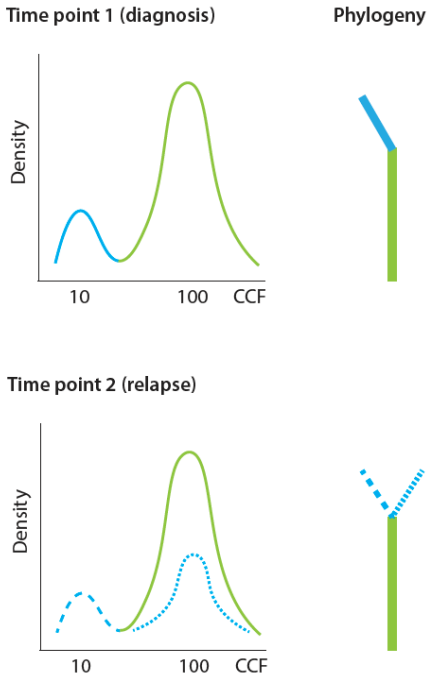


Figure 52. Subclonal reconstruction using multiple time points. Multiple samples can better estimate the subclonal architecture of tumors, dissecting subclones that might have similar frequencies at one time point (blue subclone in Time point 1) but that differ during tumor evolution (blue dashed subclones in Time point 2).

The subclonal architecture of tumor longitudinal samples in Study 4 and Study 5 was reconstructed from WGS data using a Bayesian clustering method named DPClust in the PCAWG project (Bolli et al., 2014; Dentro et al., 2017; Maura, Bolli, et al., 2019; Nik-Zainal, Van Loo, et al., 2012). It uses a hierarchical Bayesian Dirichlet process to model the mutations as deriving from an unknown number of clusters, or subclones. The properties of the clusters, the fraction of

tumor cells they represent, and the number of mutations they contribute, are also unknown. This approach can jointly estimate all unknown parameters. A binomial distribution models the mutated reads and accounts for read sampling variation while integrating copy number states and tumor purities. A Markov Chain Monte Carlo (MCMC) is used to infer putative subclones, to assign mutations to subclones, and to estimate the subclone frequencies in each sample. The MCMC samples were run for 10,000 iterations, and the first 5,000 were discarded.

The phylogenetic ordering of the subclones was inferred from the subclone frequencies in each sequential sample following the "pigeonhole principle", as previously described (Maura, Bolli, et al., 2019). This principle states that the sum of CCF of branching subclones should not be greater than the parental clone, assuming the infinite sites hypothesis (mutations occur only once and never revert to wild type). In linear evolution, the smaller subclone must be a descendant of the bigger subclone. For any possible subclone phylogeny, the pigeon principle must be followed in each sample. In multi-sample approaches, this further constrains the feasible relationships between subclones. A tolerated error can be introduced to make these principles more permissible for noisy data. We used values between 0.001 and 0.05 to account for each case's variability. Clusters with less than 100 (Study 4), 50 (Study 5), or not assigned to the reconstructed phylogenetic tree were excluded. The length of each tree branch in the phylogeny is proportional to the number of mutations assigned to the corresponding subclone. TimeScape R package (v1.6.0) was used to plot the fish plots that show the clonal evolution of tumors.

3.3.10 High-coverage, UMI-based NGS

Data analysis was performed following manufacturer's recommendations. Briefly, paired reads were trimmed using cutadapt (Martin, 2011). Trimmed

FASTQ reads were converted to unmapped BAM using Picard's FastqToSam tool. UMI information was extracted and stored as a tag using fgbio.jar ExtractUmisFromBam. Template reads were converted to FASTQ with Picard's SamToFastq. Next, template reads were mapped against the human reference genome (GRCh37) and the reads were merged with the UMI information using Picard's MergeBamAlignment.

Finally, reads were grouped by UMI and a consensus was called on grouped reads using fgbio.jar GroupReadsByUmi and CallMolecularConsensusReads, respectively. Note that a minimum of 3 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. Finally, mean coverage was determined using Picard's CollectTargetedPcrMetrics tool. Mean coverage was 23,805x.

Read counts were collected at all targeted genomic positions for all samples using bcftools mpileup. The versions and parameters used for each tool are detailed in Table 7. A custom script was used to parse the depth (DP) as well as reference and alternate supporting reads (AD). Allele frequencies from positions lacking mutations by WGS were used to model the potential background sequencing noise, which was unified according to the tri-nucleotide context of each mutation (considering the flanking 3' and 5' bases together with the variant itself). The presence/absence of the mutations of interest was assessed according to the background noise of their tri-nucleotide context, and they were annotated as high-confidence when their frequency was above the background with a probability of 95%. Mutations with supporting reads but below the 95% threshold were considered as low confidence. Variants were classified as not present when the alternate allele had no supporting reads.

Table 7. Program versions and parameters used for high-coverage UMI-based analysis.

Program	*Version*	*Parameters*
cutadapt	v1.15	-g CCTACACGACGCTCTTCCGATCT -a AGATCGGAAGAGCACACGTCTGAA -A AGATCGGAAGAGCGTCGTGTAGG -G TTCAGACGTGTGCTCTTCCGATCT -e 0.1 -O 9 -m 20 -n 2
FastqToSam, *MergeBamAlignment*	v2.10.2	default parameters
ExtractUmisFromBam	v1.0.8	--read-structure=16M+T 16M+T --single-tag=RX --molecular-index-tags=ZA ZB
GroupReadsByUmi	v1.0.8	--strategy=adjacency --edits=1 --min-map=10
CallMolecular *ConsensusReads*	v1.0.8	--min-reads=3
BWA-MEM	v0.7.15	default parameters
CollectTargetedPcrMetrics	v2.10.2	CLIP_OVERLAPPING_READS=true MINIMUM_MAPPING_QUALITY=15 MINIMUM_BASE_QUALITY=15
bcftools mplileup	v1.8	-B -Q 13 -q 10 -d 100000 -a FORMAT/DP, FORMAT/AD, FORMAT/ADF, FORMAT/ADR -O v

3.3.11 Bulk RNA-seq

Like other NGS techniques, bulk RNA-seq produces a collection of mixed reads with unknown locations. Here, the mapping against the reference genome is more challenging as it must deal with the non-contiguous transcript structure and short read lengths. After the alignment, multiple analyses can be performed including gene expression, splicing events, gene fusions, and even variant calling. During this thesis we conducted analyses for differential gene expression (Results - Chapter 3: Study 4, and methodological contributions – see Appendix) and differential splicing (see Appendix), which will be explained in this section. All the methods included in these analyses and their versions used can be found in Table 8.

First, quality assessment and trimming or filtering of reads was performed to retain only good quality and informative reads. RNA-seq preparation kits include procedures and best practices to deplete ribosomal RNA (rRNA) from the total RNA before sequencing. However, experimental techniques are not bullet-proof, which calls for posterior examination and repair. To verify if the resulting material is suitable for downstream analyses, we first applied rRNA filtering by running SortMeRNA (Kopylova et al., 2012), a local sequence alignment tool that can be used for mapping and removing rRNA contamination. Non-ribosomal RNA reads were subsequently trimmed for sequence adapters and low quality bases using Trimmomatic (Bolger et al., 2014), a tool for custom quality trimming and adapter clipping. Reads were scanned with a 4-base wide sliding window and cut when the average quality per base dropped below 20, Illumina adapters were removed, and reads with a minimum length of 50bp were kept. Quality of the original and trimmed reads was assessed using FastQC, and MultiQC was used to visualize all reports.

At this point, RNA reads are ready to be analyzed. For differential splicing analysis, we used LeafCutter (Y. I. Li et al., 2017) to quantify splicing variation by leveraging spliced reads (i.e., reads that span an intron) to evaluate differential intron usage across input samples. For gene expression analysis, two different approaches can be applied: alignment of the reads against a reference genome and posterior counting of gene-level expression, or direct transcript quantification from FASTQ files performing a pseudoalignment against a reference transcriptome for rapidly determining the agreement of reads and targets, without the need for prior alignment. With respect to the first option, we mapped the reads using STAR (Spliced Transcripts Alignment to a Reference) (Dobin et al., 2013), an aligner specifically designed to address the difficulties of RNA-seq data mapping that uses a two-step strategy: it starts with a sequential maximum mappable seed search, followed by a seed clustering and stitching step. Next, we

used the htseq-count tool from HTSeq (Anders et al., 2015) to generate gene-level expression read counts by calculating the overlap of reads with genes. Regarding the second approach, we used Kallisto (Bray et al., 2016) to quantify abundances of transcripts and tximport (Soneson et al., 2016) to summarize transcript-level estimates into gene-level counts (GRCh38.p13, Ensembl release 100).

In Study 4, we applied this last methodology, and conducted a paired differential expression analysis using DESeq2 (Love et al., 2014). To detect genes with changes in expression, regardless of low or highly variable read counts, a shrinkage of effect size was performed using the "apeglm" (Approximate Posterior Estimation for generalized linear model) method (Zhu et al., 2019). Differentially expressed genes were determined by an adjusted P value (Q) <0.01 and absolute log2-transformed fold change >1.

Table 8. Tools used for bulk RNA-seq analyses.

Program	*Version*	*Reference*
FastQC	v0.11.5	www.bioinformatics.babraham.ac.uk/projects/fastqc
SortMeRNA	v4.3.2	https://github.com/biocore/sortmerna
Trimmomatic	v0.38	https://github.com/usadellab/Trimmomatic
MultiQC	v1.7	https://github.com/ewels/MultiQC
STAR	v2.6.0c	https://github.com/alexdobin/STAR
HTSeq	v0.11.0	https://htseq.readthedocs.io/en/master/
Kallisto	v0.46.1	https://github.com/pachterlab/kallisto
tximport	v1.14.2	https://bioconductor.org/packages/release/bioc/html/tximport.html
LeafCutter	v0.2.8	https://github.com/davidaknowles/leafcutter/
DESeq2	v1.26.0	https://bioconductor.org/packages/release/bioc/html/DESeq2.html

4 Results

The results are presented following the outline of this thesis, as previously described (see Introduction - section 1.1 Thesis trajectory), and are divided into three chapters, according to the thesis' three main blocks (Figure 53). Each result corresponds to a Study, which also contains a brief introduction.

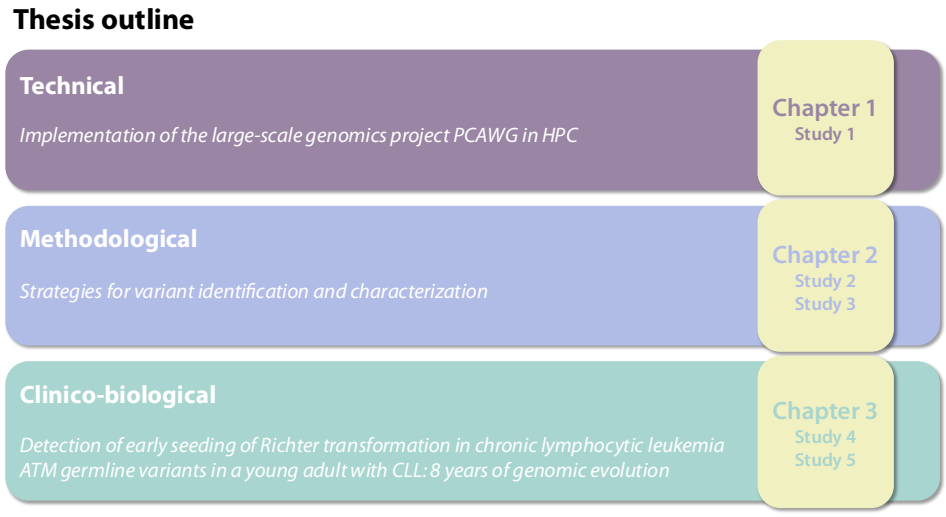


Figure 53. Thesis trajectory. Contents of the chapters within the Results section, which correspond to each part of the thesis.

First, in Chapter 1, the results obtained within the PCAWG project are presented. Here, we dealt with the most technical aspects, including new emerging technologies that defied the BSC's HPC restrictions at the time. At this point, I performed NGS analysis, but did not enter into their design nor the evaluation of their results. In Chapter 2, the results obtained within the MedPerCan project are explained. At this point, we went into the methodology itself, evaluating the performance of variant calling strategies and deepening into the characterization of somatic variation. This chapter also includes the extension of the protocols developed within the MedPerCan project to meet the needs of our studies presented in Chapter 3. Finally, in Chapter 3, these methodologies,

and the lessons learned, were applied to studies of the progression of CLL in a close collaboration with Dr. Elías Campo's group at Hospital Clínic de Barcelona/IDIBAPS. This last chapter, where we went beyond the computational aspects and engaged in the biological significance of the results, concludes the trajectory of this thesis.

4.1 Chapter 1: The Pan-Cancer Analysis of Whole Genomes infrastructure

4.1.1 Introduction

The Pan-Cancer Analysis of Whole Genomes (PCAWG) was a collaborative effort emerging from the ICGC project that joined together more than 2,800 cancer whole genomes coming from different countries and subprojects to homogeneously analyze genomic features across 38 tumor types. The final aim was to identify features associated to cancer processes, beyond each specific tumor type. Previous studies from the ICGC generated volumes of NGS data and obtained remarkable results, but each one of them focused only on one individual cancer type. At the same time, the TCGA repertoire of NGS cancer-related datasets was also growing and they envisioned a large-scale collaboration, the Pan-Cancer Atlas, that interconnected mainly whole-exome sequencing analyses of the 33 most common tumor types to increase our understanding of how tumors arise in humans (L. Ding et al., 2018; Hoadley et al., 2018; Sanchez-Vega et al., 2018).

Within the new PCAWG initiative, the goal was to jointly analyze these valuable data, coming from both the ICGC and the TCGA, and concentrate on comprehensive whole-genome sequencing analysis to better understand the molecular processes behind the origin of tumors and their progression, identifying

shared and uncommon features across cancer types. At that time, it was the largest effort done so far to put together and analyze this number of tumors. The integration of the data at this large-scale level implied many technical and conceptual challenges that were addressed during the project. Because of the difficulties imposed by the management and analysis of the data, and the complexity of the different biological implications of the results, this consortium generated different groups devoted to specific areas and tasks.

At the technical side, the PCAWG created a technical working group to coordinate the development of uniform portable software, perform uniform analyses on ~1PB of sequencing data distributed across different geographical locations and using a variety of computational resources, and to ultimately provide the community with high-quality and validated consensus variant catalogs to find answers to specific biological questions around cancer formation and progression.

As this large PCAWG cohort was derived from all the independent ICGC studies done in each country, these datasets had already been analyzed within their own jurisdictions. Ideally, the merging and unification of the different analysis (i.e., the VCFs with somatic variants) previously generated should theoretically allow us to derive biological conclusions, but the fact that these prior results came from different methodologies for variant discovery, which showed a vast heterogeneity among them, made their unification not possible from the interpretation point of view. Thus, to eliminate variations that arise from discrepant analysis and to ensure an accurate integration of the results for downstream analysis, a uniform analysis on all samples was required. This homogenized analysis was done using metadata conventions for describing raw sequencing data, and a standardized set of pipelines covering the alignment of sequencing reads, three variant calling pipelines, and filtering and merging

strategies of their results validated through target deep sequencing (Figure 54). All these analyses were performed within 13 data centers selected among the partners, which included the BSC. Each data center was devoted to part of the analysis and had to be aligned with the other data centers. These core workflows yielded high quality and harmonized somatic variants from all tumor genomes for downstream working groups to explore the biology of cancer.

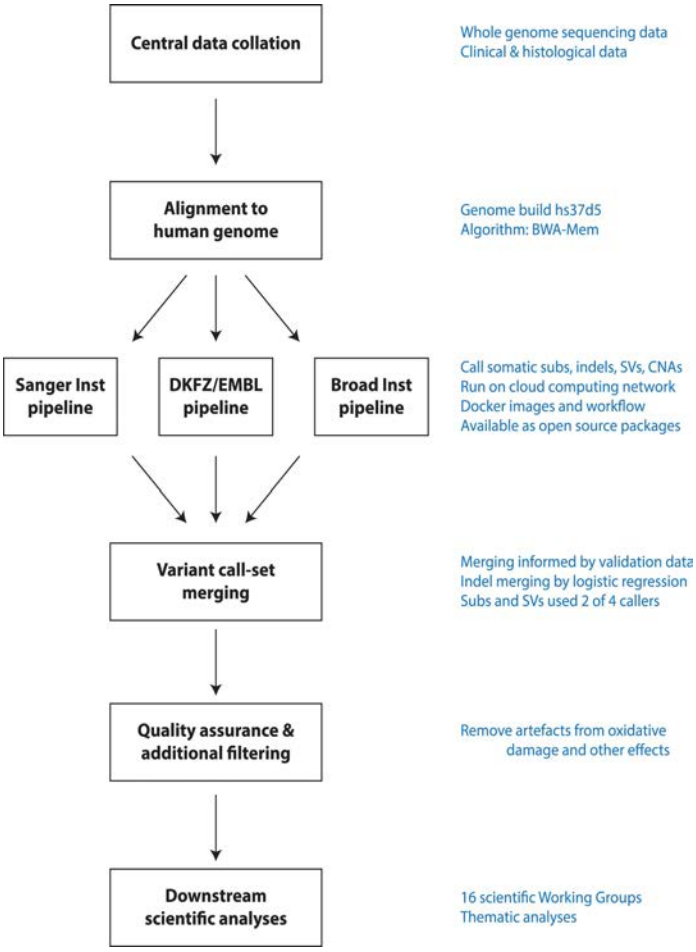


Figure 54. Flow-chart showing key steps in the analysis of PCAWG genomes. After data collection, alignment to the human genome was performed, somatic mutations were identified by three pipelines, and subsequent merging into a consensus variant set was used for downstream scientific

analyses. Subs, substitutions; DKFZ/EMBL, German Cancer Research Centre/European Molecular Biology Laboratory. Image from Campbell et al., 2020.

This (re-)analysis implied the need for dedicated infrastructures allowing automatic executions at different computing sites worldwide. From the technical point of view (Figure 55), the project started with the reformatting of sequencing data, its annotation with standardized metadata, and its submission to one of the processing data centers. Submitted data was subject to homogeneous primary analysis that included 1 WGS alignment, 3 variant calling pipelines, and 2 RNA-seq alignments. Overall, around 14 computing clouds and HPC facilities were utilized over 2.5 years. Aligned reads and variant calls were finally made available to 16 working groups for downstream analysis, who produced over 20 publications in *Nature* and affiliated journals. Finally, the generated data was made available to the broader community through the ICGC Data Portal.

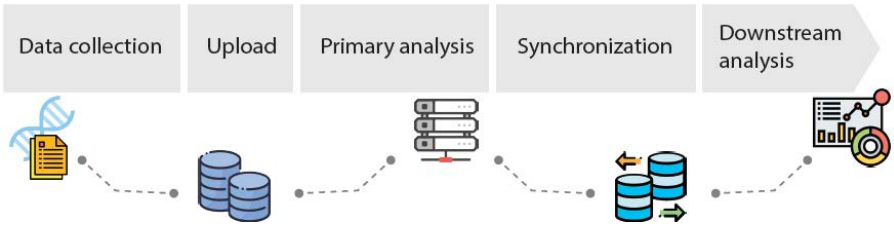


Figure 55. PCAWG project main steps. The project started with the collection, formatting, and annotation of sequencing data that was uploaded to the data processing centers. Next, the pipelines of the primary analysis were executed. The results were synchronized among the data centers, who provided the data to the working groups that performed the downstream analysis.

Each of these steps brought about technical requirements and called for innovative solutions to complete the activities of the project.

To start with, the PCAWG dataset's size and nature presented significant methodological and data management challenges. First, genomic data across

countries was subjected to different jurisdictions that imposed particular restrictions of some cohorts on geographical storage and cloud-based processing. For example, the data generated in the United States (generated within the TCGA project) could not be, by law, analyzed and re-distributed from non-US countries. This already imposed several challenges to the project and demanded specific solutions that consisted in distributing the management and analysis facilities across different countries, including the US. The size of the final dataset was estimated to be around ~1PB, but during run-time temporary files occupied more space, way past that number.

Storage, legal constraints, and compute requirements to analyze such a large dataset made it unworkable to complete the analysis at a single center. All these limitations had to be considered to organize the executions of the analyses on a network of compute sites, including academic and commercial clouds, and more restrictive HPC environments, like in the case of the BSC. The distribution of workloads among different centers made necessary the use of orchestrators and the development of portable workflows that could be transparently ported to different compute environments and architectures, providing consistent and comparable analysis results independently of the underlying platform.

As a major orchestrator of data management across the different analysis centers, we originally selected GNOS, a software made by Annai Systems that could coordinate several compute sites, which would then act as nodes and as GNOS serves. Each compute site that offered storage allocation became a GNOS server capable of accepting unaligned BAMs, aligned BAMs, and variant calling VCFs. GNOS repositories were connected and synchronized along the project in order to facilitate access to all the raw and generated data from multiple cloud environments.

At the start of PCAWG, cloud computing at scale was still novel in the cancer bioinformatics community. Most alignment and VC pipelines were executed on local and HPC clusters, and containerization (e.g., Docker or Singularity) was still a technology in their early stages. The project had to struggle with all these emerging frameworks as it matured. Over the years, pre-defined virtual machines with embedded workflows were converted into docker containers that could be run in HPC clusters more conveniently, as well as in cloud infrastructures.

Upon completion of the primary analysis, the last stage of the project involved the participation of researchers organized in thematic working groups that needed to access the data. Downloading hundreds of terabytes of data is an unsustainable model, as it is only feasible for few research centers that have enough storage and compute facilities. With this in mind, PCAWG acknowledged the need to co-locate the data with compute resources, so that researchers could bring their analysis and methods to the data, instead of the other way around.

The project embraced the use of cloud technologies, and the majority of computational sites opted to install or adapt cloud infrastructures for the project, which made their deployments much easier. Unfortunately, this was not possible at some HPC-based institutes at the time. Initially, SeqWare was used as a mechanism to encapsulate analytical workflows so that they can be run in a variety of sites. SeqWare bundles were used to deploy the workflows in worker VMs that were instantiated at each computing center. At this stage, in order to solve the sparse distribution of centers, the project decided to use cloud computing and specific virtual machines with all the required dependencies. This strategy was adopted to ensure that the same pipelines and processes that were run in each analysis center could provide the same results, independently of their architectures. Over time, Docker allowed the technical group to leverage non-cloud environments in the project as well. This lightweight virtualization

technology complemented the clouds and ensured that workflows worked identically across the diverse compute environments that were part of the project. HPC clusters could then be used by the project, even if they did not provide cloud services, as long as they could be modified to enable Docker container executions. PCAWG workflows included many PCAWG-specific elements that limit their usability outside of the project, and even within the project at more restrictive centers. For example, many workflows assume that input data should be staged in from GNOS repositories and, likewise, the results should be uploaded back to GNOS. To enable the execution in limiting environments (e.g., without internet access) and the long-term usability beyond the project, most workflows were extended or simplified to be run under these settings.

The project represented one of the major initiatives at the time. Although a (big) part of the effort was dedicated to the technical aspects, the central goal of the project was, of course, to shed light on the biological processes driving cancer. On the whole, the integrative analyses of whole-genome sequencing studies brought us closer to the understanding of the causal molecular alterations of cancer. Analyses of non-coding somatic mutations could identify the driving role of non-coding point mutations and structural variants to the cancer phenotype and the role of germline variants in patterns of somatic mutations. Further integration with transcriptomic data provided a comprehensive catalog of RNA-level alterations in cancer and reported the effect of somatic alterations on transcription. Inference of tumor composition and evolution recognized the ubiquity of ITH and different evolutionary trajectories across multiple cancer types. It also gave us insights into the timing of DNA alterations and mutational processes that shape the cancer genome, revealing that driver mutations can be already present years before diagnosis (Alexandrov et al., 2020; Calabrese et al.,

2020; Campbell et al., 2020; Gerstung et al., 2020; Y. Li et al., 2020; Rheinbay et al., 2020).

4.1.2 Study 1: Implementation of the PCAWG infrastructure at the BSC

*During the PCAWG project I was involved in all the data-oriented tasks assigned to the BSC, which is summarized in (*Figure 54*) and included the main steps of WGS analysis. Additionally, as part of the Spanish ICGC-CLL consortium, the CLLGenome project, the BSC was in charge of managing the data derived from this project and of its contribution to the PCAWG project. More precisely, my work consisted in:*

- *Preparation of harmonized raw sequencing data and metadata and submission to the GNOS server of the CLL dataset, which included 100 donors with whole-genome sequencing, and a subset with RNA sequencing.*
- *Execution of the core alignment and variant calling pipelines (Sanger and DKFZ/EMBL) at the HPC premises of the BSC.*
- *Upload of the alignment and variant calling results to the corresponding GNOS servers.*
- *Synchronization of the GNOS server throughout the project's lifetime and assistance to the EGA submission for the long-term archival.*

I was responsible of the technical coordination at the BSC, identifying the conflicting points between our HPC center and the cloud-based environment that the project required, especially at the initial phase. Working together with our Operations team and the PCAWG technical working group, we had to find a middle point for each matter that could risk our security and mitigate the potential vulnerabilities of virtualized technologies in our pure HPC infrastructure. The workflows and their adaptive modifications were developed and provided by other partners within the project. All operation services that required root access for

setting-up were conducted by the system administration team at the BSC. GNOS developers provided support to install GNOS instances at our site and implemented additional synchronization scripts that were specific for our center. All tasks that could be done by an unprivileged user, including data submission, automation and executions of pipelines, data uploads, and GNOS operations, were done by me, and I was also part of all the discussions to adapt the project requests into our system.

The HPC infrastructure at the BSC has strict policies that directly conflict with the requirements and logistic plans of PCAWG. For example, virtualization is not allowed in HPC clusters, and their computing nodes have no external network access. Solutions between cloud-based systems and traditional HPC had to be carefully devised. Two approaches to allow VMs, at the initial phase of the project (phase 1), and docker containers, at the last phase (phase 2), were implemented at the BSC.

As a data center of the project, the BSC set-up a GNOS server with external access used for data submission. This work was done by the operations team with the support of Annai Systems. Regular updates had to be applied to the system, including patches that required root access, and other upgrades and data fixing issues that I managed.

At the beginning of the project, to carry out the executions of the primary analysis at each computational site, workflows were packaged within VMs so they could be easily distributed, following the need of executing exactly the same methodology in each data center. To allow the deployment of VMs, an alternative infrastructure had to be implemented in our center (Figure 56). An isolated cluster was created by decoupling a whole rack of MareNostrum3 (MN3) from the rest of the machine, so that black-box VMs could not affect or endanger the performance of the rest of the system.

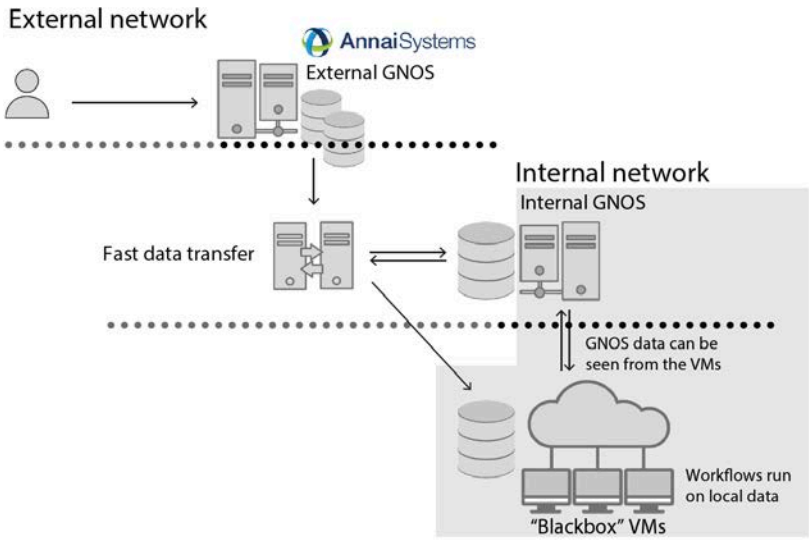


Figure 56. PCAWG infrastructure at the BSC during phase 1. Set-up at BSC's HPC premises included an isolated rack of MN3 where VMs could be deployed, an internal GNOS server that was accessible from the VMs, and an external GNOS server that was reachable from the external network and was used for data submission and synchronization with the other PCAWG data centers. Synchronization between the external and the internal GNOS was done through a data transfer server that had access to the filesystems of both servers.

The detached nodes were completely isolated from the rest, and they only had an internal network among them. This implied that they could not access any other machine that was not part of this isolated cluster, including the GNOS data server. Thus, input data had to be copied from the disk of the BSC GNOS server to a filesystem that could be seen from the VMs. This was done by using a fast data transfer machine, which was connected to both disks. In addition to that, the workflows also required access to the BSC GNOS server to retrieve metadata and validate the outputs' metadata. To bypass the non-allowed connection between the VMs and the BSC GNOS server, an internal GNOS was set up. This internal GNOS was in the same internal network as the VMs, so it could be reached from them. Both the internal GNOS and the external BSC GNOS had to be synchronized

and specific scripts to import and export the data between them had to be developed by Annai Systems. From the HPC perspective, the downfall of this solution is that, since these nodes were physically separated, they could not be used by any other HPC application when not in use.

Whereas during phase 1 the use of VMs forced computing nodes to be physically isolated from the HPC supercomputer, in phase 2, the use of docker containers required no infrastructure changes and they could be launched as any other traditional HPC application (Figure 57). Docker containers are lighter than VMs, which makes them easier to tune and make them HPC compliant. However, due to security reasons they are usually not allowed. To allow highly exceptional Docker executions, the containers had to be forced to be run as a regular user, which was created specifically for the project, and who had no extra permits that could expose the rest of the HPC cluster. We had to audit the project's Docker images together with our system administrators, and UID (user id) settings matching an unprivileged user of the machine assured that everything inside the container ran without root access. Only supervised images were allowed to be run, and an unchangeable wrapper script around the Docker execution itself assured this point. Docker containers could only be run using this script, which only allowed audited images, and that could only be executed by the PCAWG user.

With this, docker executions were allowed and were integrated into the queue system (LSF), which also had to be adapted to enable container executions. A special queue, restricted to a single user, had to be created with special settings (e.g., maximum running jobs, extended maximum wall time). Prolog and epilog scripts that can be run with root permissions had to be developed to prepare and clean the docker environment where the workflows were executed. The prolog is executed before the user execution starts, and it was used to start the Docker

engine and load the approved docker images. After completion of the execution, the epilog cleaned the environment and stopped the Docker server.

As explained above, most of the workflows came with the assumption that input data could be downloaded from a GNOS server and that the outputs could be uploaded back to the server. However, our infrastructure did not allow external connections of any kind, and the workflows had to rely on input data that was already present at the GPFS disks. Similarly, output data could not be sent to external GNOS servers and had to be stored at the filesystem instead. This restriction entailed the adaptation of the workflows to work with local data, and the implementation of a strategy to first gather input data and place it at the local filesystem, and later collect the results from disk and upload them to the appropriate GNOS server.

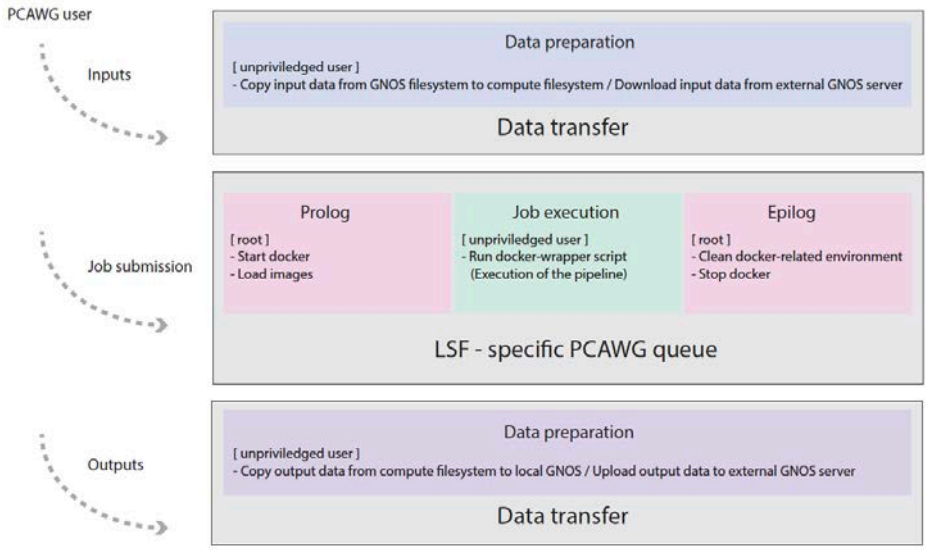


Figure 57. PCAWG Infrastructure at the BSC during phase 2. Docker allowed a more integrated solution for workflow executions. They could be run in regular compute nodes but needed a special queue that performed additional settings to start and finish a docker-enabled environment. Input and output data had to be managed independently as compute nodes do not have internet access.

Overall, three core pipelines were run at the BSC: the alignment pipeline of whole genomes, during phase 1, and two variant calling pipelines (Sanger and DKFZ/EMBL), during phase 2. We performed 10.2% of the alignments, and 28.5% and 17.2% of the DKFZ/EMBL and Sanger VC pipelines, respectively. The number of VC executions in all PCAWG centers throughout the project can be seen in Figure 58 and Figure 59, while the computational resources used at our center are detailed in the Methods (see Methods - section 3.1). Of note, the third variant calling pipeline (Broad) had a part of a proprietary software and could not be run outside of the developer's center. Moreover, TCGA data could only be stored and distributed from the United States (US) and, initially, it was planned to be analyzed there. However, the project required more computational nodes that could analyze these data as there were not enough resources at the time within the US. Logistics were done to allow international executions: external centers had to apply for the credentials to download the data, run the analyses, and upload the results back to a US site. The BSC ran many executions on TCGA data, retrieving and uploading data from/to a US-based GNOS server.

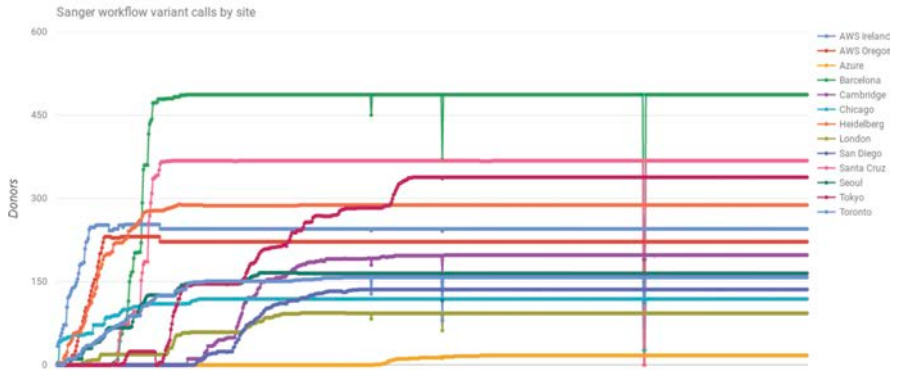


Figure 58. Executions of Sanger workflow by site. Number of completed executions of the Sanger variant calling pipeline at each site during the project.

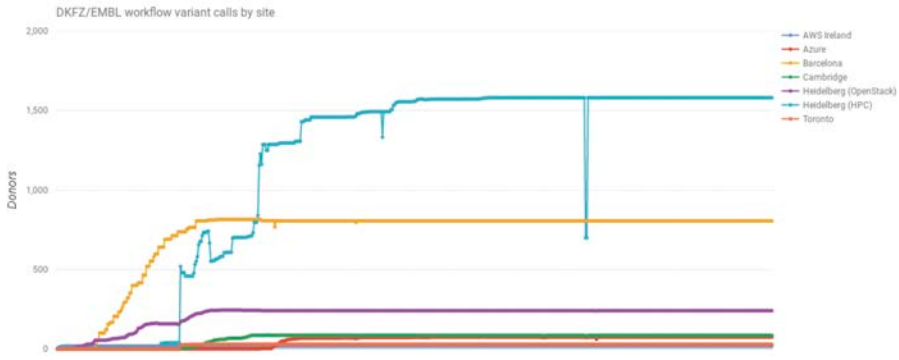


Figure 59. Executions of DKFZ/EMBL workflow by site. Number of completed executions of the DKFZ/EMBL variant calling pipeline at each site during the project.

4.2 Chapter 2: Framework for variant characterization in tumor genomes

4.2.1 Introduction

The identification of somatic variants in tumors (i.e., variant calling) is at the root of cancer genomics analysis, for both research and clinical applications. Researchers rely on variant calling results, which are at the bases of any further downstream analyses, to conduct their investigations. In clinical settings, the detection of variants can be used to identify biomarkers for diagnosis, prognostic indicators, or to guide treatment decisions, and can have a great impact as NGS is progressively introduced into healthcare routines.

The MedPerCan project devised the applicability of such analyses in the clinics and set a pilot project that would include all actors taking part in this process, from sequencing centers to data analysis centers and clinicians as end-users. The aim of the project was to create a prototype to evaluate the impact of genomic analysis in clinical decision-making in oncology. Within this context, we

designed, evaluated, and implemented variant calling strategies to analyze WES (Results - Chapter 2: Study 2). Forthcoming collaborations, during and after the project, guided the continuity and enhancement of the initial pipeline for the detection and characterization of small variants in WES throughout the project and past its completion. The extension of this framework (Results - Chapter 2: Study 3) included the integration of new data types, such as WGS or transcriptomics, and new analysis tools responding and adapted to the needs of scientific questions. All these methodologies have been applied to several published studies (see Appendix) and, more importantly, they have also been applied to a capital part of the thesis, where the effort and focus was not only on the technical and methodological aspects but rather on the biological interpretation of their results. Within this last part, we extended the cancer genomics analyses to cover questions related to the clonal dynamics of tumors and the mutational processes that may act during this evolution (Results - Chapter 3: Study 4 and Study 5).

4.2.2 Study 2: Variant calling strategies in MedPerCan

During the MedPerCan project, I coordinated the design, evaluation, and set-up of variant calling strategies for whole-exome data. Montserrat Puiggròs, Héctor Gracia, and Álvaro Férriz from the Computational Genomics group at BSC were also involved in the project. I installed the tools, with the help of the Support team at BSC, in MareNostrum4, Nord3, and StarLife, and automatized and performed their executions. I also implemented benchmarking runs to assess their performance on different settings. All WES data from the project was subjected to uniform analysis using these pipelines. Hereafter, I will present the work that I carried out during the project.

In the next sections, first, there is a brief introduction of the MedPerCan project from the organizational point of view, laying out the specific role of the BSC within the project. Next, the results of the work carried out at the BSC, which include the design of methodologies for variant identification and characterization, are presented, followed by their application to representative tumor types. Please note that some of the work presented here overlapped with the first phase of Study 3. Some benchmarking efforts were done in parallel and complemented each other.

4.2.2.1 Introduction

MedPerCan was a pilot study to assess the impact of genomic analysis in clinical decision-making in oncology and to serve as a model for the implementation of personalized medicine in Catalonia. During the project, we developed a multidisciplinary and multi-institutional circuit that started from the patient, who would go to a hospital, where they would collect specimens for genomic analyses. From there, the samples would be sent to a sequencing center, where they would sequence the samples and send the results to a data or analysis center. There, the bioinformatics analyses would be carried out, and the variants of the normal and/or tumor samples would be identified. Next, the variants would be annotated and populated into a database that would be made available to the clinicians through a user-friendly web page. The tumor board would browse the web portal, evaluate the results, and come up with some recommendations that would be sent back to the doctor. Finally, the physician would use this information for better diagnosis or improved treatment options.

Our group at the BSC acted as the data and analysis center that would perform the bioinformatics analysis to detect and characterize variants in sequenced data, more specifically, in WES. Thus, the role of the BSC within this

network was to design, evaluate, and implement a framework for variant calling and perform this analysis on the data generated by the project.

4.2.2.2 Results

Pipeline for variant identification and characterization in WES

The main pipeline developed within the MedPerCan project included a multi-caller approach for somatic variant calling of small variants (SNVs and indels) on normal-tumor paired WES data (Figure 60). Most of the datasets generated by the project fitted this analysis, but we also had to implement satellite workflows to accommodate other datasets, such as samples from patient-derived xenograft (PDX), tumor-only analyses, and special considerations for FFPE samples.

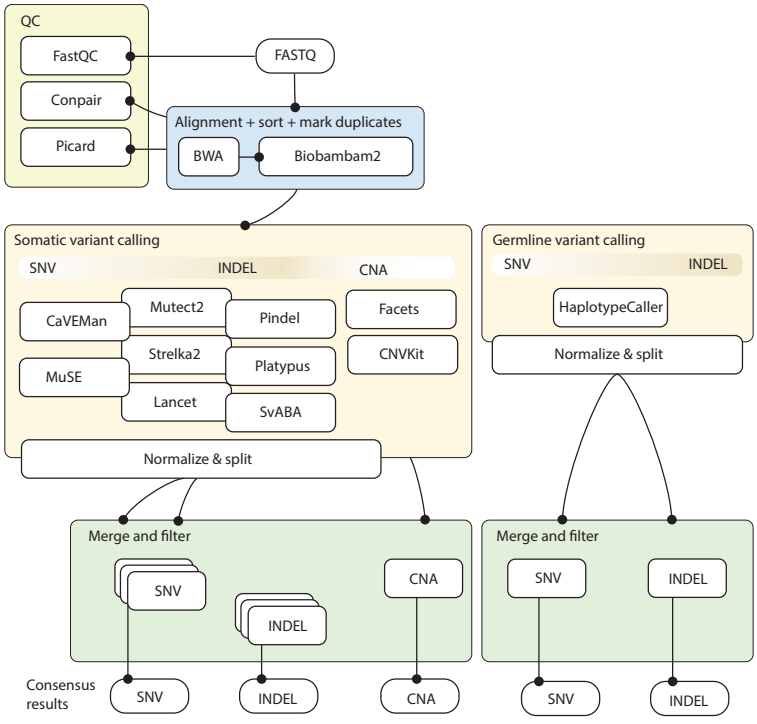


Figure 60. Main pipeline for variant identification and characterization within the MedPerCan project.

Additionally, germline variant calling was also performed on all normal samples, as some subprojects were based on inherited cancer. The results of germline variant calling benchmarks will not be shown here since they were performed by Álvaro Férriz. In brief, we evaluated different filtering strategies for the HaplotypeCaller program from GATK, including hard filters based on quality-based features (e.g., mapping or base qualities, strand bias, or read depth) and the Variant Quality Score Recalibration (VQSR) by GATK. We evaluated the results using the Platinum genomes from Illumina (Eberle et al., 2017) and, based on their performance, we selected the following hard filters: read depth <8, fisher strand bias >25.0, quality by depth <6.0, and RMS mapping quality <50.0.

The tools included in the somatic variant calling pipeline, as well as their different combinations, were benchmarked using different datasets. The results shown here are based on real data, as it captures the intricacies of tumor's complex biology as well as potential artifacts coming from real sample processing and sequencing.

For the benchmarking of the pipeline, at that time, to our knowledge, there was only one published dataset with a comprehensive characterization of somatic variants that included WES data (Griffith et al., 2015). Unfortunately, this dataset only included validated SNVs. We used the WES normal and tumor samples from this study (case AML31), and downsampled them to the average coverage that was being used within the project (i.e., tumor at 140x, normal at 90x). This first dataset showed the vast heterogeneity among different methods, with a large number of tool-specific calls (Figure 61).

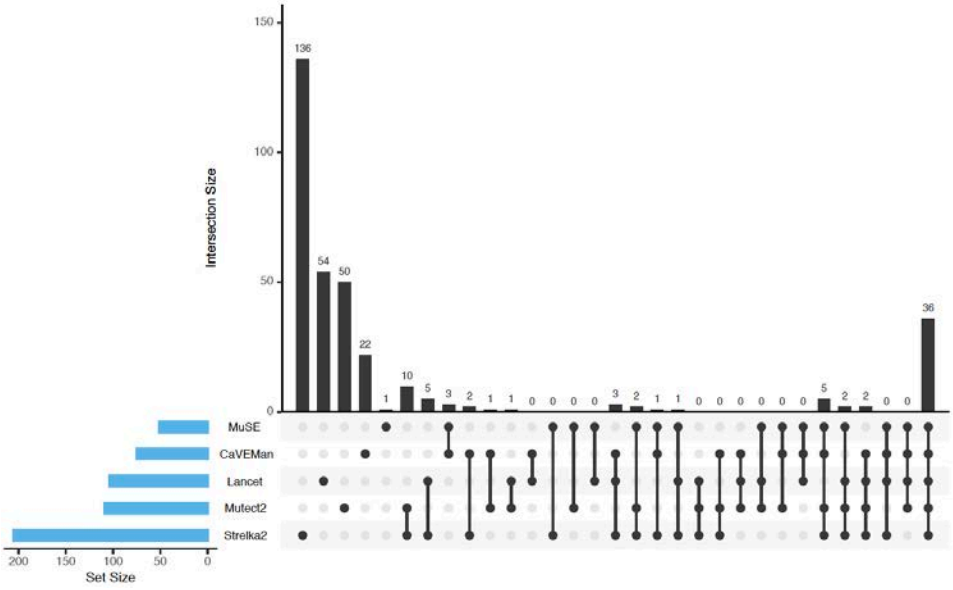


Figure 61. Comparison of SNVs results for the benchmarking dataset AML31. The upset plot shows the concordance among different variant callers. Each row represents one program. The total number of variants detected by each tool is indicated by the blue barplots on the left. The number of variants in each intersection subset is represented by black vertical bars and the total number on top.

All evaluated programs agreed on 36 SNVs, but they had tens of uniquely called mutations, except MuSE. In this case, Strelka2 was the program detecting the higher number of mutations, which is a general trend that we have seen in most of the analyzed samples (Figure 62).

In line with this, in the evaluation of AML31, Strelka2 had the highest recall but also the worst precision, while MuSE had the best overall performance with a good balance between recall and precision and the best F1-score. We also evaluated the consensus results of variants detected by a minimum number of programs and recognized that the SNVs detected by at least 3 or 4 progs (labelled 3_PROGS and 4_PROGS) yielded the best results (Figure 63 and Table 9).

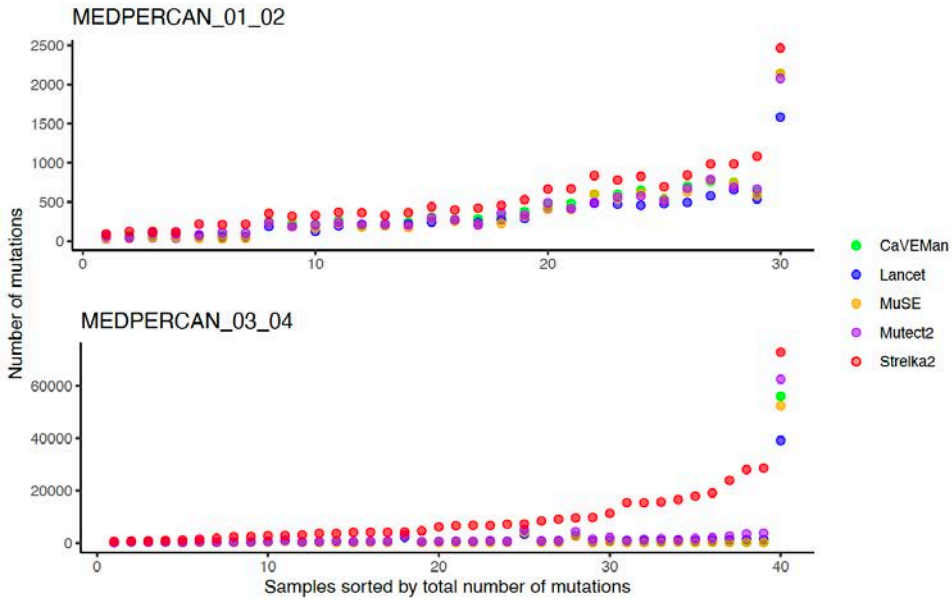


Figure 62. Overview of the number of detected SNVs. The number of SNVs detected by each program is indicated. Each dot represents a sample from the corresponding MedPerCan dataset MEDPERCAN_01_02 (top) and MEDPERCAN_03_04 (bottom).

Table 9. Benchmarking results of SNVs in AML31 WES. FP: false positives, TP: true positives, FN: false negatives, F1_score: weighted average of Precision and Recall.

VC	*Recall (%)*	*Precision (%)*	*FP*	*TP*	*FN*	*F1_score (%)*
1_PROG	93.9	13.6	291	46	3	23.8
2_PROGS	93.9	62.2	28	46	3	74.8
3_PROGS	91.8	86.5	7	45	4	89.1
4_PROGS	87.8	95.6	2	43	6	91.5
5_PROGS	69.4	94.4	2	34	15	80
Mutect2	89.8	40.4	65	44	5	55.7
Lancet	79.6	37.5	65	39	10	51
CaVEMan	85.7	56	33	42	7	67.7
MuSE	87.8	84.3	8	43	6	86
Strelka2	93.9	22.4	159	46	3	36.2

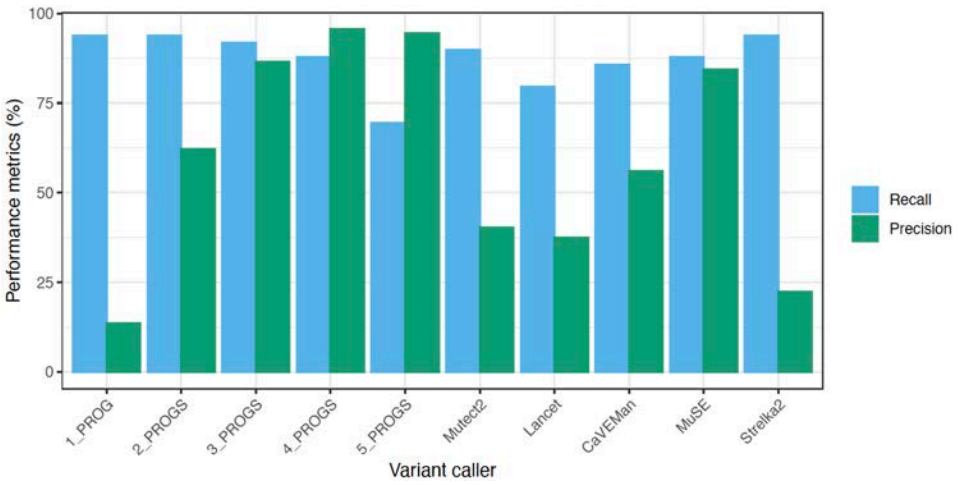


Figure 63. Benchmarking results of SNVs in AML31 WES. Blue bars represent the recall for the variants detected by each tool and those detected by a minimum number of programs, while green bars show their precision.

Due to the lack of other comprehensively characterized WES datasets for somatic variant benchmarking, we also used an approach based on orthogonal validation in which we compared the variants identified in high-coverage gene panels versus those detected by WES. Although this limited the parts of the genome and the number of variants that could be evaluated to the regions and mutations that were detected by the gene panels, it had the advantage that we could benchmark not only real data, but also part of the data that was being generated within the project. Thus, we could assess the performance of variant calling on the exact sequencing and procedures where it was meant to be used. For the evaluation metrics, variants detected in the gene panels were considered true positives, while those not seen in these high-coverage regions were considered false positives.

We identified 13 diffuse large B-cell lymphoma (DLBCL) samples that had WES from the MedPerCan project and that were also subjected to the sequencing of 106 genes in a previous study (Karube et al., 2018). The published results only

included coding non-synonymous variants, which cuts down the number of variants that can be used in the evaluation. Hence, we reanalyzed the gene panels with a validated variant calling pipeline (Nadeu et al., 2016; Rivas-Delgado et al., 2021). The comparison of the results with the WES analyses showed a significant level of precision but missed a high proportion of variants (Figure 64).

Indeed, gene panels have higher coverage than WES and can detect lower frequency variants, which are below the threshold of detection at lower coverages. Hence, we filtered the low frequency variants to have a more realistic view of the variants that can be recognized in WES.

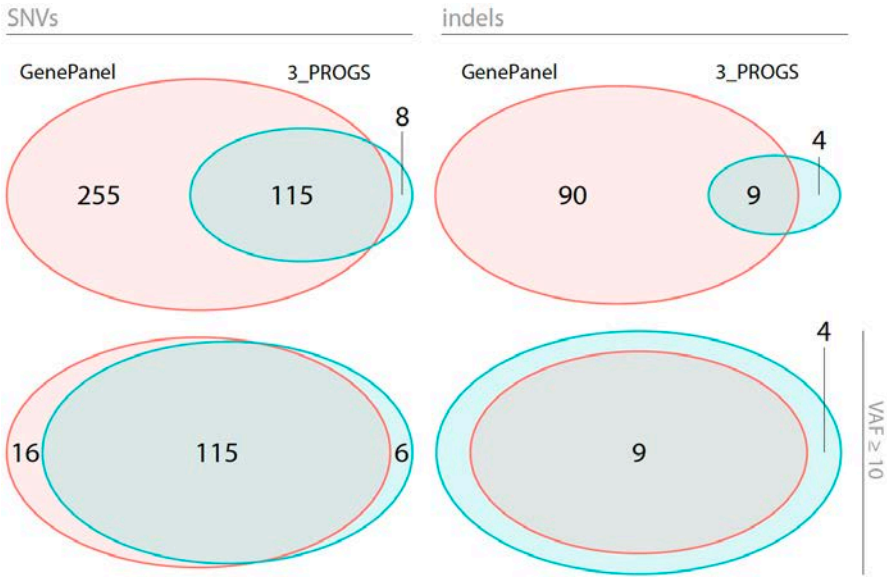


Figure 64. Venn diagrams showing the comparison of variant calling in gene panels and WES. The 3_PROGS strategy (variants detected by at least 3 programs) is shown for illustrative purposes for SNVs (left) and indels (right). The pinkish circles represent the number of variants obtained from the gene panel, while the blueish ones correspond to the WES. The results shown as is (top) and after filtering low frequency variants (bottom).

We calculated the recall, precision, and F1 score of SNVs and indels for each program as well as the consensus for each number of programs (labeled as N_PROGS: variants detected by at least N program) (Table 10 and Table 11).

Table 10. Benchmarking results of SNVs identified in WES versus SNVs from gene panels. FP: false positives, TP: true positives, FN: false negatives, F1_score: weighted average of Precision and Recall.

VC	Recall	Precision	FP	TP	FN	F1_score
1_PROG	88.6	95.1	6	116	15	91.7
2_PROGS	88.6	95.1	6	116	15	91.7
3_PROGS	87.8	95	6	115	16	91.3
4_PROGS	79.4	94.5	6	104	27	86.3
5_PROGS	52.7	94.5	4	69	62	67.6
Mutect2	58	95	4	76	55	72
Lancet	85.5	94.9	6	112	19	90
CaVEMan	84.7	94.9	6	111	20	89.5
MuSE	80.9	94.6	6	106	25	87.2
Strelka2	87.8	95	6	115	16	91.3

Table 11. Benchmarking results of indels identified in WES versus indels from gene panels. FP: false positives, TP: true positives, FN: false negatives, F1_score: weighted average of Precision and Recall.

VC	Recall	Precision	FP	TP	FN	F1_score
1_PROG	100	60	6	9	0	75
2_PROGS	100	69.2	4	9	0	81.8
3_PROGS	100	69.2	4	9	0	81.8
4_PROGS	88.9	66.7	4	8	1	76.2
5_PROGS	77.8	63.6	4	7	2	70
6_PROGS	55.6	100	0	5	4	71.4
Mutect2	77.8	58.3	5	7	2	66.7
Lancet	88.9	66.7	4	8	1	76.2
Pindel	88.9	66.7	4	8	1	76.2
Platypus	66.7	100	0	6	3	80
SvABA	100	69.2	4	9	0	81.8
Strelka2	100	64.3	5	9	0	78.3

Like in the AML31 benchmark, 3_PROGS was one of the best approaches for both SNVs and indels. However, the 1_PROG and 2_PROGS outperformed 3_PROGS for SNVs. This can be explained for the restrictions on the evaluated variants. Due to the large number of low frequency variants that could be detected in the gene panels and not in WES, mutations with VAF lower than 10% were filtered out. This can inflate the precision of variant callers, as it eliminates variants that are more difficult to detect, and more prone to be artifacts. An example of this can be seen in Study 3 (Figure 72 confirms the low frequency of false positives from a WGS benchmark).

A similar reasoning can be applied to Strelka2, which was the best individual tool and achieved the same performance as 3_PROGS. However, Strelka2's good results are most likely due to the VAF filtering used in this evaluation. Actually, this program tends to attempt to call difficult variants close to the limits of detection, which translates into a high recall and a low precision, as can be seen in the AML31 benchmark and in the WGS benchmarks in Study 3.

Next, we used another orthogonal validation dataset based on CLL, composed of 64 WES samples and a validation gene panel of 28 CLL driver genes (Nadeu et al., 2016, 2018). In this case, we used the variants reported by the publications directly, which include non-synonymous variants. Although this WES data was not part of the MedPerCan project and was sequenced differently, we thought that this dataset would be of value as it allowed us to evaluate the performance of variant callers on CLL samples, which is the focus of the research studies of the thesis (Results - Chapter 3: Study 4 and Study 5). Thus, the results and lessons learned from this analysis could be directly applied to our studies.

During the evaluation of this dataset, we realized that several variants were missed by some of the programs used due to tumor contamination in the normal sample. This is a well-known characteristic of blood cancers, which present unexpected contamination of normal samples with tumor cells. Consequently, most variant callers discard genuine somatic mutations since they are found in the normal sample and therefore interpreted as germline variants. This criterion works well with solid tumors that do not tend to present tumor cells in normal samples, but it must be reconsidered when analyzing hematological malignancies.

In line with this, we reevaluated the filtering decisions of the more stringent variant callers (i.e., Mutect2, Lancet, CaVEMan, and Pindel) and implemented a more flexible selection of somatic variants. We recovered all variants (including SNVs and indels) that were filtered out exclusively due to the presence of mutated reads in the normal sample, that had a variant allele frequency below 0.05 in the normal sample, and whose VAF difference between tumor and normal was greater than 0.2. This process allowed us to rescue 15% (11/75) of the originally missed somatic variants. The results presented from here onwards already include the application of this method.

Focusing on the 3_PROGS strategy, which includes the variants detected by at least 3 programs, and that showed one of the best performances in the DLBCL orthogonal validation and the AML31 benchmark, we obtained a precision of 85% and a sensitivity of 64% for SNVs. Next, as seen before, we recognized that the drop in recall was mainly due to low frequency variants that could be identified in the high coverage gene panels, but not in the lower coverage WES samples. Hence, we evaluated the results discarding low frequency variants (<10%) and attained a sensitivity of 83% and a precision of 83% (Figure 65).

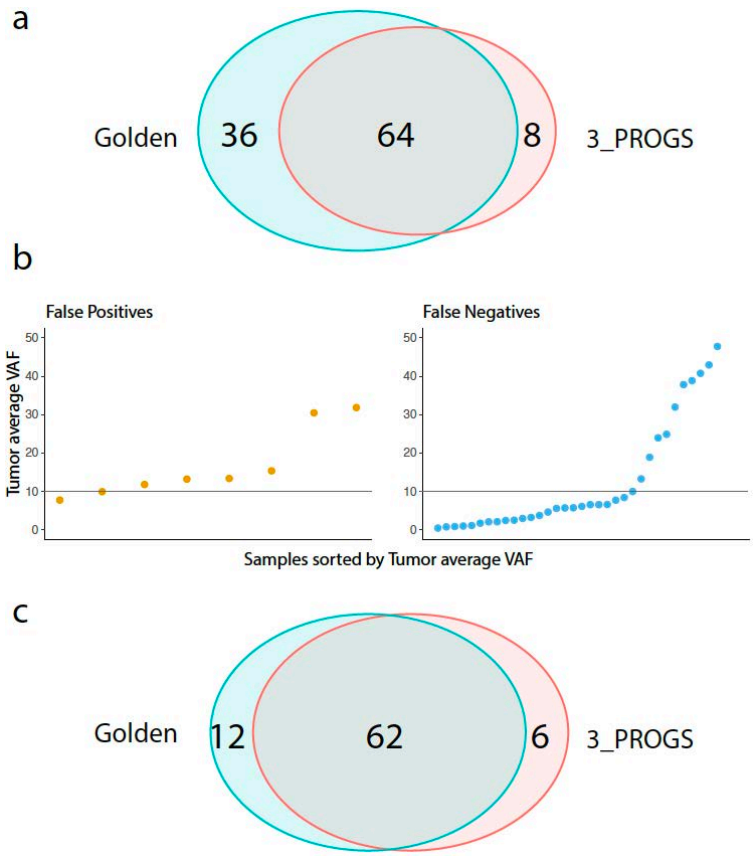


Figure 65. Performance of 3_PROGS for SNVs in CLL WES. a. Venn diagram shows the intersection of SNVs detected by 3_PROGS (pink) and SNVs detected in the high coverage gene panel considered as "truth" and named Golden (blue). b. Variant allele frequency of the false positives and false negatives, which commonly have a VAF below 10%. c. Venn diagram shows the intersection between 3_PROGS and the Golden variants having a VAF above 10%.

We analyzed the indel variants in the same way and obtained a precision of 91% and a sensitivity of 43%. Again, we found that many false negatives, or missed indels, had low frequencies, and we performed the same evaluation considering only indels with a VAF above 10%. This increased our sensitivity to 71% and maintained a specificity of 91%. Indel detection is more complex than that of SNVs, which could explain the lower sensitivity in this variant calling (Figure 66).

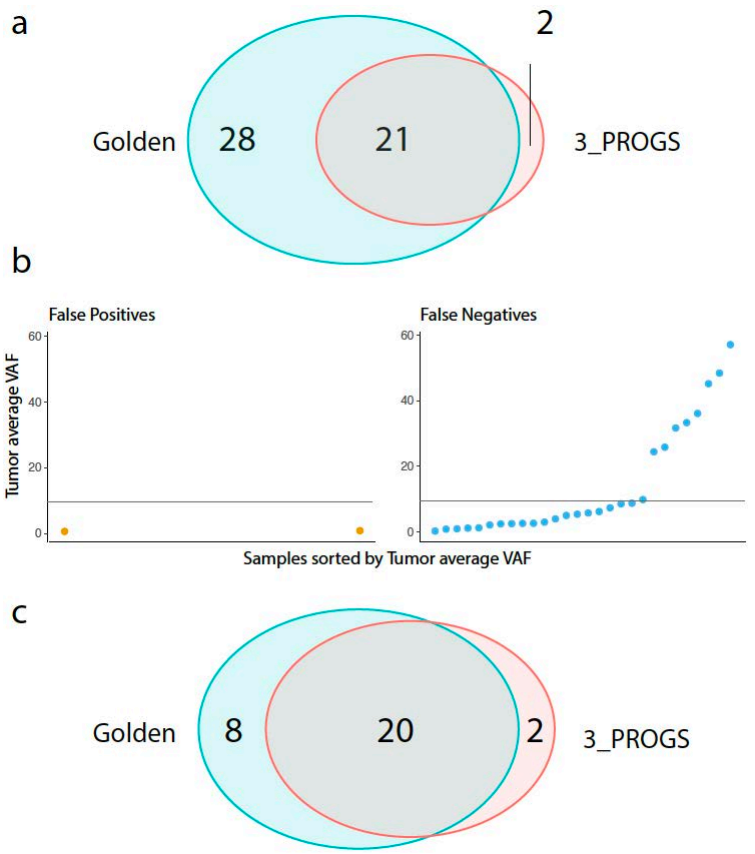


Figure 66. Performance of 3_PROGS indels in CLL WES. a. Venn diagram shows the intersection of indels detected by 3_PROGS (pink) and indels detected in the high coverage gene panel considered as "truth" or Golden (blue). b. Variant allele frequency of the false positives and false negatives, which commonly have a VAF below 10%. c. Venn diagram shows the intersection between 3_PROGS and the Golden variants having a VAF above 10%.

All the results presented up to now refer to variant calling performed on normal and tumor paired samples, which is the ideal setting for somatic variant calling. However, when the germline sample is not available one can still do a tumor-only analysis, although it is less precise and can leak a high number of germline variants.

Within the MedPerCan project, several tumor samples without a matched normal were included. Thereby, we sought to assess the validity of such analyses to understand their reliability and to what extent they can be used. We used the same 13 DLBCL samples used for the previous orthogonal validation and performed two tumor-only approaches. On one hand, we applied the same pipeline as in the normal-tumor paired samples, using a random normal sample, and filtering all variants that were present in a panel of normals created from an in-house cohort of around 600 CLL and mantle cell lymphoma (MCL) cases. On the other hand, we selected a program that was already prepared to run tumor-only analysis (i.e., Mutect2). In both cases, after the variant calling, we also filtered out all variants with a population frequency greater than 1% found in gnomAD, ExAC, or 1000genomes. The results shown here are based on SNVs.

First, we did a simple comparison between the results previously obtained by normal-tumor analysis and the two tumor-only strategies (Figure 67). As expected, both tumor-only results carried way too many mutations (most likely germline variants) that were not called by the pipeline which considered the germline sample. Due to this high number of discrepant variants, the results are not comparable, and the reliability of the tumor-only variants is questionable. Next, we tried to see if we could improve the results by reducing the span of the genomic regions, for instance focusing only on regions of interest, i.e., the gene panel captured regions. Of course, this means that we would be losing most of the variants that could be detected by WES, but at least we would grasp valuable information from the tumor-only samples. We observed that the variability between tumor-only and normal-tumor analysis is now lower, and the proportion of potential germline leakage is reduced (Figure 67).

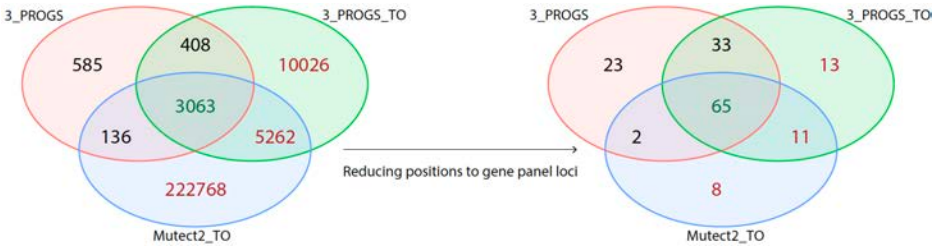


Figure 67. Comparison of normal-tumor paired versus tumor-only analyses. Results from WES normal-tumor analysis considering variants detected by at least 3 programs (3_PROGS) versus tumor-only approaches: 3_PROGS using a random normal (3_PROGS_TO) and Mutect2 in tumor-only mode (Mutect2_TO). Red numbers correspond to variants detected only in tumor-only analyses. Green variants in the middle correspond to matched results between normal-tumor paired and tumor-only analyses.

We continued to apply other filters based on the VAF and the functional impact of the variants. Here, we included the gene panel results in the comparison, which allowed us to assess the sensitivity and precision on the evaluated variants (Figure 68).

As we reduce the number of evaluated variants by their frequency, we increase the agreement with the gene panel's results, as previously explained. Moreover, restricting the variants based on their functional impact (i.e., by selecting those that can have a potential effect at the protein level) further increases the compatibility between normal-tumor paired and tumor-only analyses. Thus, tumor-only analysis might be used, with caution, to identify coding mutations in genes of interest. In our benchmark, we obtained a sensitivity of 55% and 73%, and a specificity of 87% and 80%, by using our two tumor-only analyses based on Mutect2 in tumor-only mode and our normal-tumor pipeline using a non-matched normal, respectively.

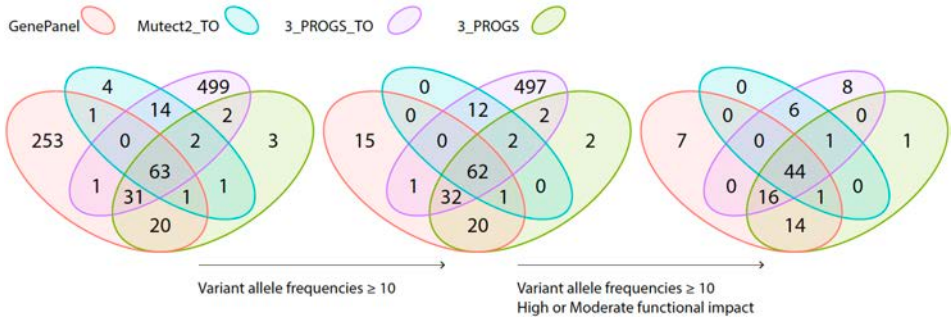


Figure 68. Comparison of gene panel, normal-tumor paired, and tumor-only analyses. Results from gene panel versus WES normal-tumor analysis considering variants detected by at least 3 programs (3_PROGS) versus tumor-only approaches: 3_PROGS using a random normal (3_PROGS_TO) and Mutect2 in tumor-only mode (Mutect2_TO). From left to right, first the results are compared as is, only restricting to the gene panel covered regions. Next, variants of low frequencies (VAF below 10%) are filtered out. Finally, only variants with a High or Moderate functional impact are kept.

As to the implementation, the main pipeline for variant identification and characterization within the MedPerCan project (Figure 60) was set-up in the BSC HPC facilities, mainly the MareNostrum supercomputer. This implementation was adjusted to our HPC regulations, respecting the established maximum wall times, parallelizing the executions to avoid sequential runs, and aiming at the efficient use of the resources.

Although the project only considered WES data, which generates relatively small files that require less resources and are faster to analyze, we anticipated the use of the pipelines for other datasets, such as WGS, and took the potential longer execution times as well as larger computational requirements into consideration. As a summary, for each tool, or task, a bash wrapper was implemented to execute all the command lines that were necessary to obtain the results (i.e., execution of the tool itself, which might include several commands, and various normalization and formatting steps). Besides, another script was also developed to submit each particular task to the Slurm queue system. Both scripts were homogenized in such

a way that all tasks could follow the same structure. The data was also organized in predefined tree directories that could be automatically identifiable from the scripts. In order to launch the selected tools or pipeline, a set of master scripts were programmed to orchestrate the executions on the input samples. As the pipeline accepted multiple inputs, GREASY was used to leverage the use of HPC for massively parallel executions. The dependencies among different tasks were managed using Slurm's dependencies. Finally, the framework supported the re-submission of failed tasks using the re-start files generated by GREASY.

Application of the methodology to representative tumor types

The goal of the MedPerCan project was to set up an operational circuit that could encompass the whole path of personalized medicine in oncology, from sample collection in a hospital to the genomic report intended to go back to the doctors and the patient, hopefully with an increased value for the management of the patient (i.e., better diagnosis or treatment options). After establishing this system, the next aim was to evaluate the clinical relevance of introducing genomic analysis in health routines. In this direction, the clinical partners (IDIBELL, IDIBAPS, and VHIO) selected a set of samples to be analyzed with a specific purpose. They covered different stages during the development of the disease and included the groups of patients who could benefit the most (i.e., cases with risk of hereditary cancer and cases who could receive treatment strategies based on their genomic profile).

IDIBELL studied the risk of hereditary cancer (colorectal/endometrial) in patients without any recognized causal variants in the current gene panel in use. They included 77 non-tumoral samples sequenced at 60x, which were analyzed by the germline variant calling pipeline.

IDIBAPS selected diffuse large B-cell lymphoma (DLBCL) cases that do not respond well to current therapies to determine whether genomic analysis could be of use to identify the patients who do not benefit from this first-line treatment. This study included 94 cases with paired normal and tumor samples and 5 cases without a matched germline sample. All samples were fresh frozen (FF), and they were sequenced at 90x and 140x, for normal and tumor samples, respectively. Normal samples were subjected to germline analysis, paired normal-tumor samples were analyzed by the somatic variant calling pipeline, and tumor-only approaches were performed on the tumor samples without normal.

The last two subprojects included studies to improve treatment selection during advanced stages of the disease. VHIO included a set of 41 relapse cases with colorectal cancer to identify molecular alterations that could be potential therapy targets. They had paired normal and tumor samples, from FFPE, and sequenced at 150x and 70x, respectively. Germline and somatic variant calling were performed. Another group from IDIBELL investigated POLE/POLD1 mutated endometrial tumors to assess if they could benefit from PD1 inhibitors. They included 42 FFPE samples, including multi-region tumors at 60x and controls at 60x from 11 patients. Germline and somatic variant calling were performed on paired normal-tumor samples and tumor-only strategies were applied to unmatched tumor samples. An additional quality assessment was done due to the poor quality of some samples.

Overall, the previously described somatic variant calling pipeline was applied to normal-tumor paired samples, including both FF and FFPE. Germline variant calling was performed on all normal samples. There was also a subset of cases without matched germline material that were analyzed using tumor-only approaches, and PDX samples that were subjected to tumor-only analyses with a prior step to distinguish human DNA from mouse DNA (see Methods - section 3.3).

After the genomic analyses conducted at the BSC, the outputs were sent back to the CRG-CNAG that was in charge of setting up a platform for an easy inspection and interpretation of the results. From there, each hospital group received the list of genomic variants identified in their samples, either through the CNAG platform, or through raw files with lists of variants produced at the BSC, to perform the corresponding downstream analysis. Some of this work has led to possible publications and manuscripts are under review or being finalized.

4.2.3 Study 3: Comprehensive characterization of tumors based on its genomic profile

Complementing the previous project, we started an independent collaboration with the group of Dr. Elías Campo for the characterization of CLL evolution with special focus on Richter transformation. For this project, we relied on the MedPerCan variant calling pipeline, and added further functionalities and filtering strategies to adapt to the needs of that particular project (Results - Chapter 3: Study 4 and Study 5), as well as other side projects (see Appendix). The results concerning the methodological aspects of this new phase will be explained within this chapter, while the application of the methods to our biomedical studies will be presented in the next chapter (Results - Chapter 3).

Following the needs of these upcoming research collaborations, we expanded the MedPerCan framework to cover new data types as well as complementary downstream analyses. Among the additions and improvements, we set-up tools for RNA-seq analysis, extended the pipeline and benchmarks to deal with WGS, and enhanced it with downstream analyses including tumor evolution and mutational signatures. We set-up and evaluated the required methods and implemented their automatic executions. I worked together with Ana Dueso from the Computational Genomics group at BSC, who contributed to the setting up of

SVs merging and consensus strategies, and Ferran Nadeu from IDIBAPS, who provided input for variant calling and downstream analyses in connection to CLL studies.

Here, I present the extension of the MedPerCan variant calling pipeline (Results - Chapter 3: Study 2) to support additional data types and analyses, within the context of our collaboration with Dr. Elias Campo's. The RNA-seq analyses as well as WGS downstream analyses, including mutational signatures and tumor evolution, are described in the Methods (see sections 3.3.8, 3.3.9, and 3.3.11). The variant calling pipeline for WGS, including its evaluation, is presented here, while the specific technical aspects can be found in the Methods (see section 3.3.2). These new strategies have been applied to several works (see Appendix), including the two studies that are part of this thesis, described in the next Results' chapter (Results - Chapter 3: Study 4 and Study 5).

4.2.3.1 Introduction

Genomic initiatives, like the MedPerCan project, though they are meant to move towards their translation into the clinics, they remain at the level of research, and their applicability in real clinical practice has yet to come. Within the MedPerCan project, we developed a framework for tumor genome analysis of WES (described in Study 2), which we continued to use and improve beyond the completion of the project. New collaborations called for additional analyses to include other molecular data, such as WGS or RNA-seq, and further characterization of somatic mutations. In this direction, we introduced new features to gain insights into tumor composition and evolution, genomic complexity, and the identification of mutational processes that contribute to the mutational spectra of tumors.

4.2.3.2 Results

Pipeline for variant identification and characterization in WGS

WGS allows a wider characterization of tumors, including not only the coding but also the non-coding region of the genome. The higher number of mutations that can be identified genome wide is advantageous for further analysis such as mutational signatures or tumor evolution. Moreover, copy number alterations and structural variants can be recognized throughout the whole genome. Small variants, including SNVs and indels, can usually be identified using the same programs as for WES, while CNAs and SVs might require new tools specific for WGS (Table 4). In line with this, we based the extended WGS framework on the initial MedPerCan pipeline, added new programs to cover CNAs and SVs, and complemented it with the full characterization of the mutational landscape of tumors. We integrated methods to explore it in terms of genomic complexity, mutational processes, and tumor subclonal structure to uncover the dynamic forces driving its evolution (Figure 69).

The design and implementation of WGS workflows, including the selection and evaluation of tools, was guided by the needs of our studies of CLL and emerging questions we wanted to elucidate (see Results - Chapter 3: Study 4 and Study 5). Variant calling was fine-tuned with additional benchmarks on the same datasets we were analyzing, and a thorough and manual investigation of the results was always applied. Overall, the complemented framework for WGS tumor analysis was enhanced to work with our CLL studies, but it could be applied to any other cancer dataset. Likewise, the protocols we followed for the examination, evaluation, and interpretation of the results could be adopted by any cancer genomics study.

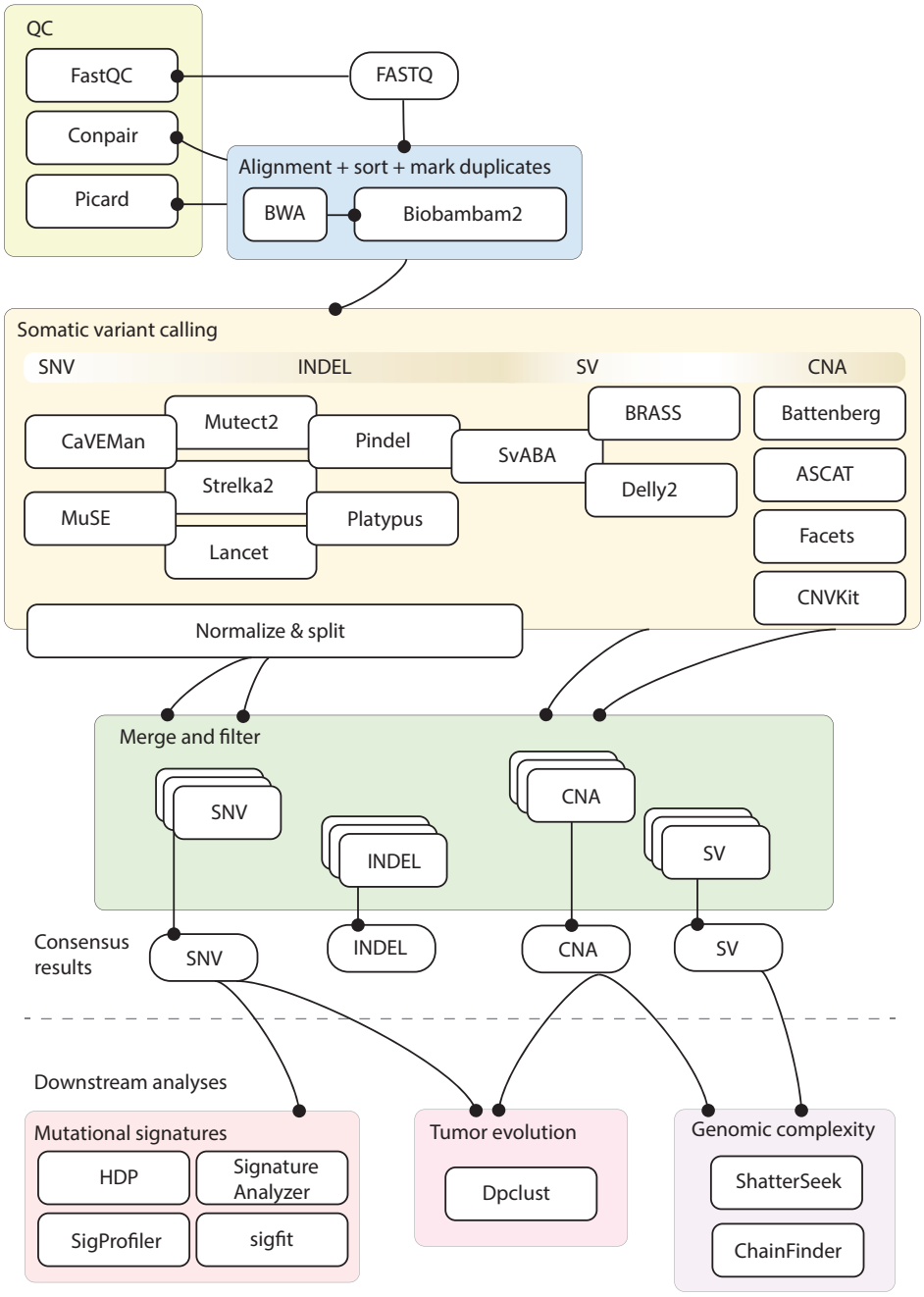


Figure 69. Framework for tumor WGS analysis. Overview of the steps and programs used to analyze WGS. Quality control and alignment precedes the central step to discover tumor variants

implemented by a multi-caller approach. Finally, downstream analyses expand the characterization of somatic variation on top of variant calling.

Variant calling is at the core of genomic analyses and a good understanding of its performance and reliability on the analyzed datasets is essential. To assess the strengths and weaknesses of variant calling strategies for WGS, we first benchmarked different tools and candidate consensus strategies using real WGS data. We worked with the well-characterized medulloblastoma sample (MB99) from published benchmarking efforts (Alioto et al., 2015), and used a downsampled version at 30x to match the coverage of our data. The Tier1 and Tier4 lists of golden variants were used to calculate the sensitivity and precision, respectively, for SNVs and indels (see Methods - section 3.2.1).

The intent of the results discussed here is to present the guidelines and the reasoning that we followed to evaluate variant calling strategies, rather than the selection of one single strategy that might be very convenient for one specific benchmarking dataset but might not be so good for other external data with slightly different characteristics. For illustrative purposes, SNVs will be used to explain and exemplify our approach in more detail.

As expected, and as we have seen before, there was a high disagreement among programs in WGS variant calling. Starting with SNVs, some variants were indeed detected by all algorithms, but the majority of them (74%) were private mutations detected by only one tool (Figure 70). The evaluation of these SNV results showed that all programs had an overall good recall, but their precision was lower and more variable. The criteria of selecting the variants detected by a minimum number of programs improved the precision, while preserving a reliable recall (Figure 71).

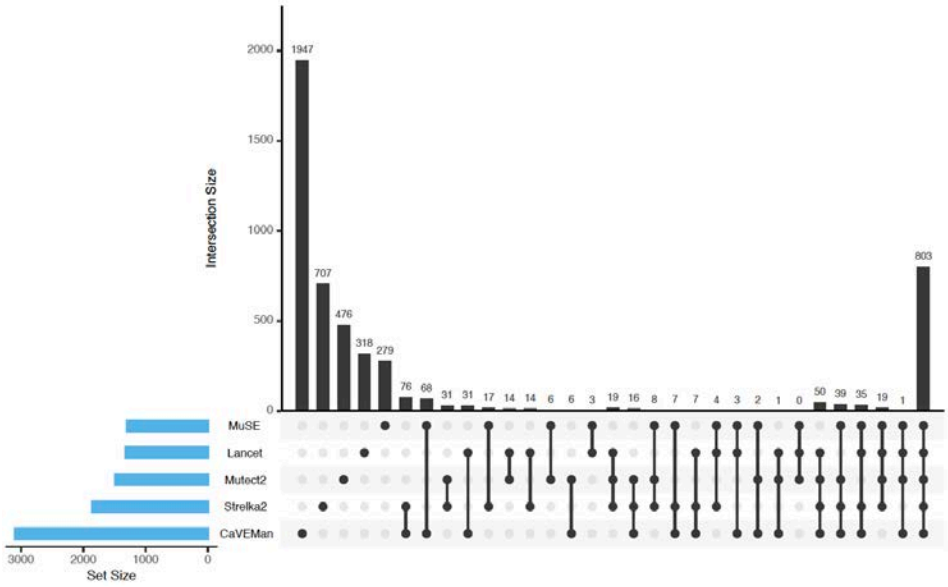


Figure 70. Comparison of SNV results for the benchmarking dataset MB99. The upset plot shows the concordance among different variant callers for SNVs. Each row represents one program. The total number of variants detected by each tool is indicated by the blue barplots on the left. The number of variants in each intersection subset is represented by black vertical bars and the total number on top.

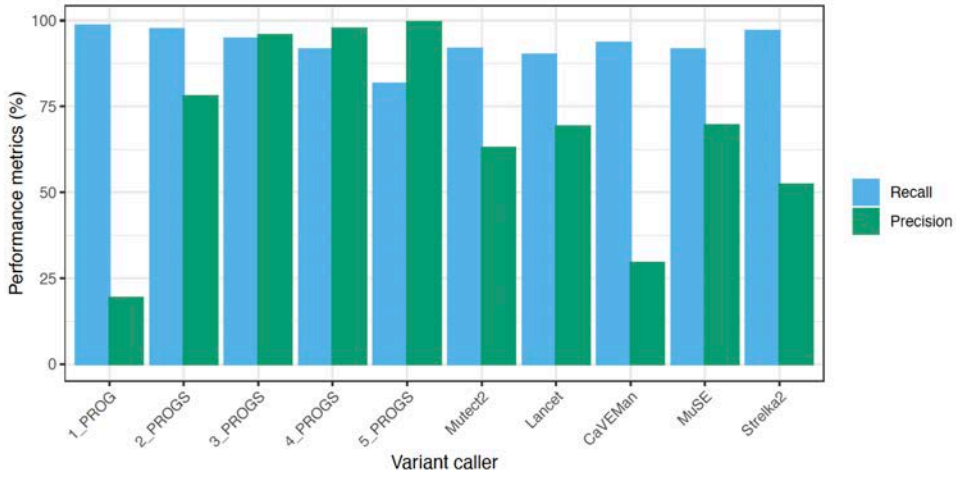


Figure 71. Benchmarking results of MB99 SNVs. Blue bars represent the recall for the variants detected by each tool and those detected by a minimum number of programs (N_PROGS), while green bars show their precision.

Clearly, the union of all program's results (labeled 1_PROG) had the best recall, as it included the highest number of variants. However, it also had the worst precision of all. On the other way around, the consensus of variants detected by all 5 programs (5_PROGS) had the best precision and the worst sensitivity. Investigation of the false positives showed that most of them came from tool-private calls and were related with low frequency values (Figure 72).

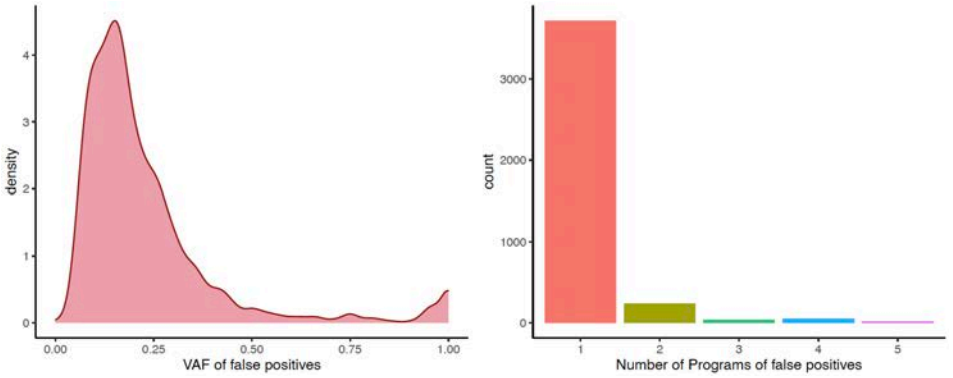


Figure 72. Characteristics of false positives. Distribution of the VAF of false positives (left). Absolute number of false positives detected by the intersection of each number of programs (right).

In order to select the best strategy for our research, we first tried to see if we could improve the results of each program individually. Digging into each program's results, we saw that specific features calculated by the tools could have a direct impact on their performance. Most variant callers provide multiple metrics for each reported variant, commonly including the number of reads supporting the mutation, the total depth, quality-related values such as mapping or base qualities, and other statistics indicating the reliability of the variant. These factors are taken into consideration to try to discern true somatic variants from germline variants, sequencing errors, and other potential artifacts. We investigated the behavior of these features on true and false positives to see if stricter filtering could improve the results. We found several candidates that had

distinct values for true and false positives, and that could possibly be used to fine-tune the programs' outputs. Some examples are detailed in Figure 73, where three metrics from three variant callers show higher values for true positives, while most false positives have the lowest values. We explored the metrics that had this kind of behavior and set different thresholds to try to filter out false positives, while keeping the true variants.

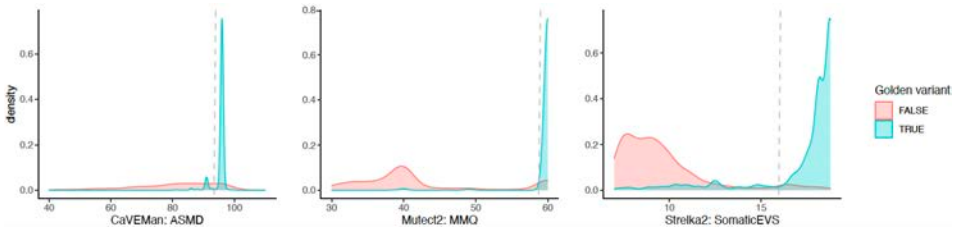


Figure 73. Distribution of program-specific features on true and false positives. Example of three metrics from three variant callers (CaVEMan, Mutect2, and Strelka2). The definition of the metrics shown is detailed in the Methods, section 3.3.2.2. The density plots show the distribution of the true positives (variants in the Golden dataset, in blue) and false positives (variants not in the Golden dataset, in pink). Dashed vertical lines indicate an illustrative threshold that would minimize false positives, while keeping the majority of true positives.

Moreover, we also examined the general profile of the variants that would be discarded when applying those filters. We explored them in terms of the pattern of the 96 mutation types (see Introduction - section 1.3.4), widely used for mutational signatures analyses, and the variant allele frequency of the mutations. Focusing on CaVEMan's results (Figure 74), we can see that the expected 96-classes profile (from Golden SNVs) is slightly different from this tool's output, as it shows 4 striking peaks that are not present in the validated variants. Now, if we compare the variants that passed the additional filters against those that were filtered out, we can see that the profile of the conserved variants is more similar to the golden profile, while the filtered-out mutations mainly correspond to the discrepant peaks. Besides, the frequency of the kept variants

covers from clonal to lower frequency variants, but virtually all filtered-out variants have very low frequencies, indicating that they are supported by a low number of mutated reads and that they are more likely to be artifacts.

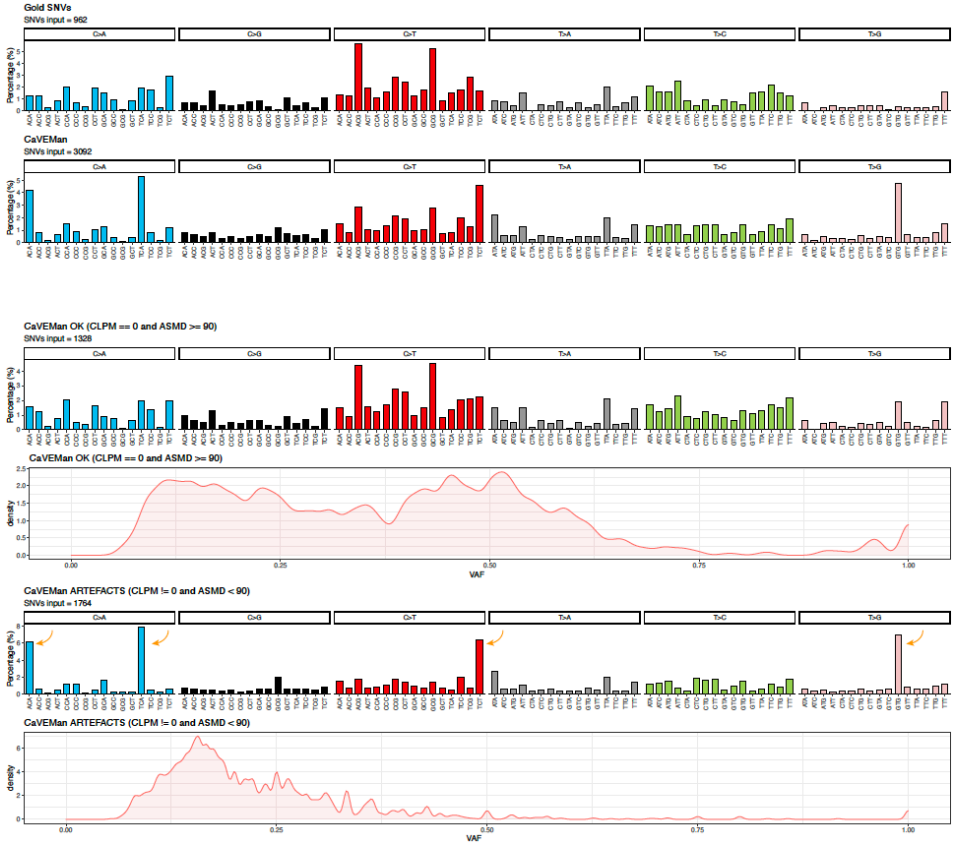


Figure 74. Characterization of the variants detected by CaVEMan in MB99. The two upper rows illustrate the mutational profile of the golden SNVs and that obtained from the CaVEMan results, which shows some characteristic peaks that are not present in the truth set. Next, the "CaVEMan OK" plots show the profile and VAF of the variants that would be kept after filtering, respectively, while the last two plots display the profile and VAF of the variants that are filtered out.

We followed the same procedure for the rest of the features and programs and obtained similar results. We also inspected real sample data from our in-house cohorts to evaluate the impact of this filtering strategy not only on a single

benchmarking dataset, but also on multiple samples from our studies. This analysis allowed us to confirm that variant filtering based on the selected program's metrics would have equivalent results on our own data.

Next, we checked the agreement among programs after applying our filtering strategy per program (see Methods - section 3.3.2.2) and observed that we reduced a large number of unique variants (75% of them were removed).

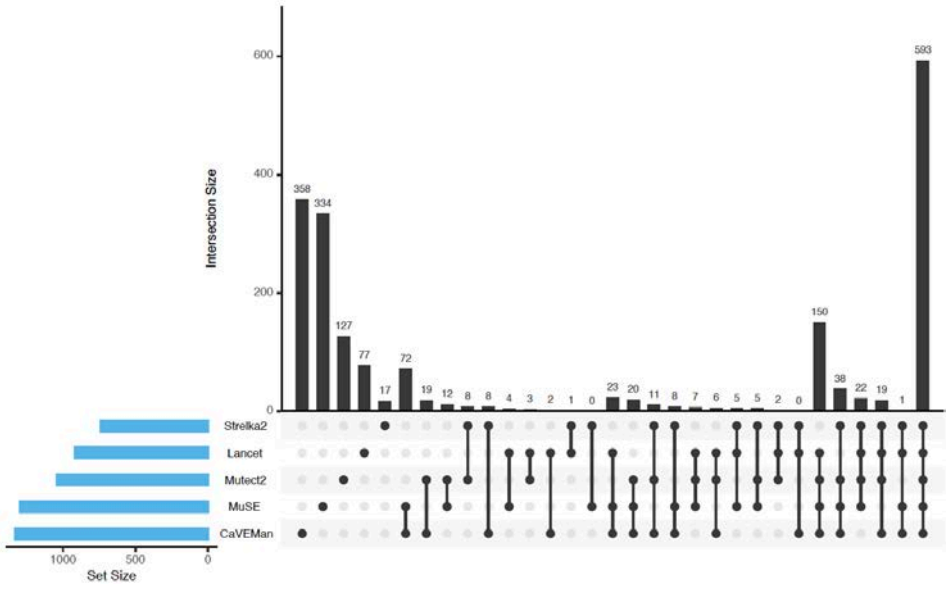


Figure 75. Comparison of SNV results for the benchmarking dataset MB99 after filtering. The upset plot shows the concordance among different variant callers after filtering each program's results. Each row represents one program. The total number of variants detected by each tool is indicated by the blue barplots on the left. The number of variants in each intersection subset is represented by black vertical bars and the total number on top.

Altogether, we concordantly observed that our strategy was removing unique calls, mainly at low frequencies, and often corresponding to striking peaks in the 96-mutation classes that were not expected. Consistently, the evaluation of the results using the golden variants showed a significant improvement on the

precision of each program, while their recall was only affected by a small decrease (Figure 76).

Based on this and the overall exploration of our own data, we evaluated the results of different combinations of programs. The selected pipeline (named Filtered2) included 3 programs (CaVEMan, Mutect2, and MuSE), applied the previously explained filters to the first two, and kept the SNVs detected by at least two algorithms. We did not base our choice solely on the limited benchmarking datasets, which could lead to overfitting of the proposed solution, providing very good results on this data at the expense of losing efficiency on other diverse datasets. We rather examined a variety of our own samples, together with our benchmarks, to see the way in which the programs behaved. In the MB99 dataset, our Filtered2 approach had a recall of 93.6% and a precision of 93.2%, which outperformed all single programs as well as most of their combinations (i.e., considering variants detected by a minimum number of programs) (Figure 76). In the case of the MB99 benchmarking dataset, the intersection of 3 programs out of 5 without any additional filtering achieved a slightly higher recall and precision. However, we found that Filtering2 was more conservative among our samples and less computationally demanding.

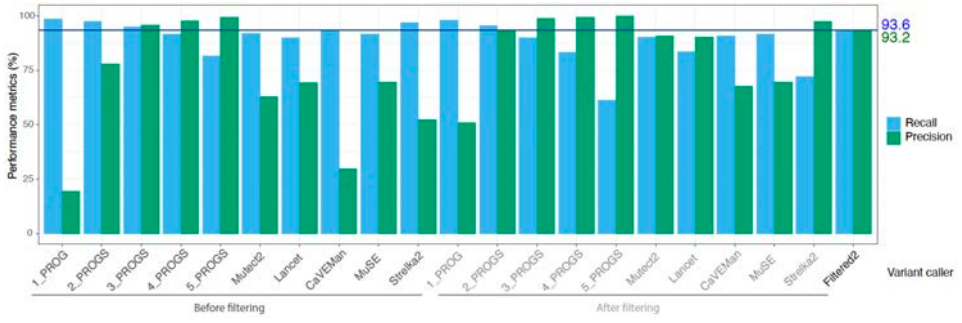


Figure 76. Benchmarking results of MB99 SNVs before and after filtering, and comparison to the Filtered2 strategy. Performance (recall in blue, precision in green) of variant callers as well as their combinations considering variants detected by a minimum number of programs before filtering (left)

and after filtering (right). The selected strategy Filtered2 is shown on the right and its precision and recall are indicated on the right. The horizontal lines correspond to these values for easy comparison with the rest of the results.

Following the SNV analysis, we examined the performance of indel detection using the same benchmarking for WGS (MB99). We applied the same strategy based on 6 variant callers and additional filtering (see Methods - section 3.3.2.2).

Similar to SNVs, we selected 4 programs (Pindel, Platypus, Mutect2, and SvABA), applied the corresponding additional filters, and kept the variants detected by at least 2 programs. Not surprisingly, the performance achieved for indels was much lower than that of SNVs. In the MB99 sample, the Filtered2 strategy had a recall of 67.7% and a precision of 82.3% and was slightly outperformed by the variants detected by 3 or more programs after filtering. The intersection of 3 or more programs without any additional filtering also showed reasonably good results (Figure 77).

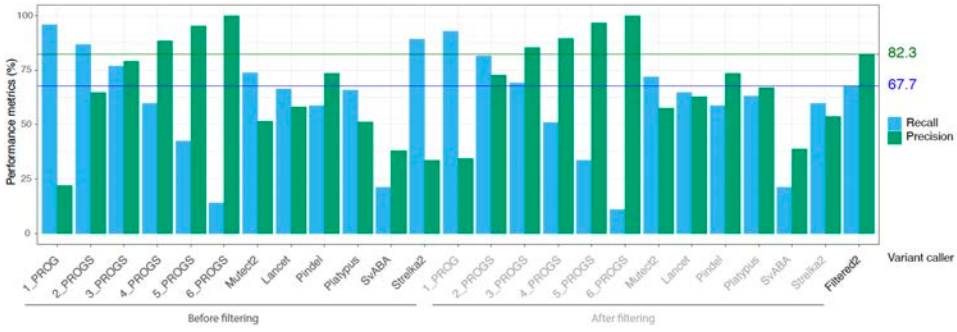


Figure 77. Benchmarking results of MB99 indels before and after filtering and comparison to the Filtered2 strategy. Performance (recall in blue, precision in green) of variant callers as well as their combinations considering variants detected by a minimum number of programs before filtering (left) and after filtering (right). The selected strategy Filtered2 is shown on the right and its precision and recall are indicated on the right. The horizontal lines correspond to these values for easy comparison with the rest of the results.

To assess the performance of the Filtered2 strategy on our cohort, we performed an orthogonal validation of the pipeline against high coverage gene panels (Nadeu et al., 2018) including both SNVs and indels (Figure 78).

We identified 29 CLL samples that were subjected to both whole-genome sequencing and targeted sequencing of 28 CLL driver genes. We considered the variants detected at a minimum VAF of 10% in the gene panels (48 mutations) as the golden or *truth* variants. We obtained a sensitivity of 93% and a precision of 88%. The missed variants corresponded to indels which are well-known to be more difficult to detect.

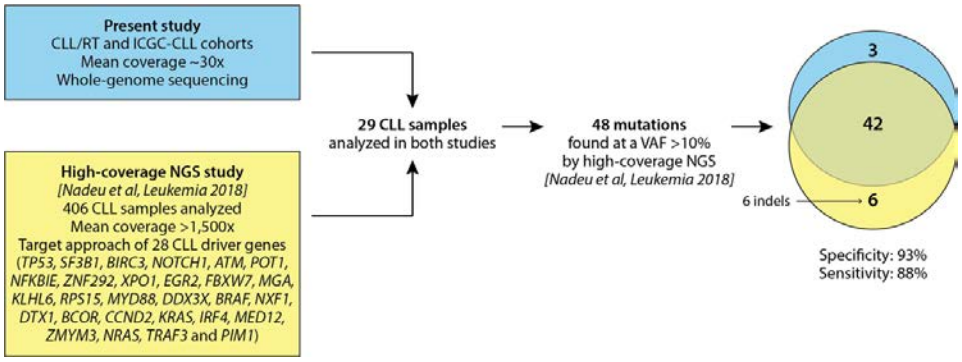


Figure 78. Orthogonal validation of the Filtered2 pipeline. Selection of 29 CLL samples subjected to both WGS and high-coverage gene panel (left). Evaluation of the performance in WGS compared to high-coverage gene panels (right). The Venn diagram shows the intersection between the WGS results (in blue), and the variants identified by the gene panel (yellow).

Finally, to assess the performance of SV variant calling, we used a benchmarking dataset created from a cell line that included a normal and a tumor paired samples (COLO829) subjected to multi-platform sequencing and validation for somatic structural variation detection (Espejo Valle-Inclan et al., 2022). As in the other mutation types, we applied custom filters to the raw results of the variant callers. In this case, the filters were mainly based on the mapping quality

of the reads to try to preserve variants with high confidence (see Methods - section 3.3.2.2).

In this dataset, the filtering step discarded many false positives from Delly2 and Brass, while preserving most of the true positive variants in Delly2 but not in Brass. For SvABA, it only filtered out a few false and true positives (Figure 79).

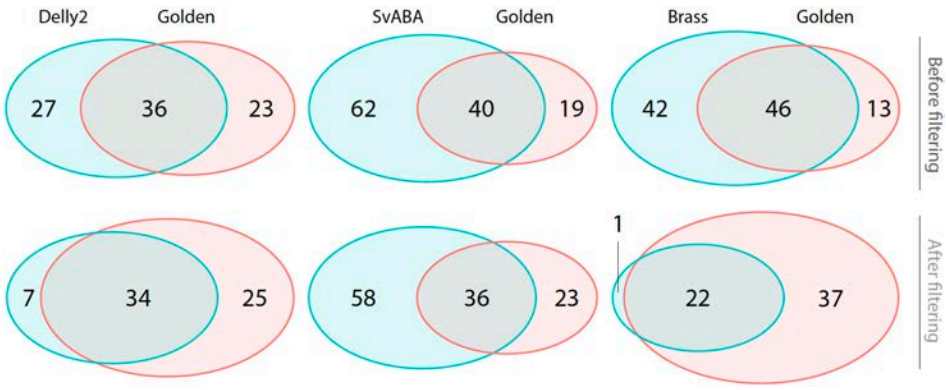


Figure 79. Comparison of variant calling results of SVs in COLO829. The results of each program (Delly2, SvABA, and Brass) is compared to the Golden "truth" variants without any additional filtering (top) and after the selected filters (bottom). The Venn diagrams show the intersection between the program's results (in blue) and the Golden variants (in pink).

We followed a similar merging strategy as for SNVs and indels, but since SVs are more difficult to detect, we applied a more flexible criteria to try to avoid a significant drop in sensitivity. Again, we selected the variants detected by at least two programs but only required one of them to pass the additional filters (see Methods - section 3.3.2.2). We refer to this strategy as Filtered2. In this particular benchmark, this approach achieved the best performance (Figure 80). When analyzing real data, since SVs are the most challenging to detect, we manually curated the results and confirmed them by visual inspection of each SV call in IGV.

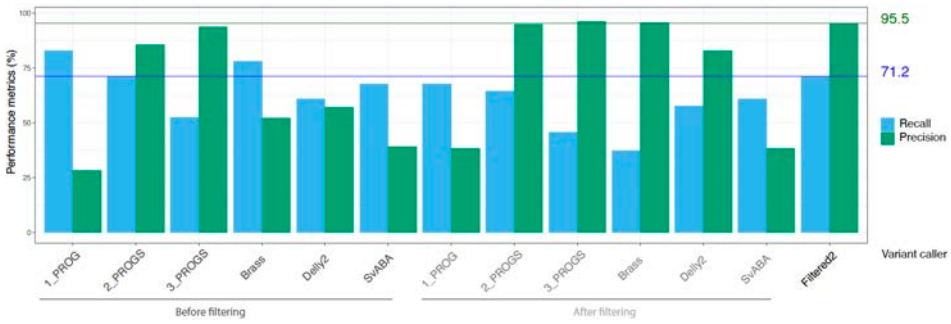


Figure 80. Benchmarking results of SVs in COLO829 before and after filtering. Performance (recall in blue, precision in green) of variant callers as well as their combinations considering variants detected by a minimum number of programs before filtering (left) and after filtering (right). The selected strategy was Filtered2 and its precision and recall are indicated on the right. The horizontal lines correspond to these values for easy comparison with the rest of the results.

4.3 Chapter 3: Application of cancer genome analysis to tackle biological questions

4.3.1 Introduction

This chapter concludes the thesis trajectory as it allowed me to apply the methodological aspects previously described, along with the expertise I gained on cancer genomics, to solve specific biological and clinical questions within oncology. This is where I brought together the computational perspective with the specific needs of biomedical scenarios and focused on the biological meaning of the results, rather than the technicalities of the analyses. In these studies, the procedures to analyze tumor genomes were aimed at finding answers to the biology and evolution of CLL.

The research presented in this chapter (Study 4 and Study 5) was carried out in a close collaboration with Dr. Elías Campo's group at Hospital Clínic de Barcelona/IDIBAPS. For each study, I first explain my contribution to the work

presented, followed by an introduction of the motivation, current knowledge in the field, and our approach. Lastly, I present the results of our research.

4.3.2 Study 4: Richter transformation study

This study has been the core of my thesis. It allowed me to go beyond the technical and methodological aspects of cancer genome analysis and deepen into the biological interpretation of their results to tackle unmet biomedical needs of Richter transformation (RT) in chronic lymphocytic leukemia (CLL). This work has been done in the collaboration with Dr. Elías Campo and Dr. Ferran Nadeu from IDIBAPS/Hospital Clínic de Barcelona. I have actively participated in the project since the very beginning, working hand in hand with Dr. Ferran Nadeu.

My role in this work started on the computational and bioinformatics analyses of NGS data. More in detail, I have collected and managed all data, conducted all WGS, high-coverage UMI-based NGS, bulk RNA-seq computational analyses, and written the corresponding methods. Pursuing my interest on the biological significance of such analyses, I also got into the interpretation of their results together with Dr. Ferran Nadeu, who gave me constant feedback on the more biological and clinical interpretation of the downstream analyses.

Overall, I have pushed and thoroughly followed the project, and I contributed to the writing and conception of the publication, supplemental material, and response to the reviewers.

The following subsection is an introduction of the overarching topic and aim of the project. Next, the results of the study are presented as they appear in the publication (see Appendix): Integrative genomics characterization of RT, Novel mutational processes active in RT, Dormant seeds of RT at CLL diagnosis, The $OXPHOS^{high}$-BCR^{low} transcriptional axis of RT, and OXPHOS and BCR activity in RT. I

contributed to the first 4 blocks, particularly: I carried out all the WGS analyses for the genomic characterization of RT, took responsibility of the mutational signatures analyses, conducted the study of tumor evolution based on WGS, and contributed to the bulk RNA-seq analyses of the fourth part. My work will be explained more extensively, while the rest will be introduced as it appears in the manuscript for the sake of contextualization and understanding of the whole research. The publication can be found in the Appendix.

4.3.2.1 Introduction

Clonal evolution (Cairns, 1975) plays a pivotal role in tumor initiation, progression, therapy resistance, and relapse in leukemias (Ferrando & López-Otín, 2017) and solid tumors (Greaves & Maley, 2012) as a result of the emergence and/or selection of fitter subclones (Dentro et al., 2021; Nowell, 1976). A better understanding of the underlying forces driving these dynamics might help us predict treatment response, prevent poor outcomes, and achieve an overall better management of patients with anticipation-based treatment strategies. Many tumor evolution studies are restricted to low resolution bulk sequencing and single or scant time points, limiting the analyses to an underestimation of the actual tumor cell subpopulations (Gerstung et al., 2020).

As a model of tumor evolution, we have used chronic lymphocytic leukemia (CLL), a usually indolent neoplasia of mature B-cells, though it can transform into a deadly cancer, usually in the form of diffuse large B-cell lymphoma (DLBCL). Richter transformation (RT) (Richter, 1928) in CLL represents a paradigmatic model of cancer evolution, where a slow growing malignancy transforms into a high-grade lymphoma associated with dismal clinical outcomes and unmet clinical needs. It is found rarely in treatment-naïve patients, but affects up to 20% of cases after chemoimmunotherapy (CIT) and newer targeted therapies (W. Ding, 2018).

RT can occur within a few months after treatment initiation (Ahn et al., 2017; Jain et al., 2015; Maddocks et al., 2015), suggesting the selection of minor subclones already present before therapy initiation (Landau et al., 2017).

The genetic makeup of RT has been mainly characterized on patients after CIT using whole exome sequencing and targeted approaches (Chigrinova et al., 2013; Fabbri et al., 2013; Rossi et al., 2011) or FFPE whole genome sequencing (Klintman et al., 2021). These studies have identified a mostly linear model of evolution from CLL to RT, where the predominant CLL clone acquired approximately 20 coding mutations per tumor. The most recurrently altered genes at the time of transformation were *TP53* disruption, *MYC* amplifications, and *CDKN2A* deletions, which were exclusive of RT. Genomic complexity was assessed according to the number of copy number alterations and was found to be intermediate between CLL and DLBCL. Risk factors were also evaluated and include CLL carrying immunoglobulin genes that belong to stereotype subset #8 (see Introduction - section 1.5.3), which is associated with aggressive disease, *TP53* alterations, and *NOTCH1* mutations. In the context of novel inhibitors, RT is less studied and often lacks *BTK* or *PLCG2* and *BCL2* mutations known to confer resistance under ibrutinib and venetoclax, respectively (Innocenti et al., 2018).

Overall, recurrent alterations and risk factors of RT have been identified, but the genomic and epigenomic mechanisms leading to RT after CIT (Beà et al., 2002; Chakraborty et al., 2021; Chigrinova et al., 2013; Fabbri et al., 2013; Klintman et al., 2021; Rossi et al., 2011; Scandurra et al., 2010) and, specially, under targeted agents (Anderson et al., 2017; Herling et al., 2018; Kadri et al., 2017; Miller et al., 2017) remain elusive. A more comprehensive characterization of RT might identify new alterations and molecular mechanisms underlying this transformation and might help us reconstruct the evolutionary trajectories of RT, recognize actionable

pathways, and determine features for early diagnosis that might allow anticipation-based therapies.

We have performed a thorough characterization of RT, including multiple layers (genome, epigenome, and transcriptome) at different resolution levels (bulk and single cell), to reconstruct its evolutionary history, to uncover the molecular processes driving this transformation, and to identify potential factors for early detection that might anticipate its manifestation.

4.3.2.2 Results

Integrative genomic characterization of RT

To achieve these goals, we have integrated the characterization of bulk whole genome, transcriptome, and epigenome, complemented with single-cell DNA and RNA sequencing analyses and functional experiments, of 19 CLL patients developing RT before or after several treatment lines, including targeted therapies or chemoimmunotherapy (see Methods - section 3.2.2; Figure 42, Figure 43, and Figure 44).

We have performed whole genome sequencing (WGS) of 53 samples (mean coverage 33x) and 1 whole exome (mean coverage 119x) of spatial or longitudinal samples of 19 patients (including up to 6 time points per patient). In the majority of cases, their CLL transformed into a diffuse large B-cell lymphoma (RT-DLBCL, n=17), and in two cases it transformed into plasmablastic lymphoma (RT-PBL, n=1), or prolymphocytic leukemia (RT-PLL, n=1), respectively. The RT occurred simultaneously at the time of CLL diagnosis (n=3) or after up to 19 years following different lines of treatment with CIT (n=6) and targeted agents (n=10; BCR inhibitors: ibrutinib n=6, duvelisib n=2, idelalisib n=1; BCL2 inhibitor: venetoclax

n=1). All RT were clonally related to the CLL (see Introduction - section 1.5.7), as assessed by their immunoglobulin genes, 15 cases had unmutated IGHV (U-CLL) and 4 mutated IGHV (M-CLL). Matched normal samples were available in 12 cases allowing a complete analysis of somatic variation, while a restricted bioinformatic analysis was performed in cases lacking germline samples (see Methods - section 3.3.2.3).

A concordant increase in complexity from CLL diagnosis to relapse and RT was observed in bulk WGS, including somatic mutations (SNVs and indels), CNAs, and SVs, and also in the epigenome, analyzing DNA methylation, H3K27ac histone modification, and chromatin accessibility (ATAC-seq) (Figure 81). The WGS analysis showed a mutational burden of 1.8 mutations per megabase and a median of 18 CNAs and 37 SVs in RT that surpassed the 1.1 mutations/megabase, 4 CNAs, and 5 SVs observed at CLL diagnosis (Figure 81.a). No major differences were seen among RT at CLL diagnosis or after different therapies (Figure 81.b).

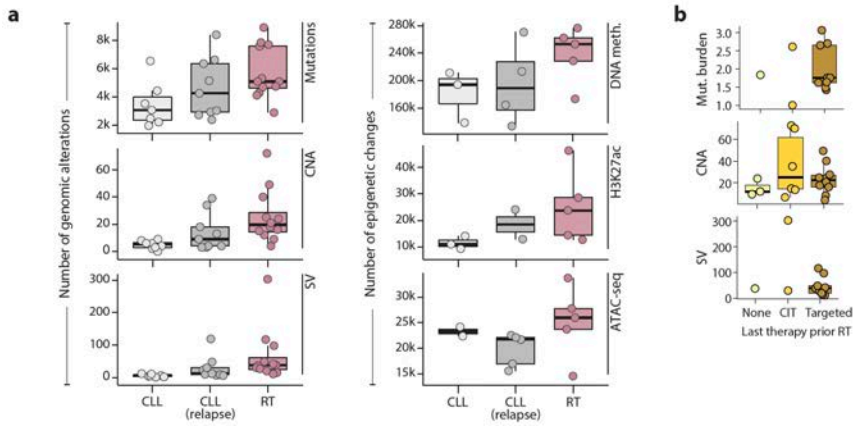


Figure 81. Genetic and epigenetic changes from CLL to RT. a. Increase in genomic alterations [mutations, including SNVs and indels, CNAs, and SVs] and epigenetic changes [number of DNA methylation, ATAC-seq, and H2K27ac changes] compared to normal naïve and memory B cells over the disease course. b. Mutational burden (mutations/megabase), CNAs, and SVs found in RT stratified according to therapy prior to transformation. Targeted, targeted therapies.

This comprehensive genomic characterization recognized novel driver genes and mechanisms, expanding the list of altered pathways in RT (Anderson et al., 2017; Chakraborty et al., 2021; Chigrinova et al., 2013; De Paoli et al., 2013; Fabbri et al., 2013; Herling et al., 2018; Klintman et al., 2021; Rossi et al., 2011, 2012; Scandurra et al., 2010; Villamor et al., 2012) (Figure 82, Figure 83, and Figure 84). The main alterations affected cell cycle regulators (17/19, 89%), chromatin modifiers (79%), MYC (74%), NF-κB (74%), and NOTCH (32%) pathways.

These genomic variations were simultaneously present in most cases, except for MYC and NOTCH altered pathways, which only co-occurred in 2/19 cases (Figure 82). Alterations in some genes such as *TP53, NOTCH1, BIRC3*, *EGR2,* and *NFKBIE* were usually present and clonally dominant from the sample at CLL diagnosis, while others were only detected during the disease course or at the time of RT (e.g., *CDKN2A/B*, *CDKN1A/B*, *ARID1A*, *CREBBP, TRAF3,* and *TNFAIP3*) (Figure 82).

Among recurrent CNAs found either in CLL or RT samples (n ≥5), deletions of 9p (*PTPRD* and *CDKN2A/B*) and deletions of 15q (*MGA*) were enriched in RT whereas deletions of *ATM* (11q), *TP53* (17p), and 13q14 were found at similar frequencies in CLL and RT (Figure 83).

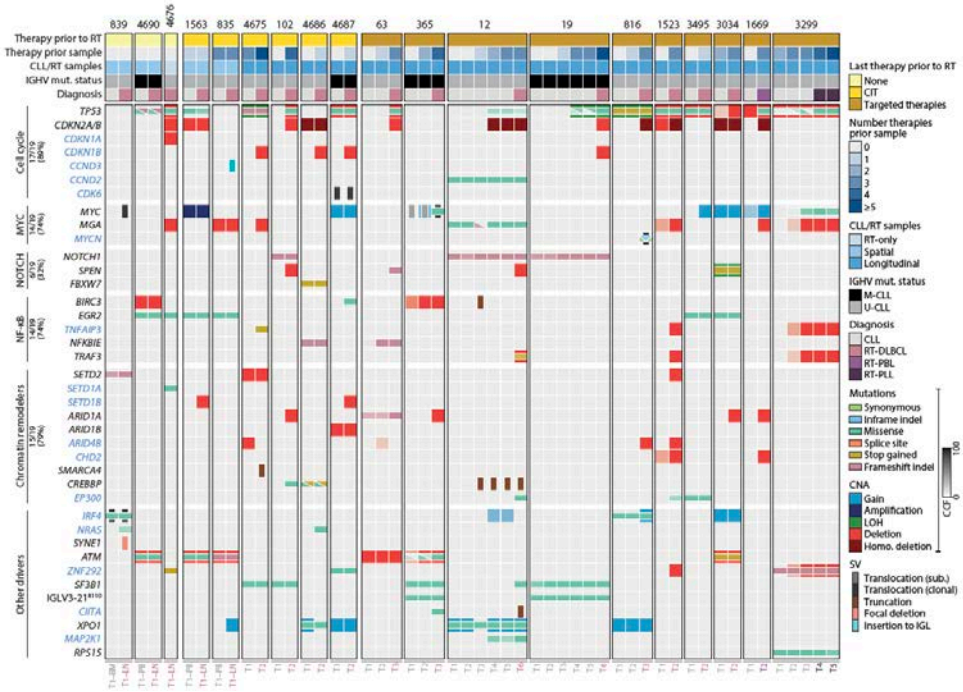


Figure 82. Landscape of driver alterations from CLL to RT. Oncoprint shows the list of putative driver alterations. Samples, grouped by patient (id on the top), are represented by columns while genes are displayed in rows. Novel drivers in RT are labeled in blue. Genes are grouped according to their biological function or if they were previously described as potential driver genes in CLL and/or mature B-cell lymphomas. Metadata including the type of therapy prior to RT, number of treatment lines before each sample, the spatial/longitudinal nature of the CLL-RT samples analyzed, the mutational status of the IGHV, and diagnosis is detailed in the upper rows. In the main plot, mutations (SNVs and indels) are depicted with horizontal rectangles, CNAs using the background color of each cell, and SVs with vertical rectangles. The transparency of the color of mutations and CNAs indicates their cancer cell fraction (CCF). For cases lacking the normal sample (case id indicated in gray), the CCF of the alterations could not be inferred and a CCF of 100% was used for illustrative purposes.

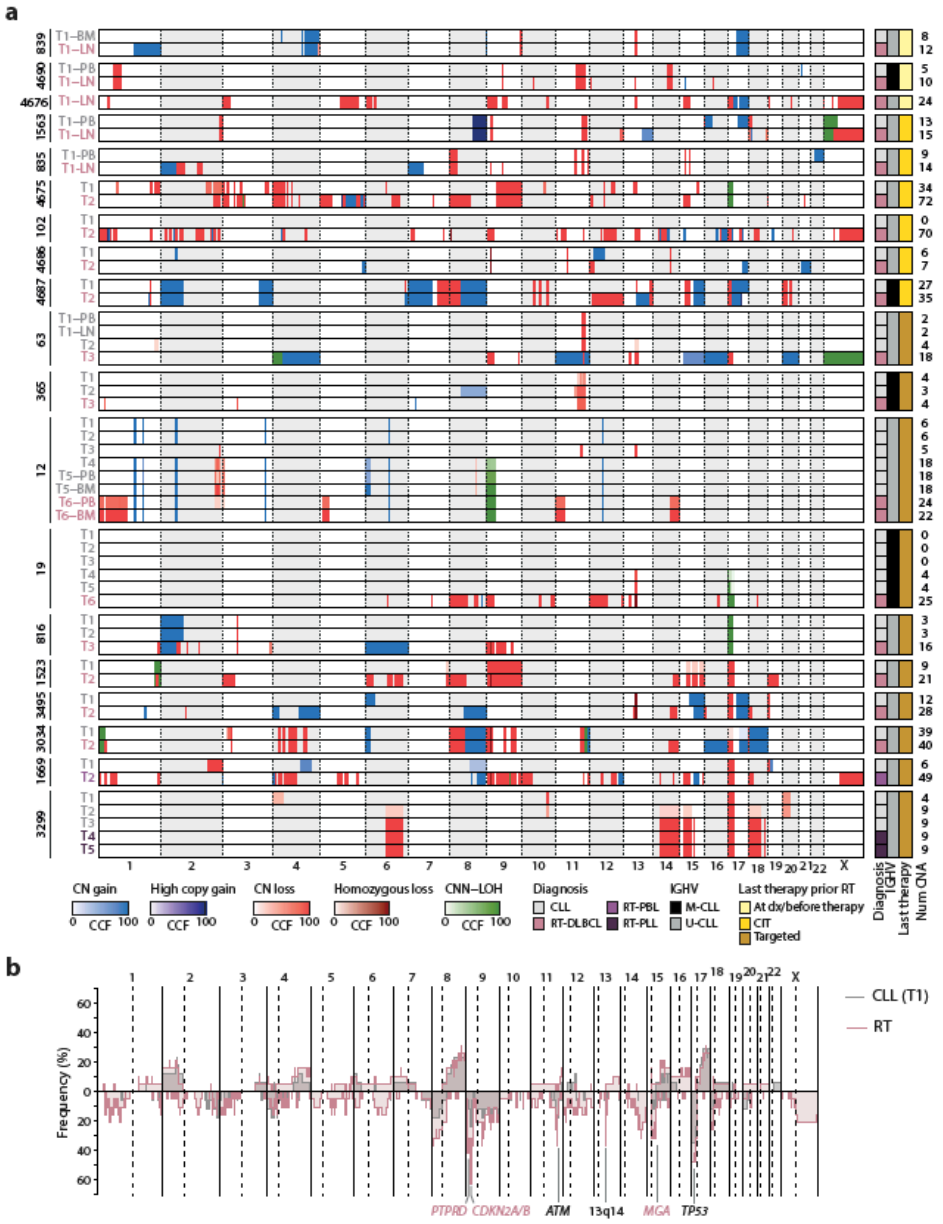


Figure 83. CNA profiles from CLL to RT. a. Copy number landscape grouped by patient. Diagnosis, IGHV mutational status, prior therapy, and total number of CNAs are indicated for each time point. The type of each CNA is indicated by its color, and the transparency is proportional to its CCF (when available). b. Aggregated copy number profile of RT vs CLL. The first CLL samples (time point 1, T1) were considered. The plot shows the percentage of samples with gains (up) and losses (down).

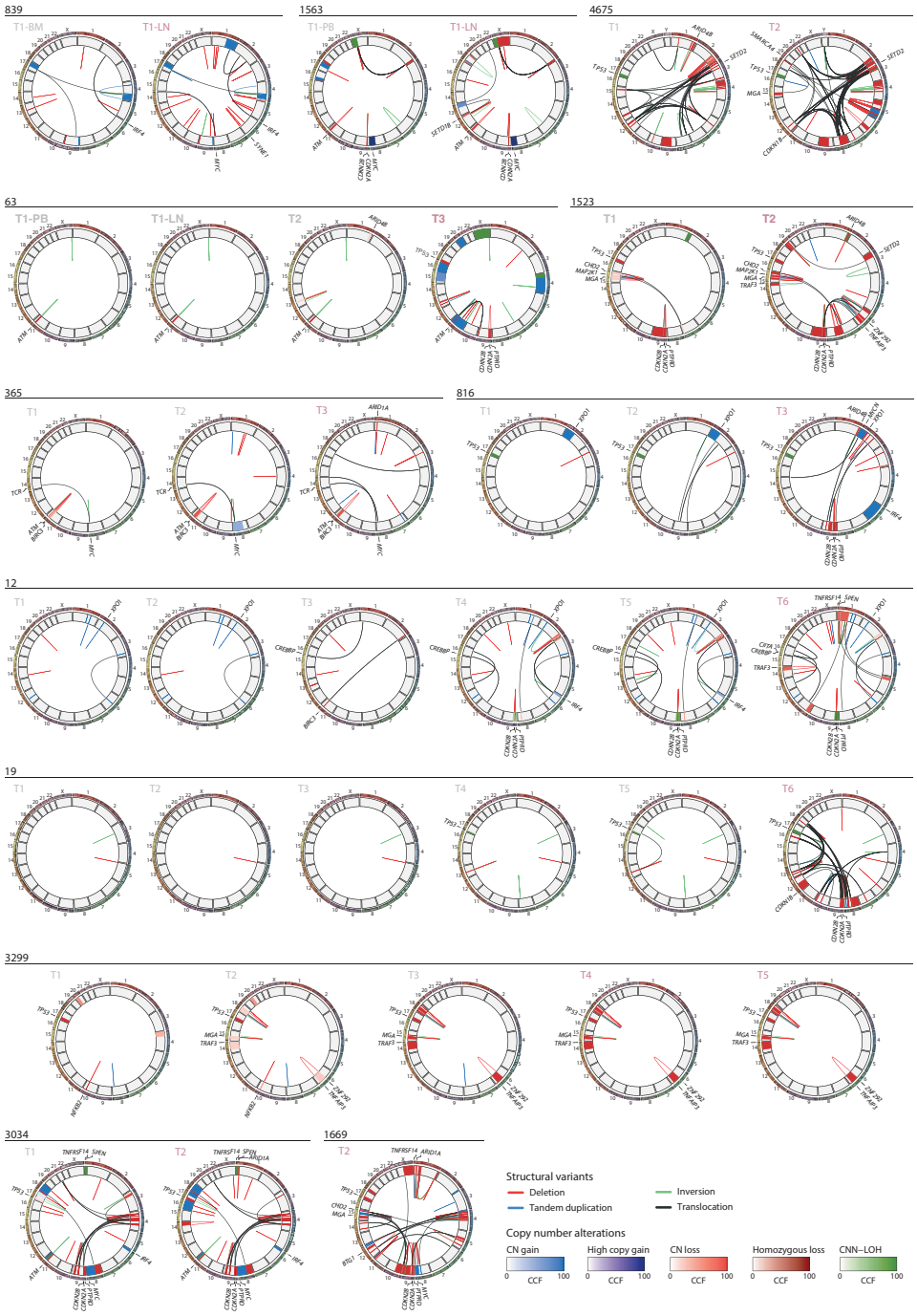


Figure 84. Structural variants in CLL and RT. Circos plots illustrating the CNAs and SVs of each sample, grouped by patient. Chromosomes are displayed in the outer circle. The next ring indicates the CNAs

colored by their type. The transparency is proportional to their CCF. The inner circle represents SVs, linking together the breakpoints of the affected loci. Candidate driver genes affected by CNAs and/or SVs are annotated.

We identified novel alterations, including deletions targeting *CDKN1A* and *CDKN1B* in 5 RT associated with downregulation of their expression (Figure 85.a), one immunoglobulin (IG)-*CDK6* translocation (Figure 85.b) and one *CCND2* mutation already present at CLL diagnosis, and a *CCND3*-IG translocation acquired at RT (Figure 85.c). Similarly, we also detected a *MYCN*-IG translocation that correlated with the overexpression of the gene (Figure 85.d). Most chromatin remodeler genes were altered by deletions and reduced their expression. Intriguingly, some of these genes were also downregulated in RT cases lacking these deletions, suggesting that other mechanisms might converge into similar transcriptomic profiles. Novel alterations in this group were deletions of *ARID4B* and truncations of *CREBBP* (Chitalia et al., 2019) and *SMARCA4* (Klintman et al., 2021) by translocations and chromoplexy (Baca et al., 2013) (Figure 85.f-g). We also recognized recurrent *IRF4* alterations in RT, which have been associated with increased MYC levels in CLL (Benatti et al., 2021).

Contrary to previous literature (Kadri et al., 2017), *BTK*/*PLCG2* or *BCL2* mutations were not identified in any RT sample after treatment with BCR or BCL2 inhibitors, respectively. Interestingly, the two M-CLL cases developing RT after targeted therapies carried the IGLV3-21^{R110} mutation, which triggers cell-autonomous BCR signaling (Minici et al., 2017) (Figure 82).

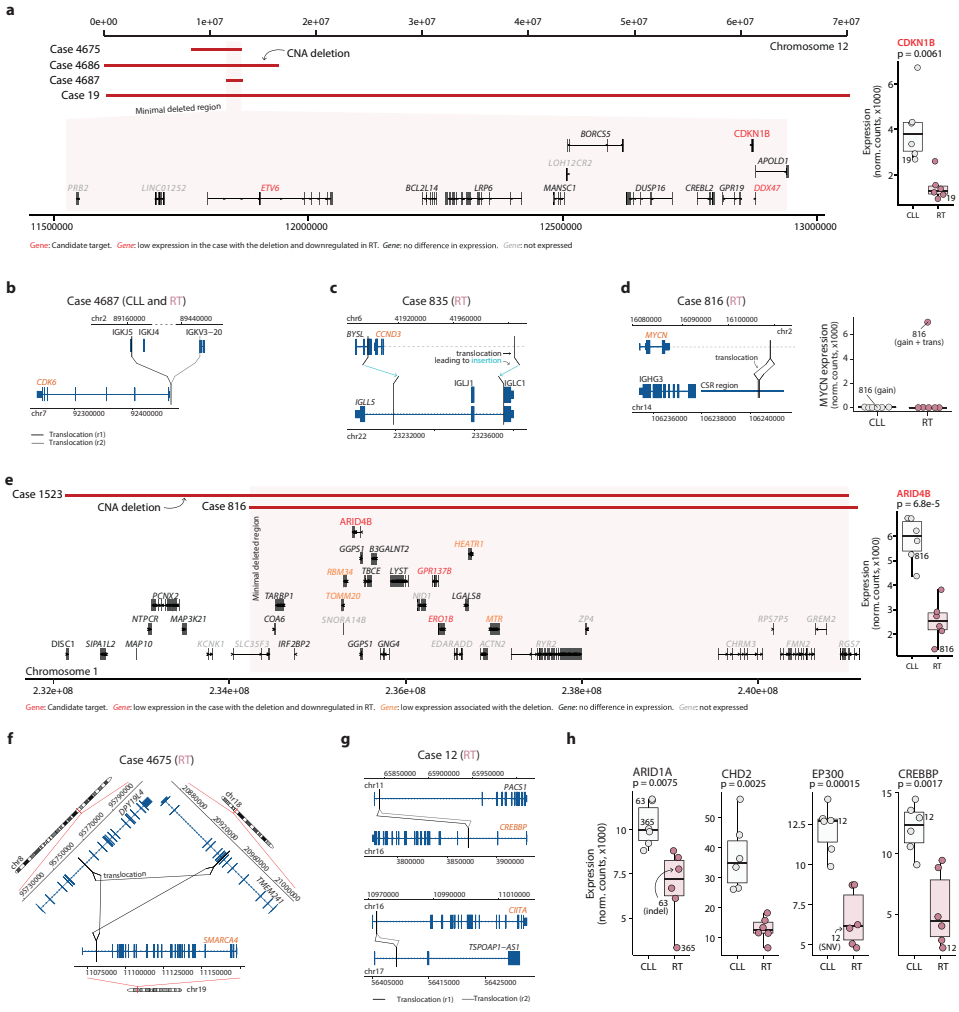


Figure 85. Structural variants and copy number alterations affecting driver genes in RT. a. Deletions in chr12 identified in four cases with the minimal deleted region affecting CDKN1B. Its expression in CLL and RT sample pairs is shown on the right. The case carrying the deletion at time of RT is labeled in the boxplot. b. Reciprocal translocation juxtaposing CDK6 next to IGKJ5 in case 4687. c. CCND3 insertion next to the constant region IGLC1 in the RT of case 835. d. Reciprocal translocation between MYCN and class switch recombination (CSR) region of IGHG3 in the RT sample of case 816 [left]. MYCN expression based on bulk RNA-seq [right]. e. Deletion in chr1 affecting two cases with the minimal deleted region targeting ARID4B. Its expression in CLL and RT sample pairs is shown in the boxplot (right). f. Chromoplexy disrupting SMARCA4 in the RT sample of case 4675. g. Reciprocal translocations truncating CREBBP and CIITA in the RT sample of case 12. h. Expression levels of known

and novel RT-driver genes in CLL and RT paired samples. Cases carrying deletions/mutations at the time of RT are indicated.

In addition to the high frequency of CNAs, which has already been described in RT (Beà et al., 2002; Fabbri et al., 2013) (Figure 83), we observed a high number of structural variants that clustered into complex structural rearrangements (Figure 84). Chromothripsis (Stephens et al., 2011), a rare event in CLL (Puente et al., 2015), was found in 8 RT, involved 1 to 3 chromosomes per event, and usually targeted one or more driver genes, as previously seen in other tumor types (Cortés-Ciriano et al., 2020; Maura, Bolli, et al., 2019) (Figure 86). In line with this, *CDKN2A/B* and *CDKN1B* where targeted by chromothripsis in 5 and 1 cases, respectively, and *MYC*, *MGA*, *SPEN*, *TNFAIP3,* as well as chromatin remodeling genes (*ARID1A*, *CHD2*, and *SETD2*) in additional cases (Figure 87).

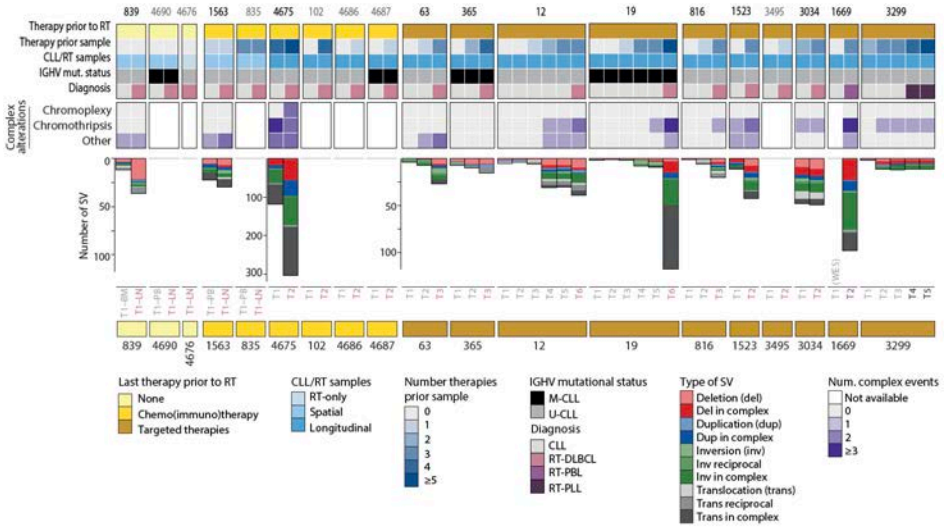


Figure 86. Structural variants and complex rearrangements in RT. a. Upper rows indicate the case and associated metadata. Next, the number of complex structural alterations in each sample is shown, together with the total number of SVs of each type (deletions, duplications, inversions, and translocations). The color indicates the type, while the transparency indicates if they belong to a complex structural rearrangement.

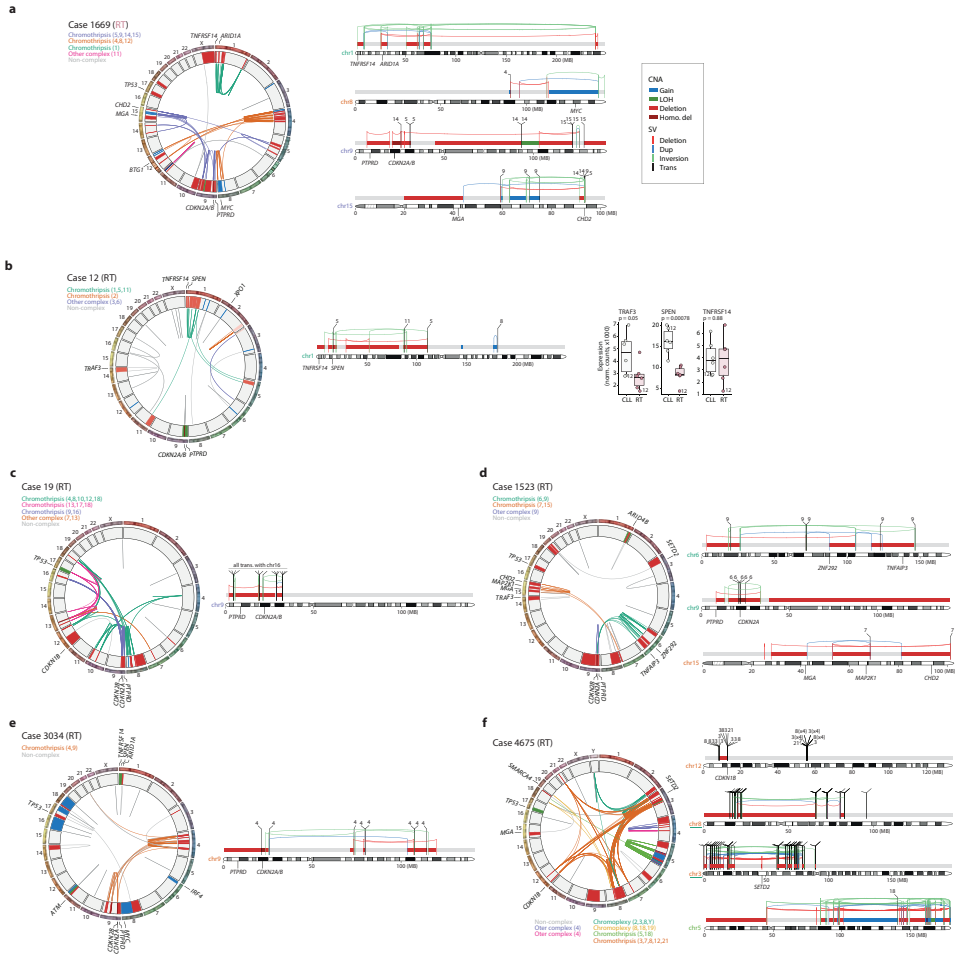


Figure 87. Complex genomic rearrangements affecting driver genes. a. The circos plot [left] displays the SVs (links) and CNAs (inner circle) found in the RT sample of case 1669. CNAs are colored by type while SVs according to their occurrence within specific complex events. Target driver genes are annotated. Chromosome-specific plots [right] illustrate selected complex rearrangements affecting one or more driver genes with CNAs and SVs colored by type. b. Complex rearrangements in case 12 and affected expression of targeted genes [right]. c-f. Additional cases with complex events.

Taken together, our analyses expand the catalog of driver genes, pathways, and mechanisms involved in RT, and identify a similar distribution of these alterations in RT arising after different treatments, suggesting that therapy-

specific pressure is not a major determinant of the driver genomic landscape of these tumors.

Novel mutational processes in RT

We next performed a mutational signature analysis to explore the mutational processes that could shape the genome of CLL and the increased mutational burden and genomic complexity of RT.

First, we integrated the CLL (time point 1) and RT samples from our CLL-RT cohort with 147 CLL obtained prior to therapy from the ICGC-CLL cohort and an independent cohort of 27 CLL post-treatment samples. An unsupervised analysis based on the percentage of the 96 classes of point mutations showed that the mutational profile of RT was notably different from the M-CLL and U-CLL (Figure 88). As shown in the plot, the first component differentiates M-CLL from U-CLL, which are known to be two different entities (see Introduction - section 1.5.3) and have different mutational profiles (i.e., mutational signature SBS9 is only found in M-CLL). The second component separates RT from CLL samples, which led us to think that there might be different mutational processes operating in RT.

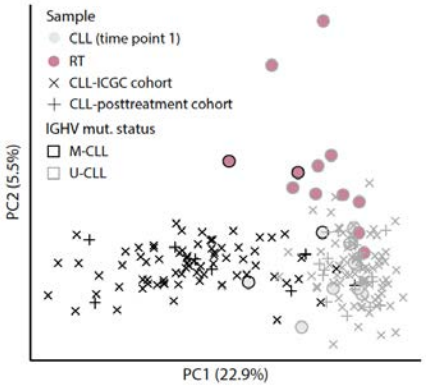


Figure 88. Mutational profile of CLL and RT. Principal component analysis of the 96-mutation profile of CLL and RT. Samples from the CLL-RT study are represented as circles, while CLL samples from the

CLL-ICGC and CLL-post-treatment cohorts are represented as crosses and plus signs, respectively. Outline colors indicate the mutational status of the IGHV of each case.

Thereafter, we extracted de novo the mutational signatures from genome-wide and clustered mutations identified in the CLL-RT and ICGC-CLL cohorts altogether. The resulting signatures were assigned and/or decomposed into reference signatures from COSMIC, obtaining 10 and 2 genome-wide and clustered mutational signatures, respectively (Figure 89).

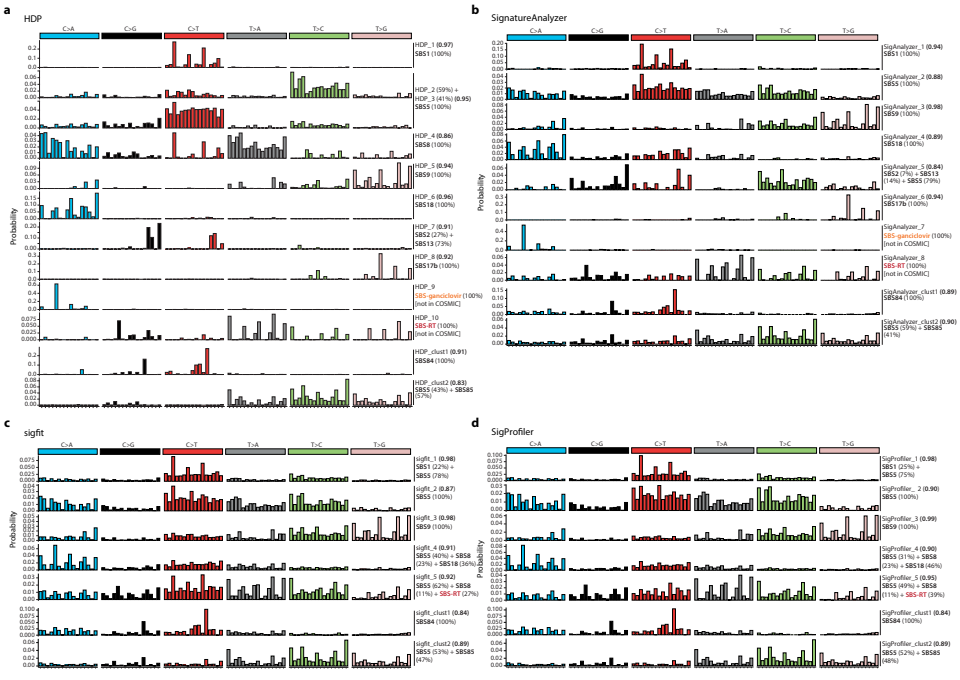


Figure 89. Mutational signatures extraction. Signatures extracted by the Hierarchical Dirichlet Process (HDP) (a), SignatureAnalyzer (b), sigfit (c), and SigProfiler (d). Decomposition of the extracted signatures into COSMIC signatures together with their percentage contribution are shown at the right of each profile. The cosine similarities between the extracted and reconstructed signatures are shown in brackets.

Among the ten genome-wide processes, 5 were previously identified in CLL and DLBCL [SBS1 and SBS5 (clock-like), SBS8 (unknown etiology), SBS9 (attributed

to polymerase-eta), and SBS18 (possibly damage by reactive oxygen species)] and 3 had been only found in DLBCL [SBS2 and SBS13 (APOBEC enzymes), SBS17b (unknown)]. Note that APOBEC-related signatures (SBS2 and SBS13) were not precisely extracted by the algorithms but manually identified based on their remarkable contribution among RT-private mutations of one case (839). One signature had been recently described in other tumors exposed to ganciclovir (nucleoside analog used primarily in treating cytomegalovirus infections) (de Kanter et al., 2021) and was named SBS-ganciclovir. Finally, another signature, characterized by [T>A]A and, in a lower degree, [T>C/G]A mutations, was considered a novel mutational process and named SBS-RT (Figure 90). This signature has not identified previously in any tumor type including CLL and DLBCL (Alexandrov et al., 2020; Arthur et al., 2018; Kasar et al., 2015; Kucab et al., 2019; Maura, Degasperi, et al., 2019; Puente et al., 2015).

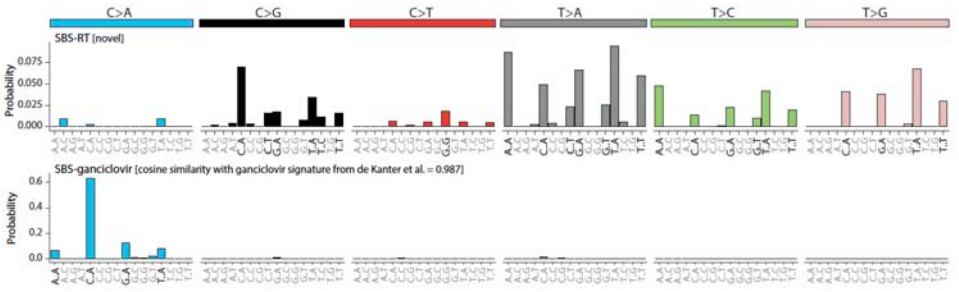


Figure 90. Signatures de novo identified in CLL-RT not reported in COSMIC. The main peaks of each signature are labeled in black.

SBS-RT was directly identified by two independent algorithms (HDP and SignatureAnalyzer) with a high cosine similarity of 0.947, and is required on top of all COSMIC signatures to properly reconstruct two signatures extracted by SigProfiler and sigfit, respectively (Figure 89 and Figure 91.a). To determine if SBS-RT represents a novel signature not identified in previous studies, we performed pairwise comparisons between SBS-RT and all COSMIC signatures, as well as a

compendium of environmental agents associated signatures (Kucab et al., 2019). Virtually none of these signatures had a cosine similarity >0.6 with SBS-RT, suggesting that SBS-RT did not correspond to any reference signature (Figure 91.b). Next, we explored if SBS-RT could be a combination of already known signatures. We decomposed it into "N" COSMIC and environmental agents related signatures (Kucab et al., 2019) using an expectation maximization approach as previously described (Lee-Six et al., 2019). The best reconstituted signature was composed of 4 COSMIC signatures and its similarity with SBS-RT was not enough to consider them equivalent. Their cosine similarity was 0.79, and, furthermore, visual inspection of the reconstructed signature showed that it was missing several SBS-RT representative peaks (Figure 91.c). Taken together, these results strongly suggest that SBS-RT is a novel mutational signature. In line with this, additional findings described below, allowed us to associate this signature with potential mutational processes.

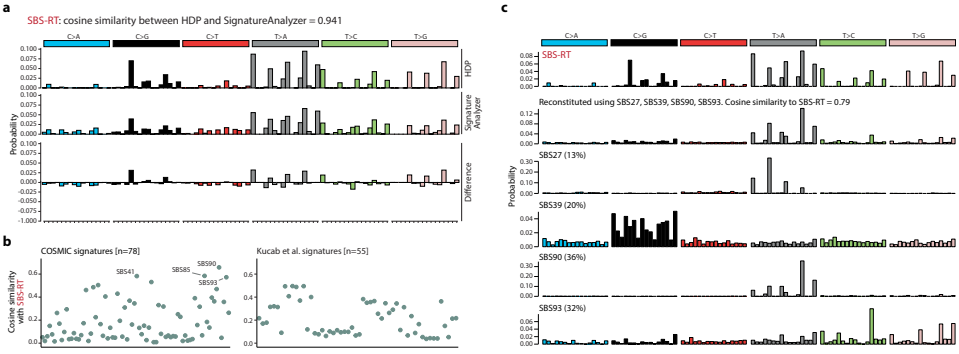


Figure 91. Novel mutational signature SBS-RT. a. Comparison of SBS-RT extracted by HDP and SignatureAnalyzer. Based on their high cosine similarity (0.941), we considered that both signatures represented the same mutational process and selected the one extracted by HDP for downstream analyses. b. Pairwise comparisons of the SBS-RT with known signatures from COSMIC and environmental agents (Kucab et al., 2019) *c. Decomposition of SBS-RT into "N" COSMIC and environmental agents signatures using an expectation maximization approach. The low cosine similarity (<0.80) between SBS-RT and the best reconstructed signature suggests that SBS-RT represents a novel mutational signature.*

Further genomic and chromatin-based characterization of this novel signature showed that SBS-RT mutations were present in all different chromatin states and early/late replicating regions although with a moderate enrichment in heterochromatin/late replication (Figure 92.a-b), a lack of replication and transcriptional strand bias (Figure 92.c-d), and a modest correlation between SBS-RT and SBS5 clock-like mutations (R=0.74, p=0.16) (Figure 92.e). Note that contribution to chromatin states and replication timings was done based on the mutations from the "CLL subclone" and the "RT subclone" identified in the reconstruction of the clonal composition and evolution by WGS.

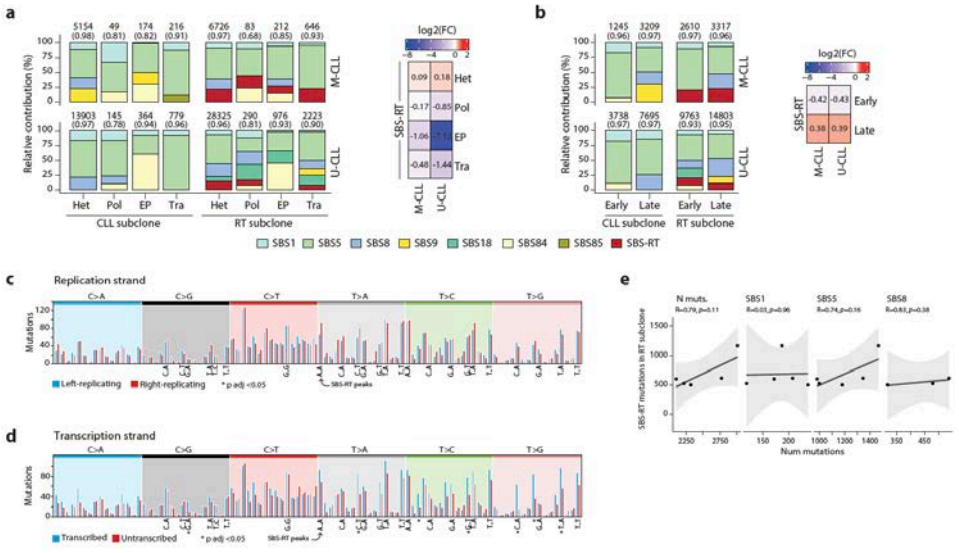


Figure 92. Characterization of SBS-RT. a. Activity of the mutational processes found in CLL and RT subclones in the different regulatory regions of the genome: heterochromatin (Het), polycomb (Pol), enhancer/promoter (EP), and transcription (Tra). Mutated and unmutated IGHV cases were analyzed separately. The heatmap [right] shows the log2-fold change of the observed vs the expected number of SBS-RT mutations per region. b. Contribution of the identified mutational processes in early and late replication regions. The heatmap [right] shows the log2-fold change between the observed and expected SBS-RT mutations per region. c-d. Replication (c) and transcriptional (d) strand bias of the mutational profile of RT subclones with SBS-RT. The number of mutations in the right and left replicating (c) or transcribed and untranscribed (d) strands are indicated with red and blue bars,

respectively. Specific SBS-RT peaks are indicated with their context at the x axis. Significant asymmetries are indicated with asterisks. e. Correlation of SBS-RT with the total number of SNVs and other mutational processes identified in RT subclones.

During the first review of our manuscript, one of the reviewers asked us to speculate what might be driving this new mutational process (SBS-RT), which led us to perform a throughout revision of the mutational signature analysis. We refined some considerations and preliminary hypotheses based on very recently published results, hinting at an in-depth exploration of all treatment lines that the patients received.

In the first submission, we considered the de novo extracted signature, now identified as SBS-ganciclovir, as a potential artifact and excluded it from further analyses. We initially flagged it as a technical artifact because it was only present in one case, had a striking 1-peak profile, and was similar to a COSMIC signature related to potential sequencing artifacts (SBS53, cosine similarity of 0.90). However, a posterior publication describing SBS-ganciclovir (named SBSA in the original paper) related it to treatment with the antiviral ganciclovir (de Kanter et al., 2021), which made us reconsider this preliminary decision. Indeed, our extracted signature had a very high similarity to this novel signature (cosine similarity of 0.987), and it was found in the RT sample of 1 case (case 4675), which had received valganciclovir (a prodrug of ganciclovir) due to cytomegalovirus reactivation at three different time points before RT (Figure 93 and Figure 44).

This finding brought our attention to all the treatments that our patients had received. Two patients had been treated with melphalan, an alkylating agent associated with a previously described signature (Maura et al., 2021), here named SBS-melphalan (originally referred to as SBS-MM1). On that account, we included it in the analysis based on its contribution to 3 RT samples, 2 of which correspond to patients who had received melphalan as a conditioning of their allogeneic stem

cell transplant 1.9 and 4.2 years prior to RT (cases 4675 and 1523), respectively. The third RT sample with SBS-melphalan contribution was case 102, which did not receive melphalan and was the one with the lowest number of mutations assigned to SBS-melphalan, suggesting that another non-identified mutational process might be hampering the results in this case. We speculated that this mutational signature could indeed be present in melphalan-exposed tumors but that it could had not been extracted de novo in our analysis due to the low number of cases treated with this drug and its slight similarity, although remarkably different, to the novel SBS-RT (cosine similarity = 0.549).

Next, we determined the prevalence of each identified mutational signature in each sample analyzed. This analysis revealed that SBS-RT was present in the RT sample of 7/18 cases, 1/6 after CIT and 6/10 under multiple treatments including targeted therapies and was detected in all subtypes of transformation (RT-DLBCL, RT-PBL, and RT-PLL) (Figure 93). It was also found in CLL samples before RT in two cases (12 and 3299), but not in other CLL samples from the reanalyzed pre-treatment ICGC-CLL cohort nor in the independent CLL-post-treatment cohort (Figure 94). Note that none of the cases included in these two supplementary cohorts had evidence of RT, with a median follow-up of 9.8 years (range 0.2-30.4).

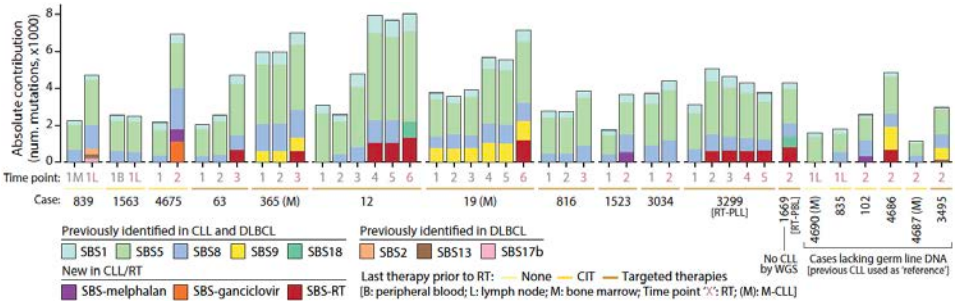


Figure 93. Contribution of mutational processes in CLL and RT. Each bar corresponds to a sample, grouped by patient. The colors indicate the contribution of the mutational signatures to the

mutational profile of each sample. RT time points are marked in rose. Last therapy prior to RT is indicated by horizontal color lines. CIT, chemoimmunotherapy.

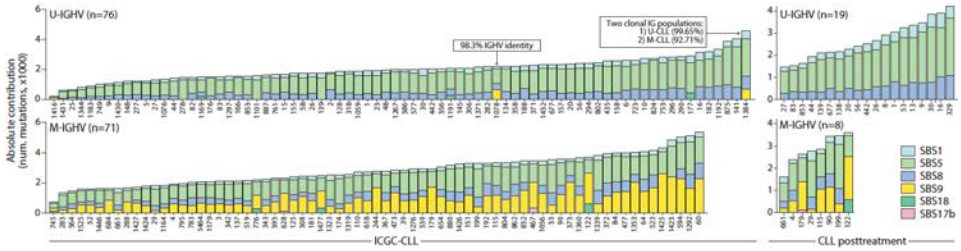


Figure 94. Contribution of mutational processes in CLL diagnosis and post-treatment samples. Analysis of mutational signatures in two additional cohorts: ICGC-CLL cohort (left) including 147 CLL samples at diagnosis, and CLL post-treatment cohort (right) of 27 samples of CLL patients after treatment. Cases are grouped based on their IGHV mutational status and sorted based on their number of mutations.

Although we first hypothesized that SBS-RT could correspond to a RT-specific mutational process, due to its absence in CLL and DLBCL samples of non-RT cases, an in-depth review of the treatment lines that the patients received suggested that SBS-RT might represent the footprint of an early-in-time therapy.

Intriguingly, all cases showing contribution of the novel SBS-RT at time of RT had been treated with the alkylating agents bendamustine (n=5) or chlorambucil (n=2) during their CLL course at a median of 2.9 years (range from 0.7 to 6.8 years) before RT. Contrarily, cases without SBS-RT had never been exposed to these drugs (Figure 95). This observation allowed us to narrow down the potential mutational process behind SBS-RT, which seemed to be related to the exposure to these alkylating agents.

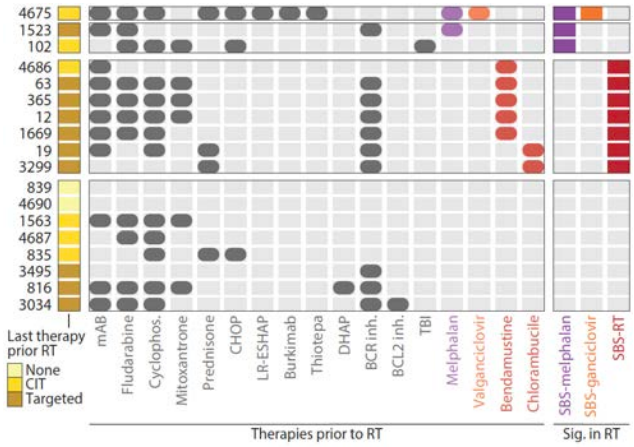


Figure 95. Therapies prior to RT. Representation of the treatments that each patient received before RT. The presence/absence of mutational signatures SBS-melphalan, SBS-ganciclovir, and SBS-RT at time of RT is shown on the right. mAB, monoclonal antibody. TBI, total body irradiation. Inh, inhibitor.

To better time the activity of each mutational process, we reconstructed the phylogenetic tree of the tumor subclonal composition for 11 patients with multiple synchronous (n=2) or longitudinal (n=9) and matched germline samples, 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". As previously described, 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 cases (365 and 19) representing early events. DLBCL-related signatures, and treatment-related signatures (SBS-ganciclovir, SBS-melphalan, and SBS-RT) were found in single RT subclones in 6 cases while 2 cases carried two simultaneous subclones with SBS-RT (cases 12 and 19) (Figure 96.a). SBS-RT contributed with 28.6% of the mutations acquired in RT (mean 679, range 499-1,167), similar to DLBCL-related signatures and SBS-melphalan/SBS-ganciclovir, led to the higher mutational burden compared to RT lacking these signatures, and it was occasionally associated with coding mutations in driver genes (*EP300* and *CIITA*) (Figure 96.a and Figure 96.b).

RT subclones were also characterized by the acquisition of kataegis (Nik-Zainal, Alexandrov, et al., 2012) that was mainly found within the IG loci, mediated by the 2 mutational signatures identified in localized regions of the genome related to direct and indirect effects of activation-induced cytidine deaminase (AID) activity (SBS84 and SBS85, respectively) (Alexandrov et al., 2020; Kasar et al., 2015). These kataegis led to the acquisition of mutations in the rearranged V(D)J genes in 5 RT (1 after CIT, 4 after targeted therapies) (Figure 96.c). This canonical AID activity in RT is in line with the acquisition of SBS9 mutations in two RT samples (4686 [CIT] and 3495 [targeted therapies]) and SVs mediated by aberrant class-switch recombination or somatic hypermutation in 6 RT (1 before therapy, 2 CIT, 3 novel agents), which targeted *MYC*, *MYCN*, *TRAF3*, and *CCND3* (Figure 82).

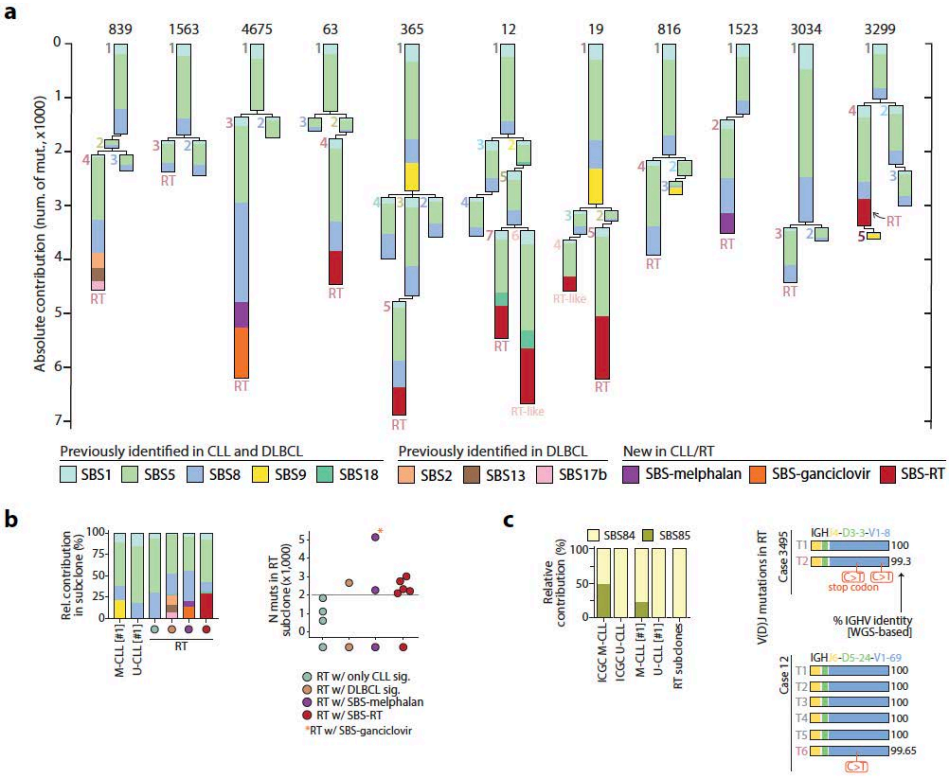


Figure 96. Mutational processes in Richter transformation's clonal evolution. a. Phylogenetic relationship of subclones and contribution of each mutational signature to their mutational profile.

b. Relative contribution of mutational processes in CLL (#1) and RT subclones [left]. Number of mutations in RT subclones [right]. RT subclones are defined by the mutational signatures identified. c. 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].

To better characterize and validate the timing of mutations introduced by SBS-RT, we applied a high-coverage, unique-molecular identifier (UMI)-based next-generation sequencing (NGS) approach in longitudinal samples of cases 12, 19, and 63. 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 mutations not associated with SBS-RT were detected earlier during the disease course. These results suggest a causal link between the exposure to these drugs and SBS-RT (Figure 97.a).

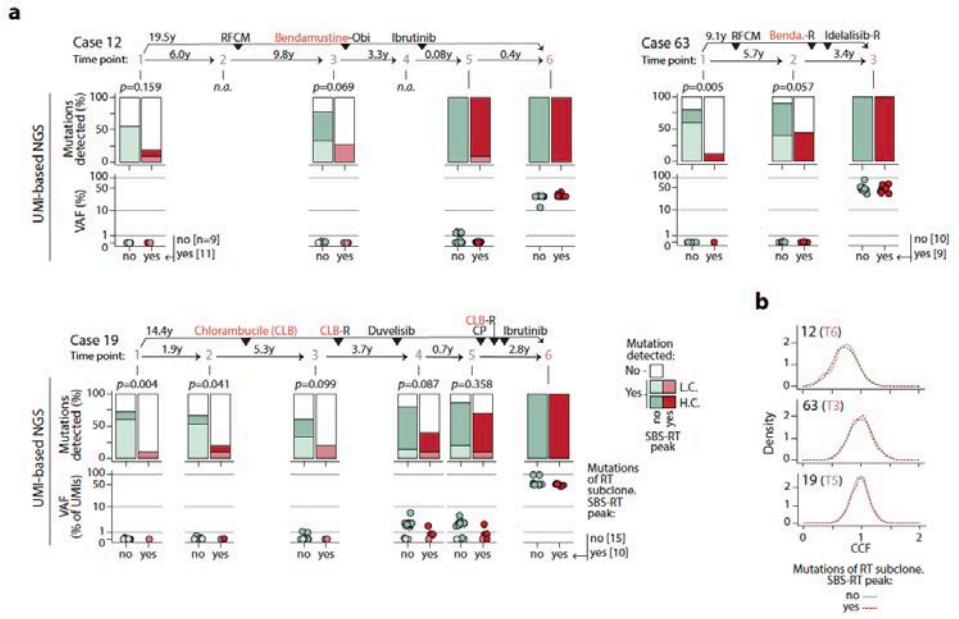


Figure 97. Timing of RT mutations by high-coverage UMI-based analysis. a. Detection [top] and variant allele frequency (VAF) [bottom] of mutations assigned to the RT subclone during the disease course in cases 12, 63, and 19 by high-coverage, UMI-based NGS. Mutations are grouped according to the main peaks of SBS-RT. P values by Fisher's test. L.C., low confidence; H.C., high confidence. b.

Distribution of the cancer cell fraction (CCF) of the SNVs assigned to the RT subclone based on WGS and stratified according to the main peaks of the SBS-RT.

The identification of SBS-melphalan, SBS-ganciclovir, and SBS-RT in RT supports the model of a single-cell that can carry the footprints of cancer therapies, which can only be detected after its expansion (Figure 97.b). Contrarily, the non-detection of SBS-RT in the 27 CLL-post-treatment cases (7 treated with bendamustine or chlorambucil) suggests that CLL relapse in these cases might be driven by the simultaneous expansion of different subclones, hindering the discovery of SBS-RT by bulk sequencing (Pich et al., 2019; Rustad et al., 2020).

As previously explained, SBS-RT mutations were found in CLL samples prior to the transformation. In case 3299, it was only present in the RT subclone, which suggests that the RT subclone was already present and could be detected before its clinical manifestation (Figure 93 and Figure 96.a). In addition, SBS-RT was 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 "transformed" or "RT-like" cells that could have been missed by the routine analysis. To prove our hypothesis, we reanalyzed the flow cytometry data available for case 12 at different time points. From time point T4, we detected two cell subpopulations differing in size and specific surface markers (likely CLL and RT-like subclones), whereas at T5 we detected an additional subpopulation of larger cells (RT subclone) (0.2% cells) which expanded at T6 outcompeting the previous RT-like subpopulation (Figure 98). These dynamics correspond to the subclonal evolution identified by WGS, which showed that the RT-like and RT subclones diverged from a cell carrying deletion of *CDKN2A/B* and truncation of *CREBBP (subclone #5)*, each acquiring more than 2100 specific mutations (Figure 96.a).

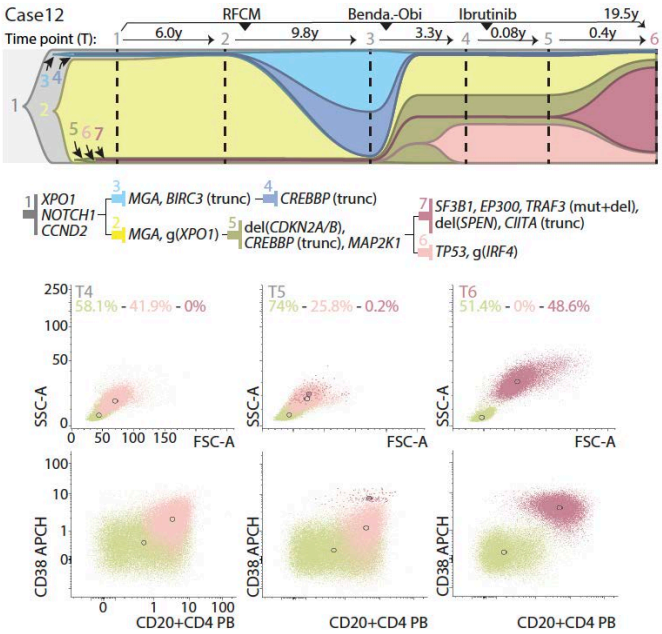


Figure 98. Clonal evolution along the disease course in case 12 inferred from WGS [top]. Abbreviations for treatment regimens are detailed in Figure 44. Each subclone is depicted by a different color and number, and its CCF is proportional to its height in each time point (vertical line). The phylogeny of the subclones with the main driver alterations is shown [middle]. Flow cytometry analysis for time points (T) 4, 5, and 6 [bottom]. The size of the cells (FSC vs. 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, respectively). Numbers along axes are divided by 1000.

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 points to the clonal expansion of a pre-existing single cell that later in time has the capacity to expand and become the dominant clone leading to RT.

Dormant seeds of RT at CLL diagnosis

Following the previous idea, that the cell which expands at the time of RT might already be present long before transformation, we sought to further explore the evolutionary trajectories from CLL to RT. We first capitalized on the 9 cases with fully characterized longitudinal WGS to confirm the presence of minor RT subclones in early steps of CLL evolution.

The longitudinal nature of our study, including from 2 to 6 time points per case, allowed us to evaluate the intra-tumoral heterogeneity of these tumors, reconstruct their subclonal architecture, infer the phylogeny of these subpopulations of cancer cells, and understand their dynamics along the course of the disease.

The RT subclone was predicted to be present at low cancer cell fraction (CCF) in the preceding CLL samples in 5 of the 9 cases, and only detected at time of transformation in the remaining 4 (Figure 99).

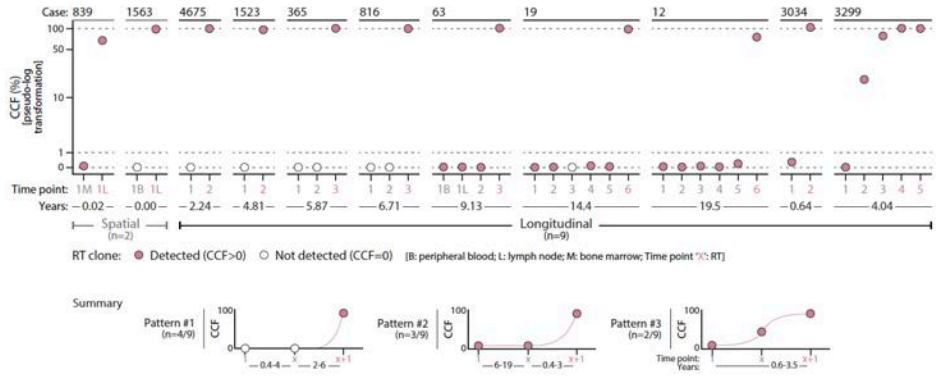
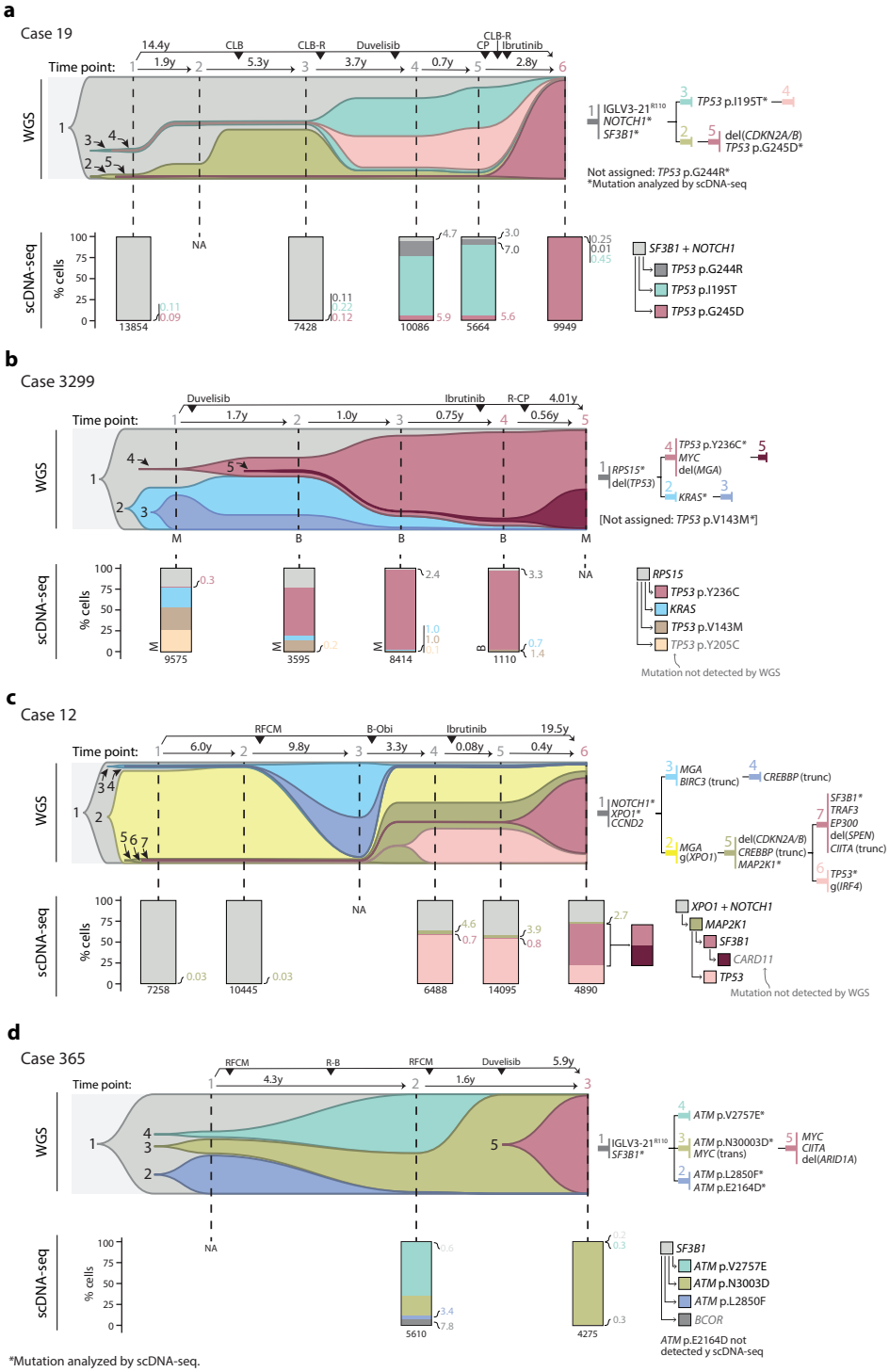


Figure 99. Early seeding of Richter transformation. Evolution of the RT subclone along the disease course based on WGS. The time lapse between the first and last sample analyzed is shown on the bottom. RT time points are marked in rose. The summary of the three patterns observed is shown on the bottom.

Among them, we identified two patterns of evolution based on the rapidness of expansion of the RT subclone. In 3/5 cases, the RT subclone was detected at time of CLL diagnosis, 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 (cases 12, 19, and 63) (Figure 99). In the other 2 cases, the RT subclone was also detected in the first CLL sample analyzed but promptly expanded driving the RT 0.6 and 3.5 years later in cases 3034 and 3299 (RT-PLL), respectively (Figure 99, Figure 100, and Figure 101).

Given the limitations of WGS data, we used a battery of more sensitive methodologies to confirm the subclonal architecture and dynamics of CLL evolution to RT that we initially inferred from WGS. These techniques included DNA- and RNA-based high-coverage sequencing of the immunoglobulin gene, scDNA-seq, and scRNA-seq. Note that we have also performed high-coverage, unique molecular identifier (UMI)-based NGS to time the acquisition of mutations of the RT subclone over time, which indirectly allowed us to validate the presence of these mutations in early time points. All these procedures confirmed the evolutionary trajectories inferred from WGS and thus supported the predicted early seeding of RT.

We performed scDNA-seq of 32 CLL driver genes in 16 longitudinal samples of 4 cases (12, 19, 365, and 3299) to validate the inferred evolutionary histories of RT (we obtained 202,210 cells passing filters, mean of 12,638 cells per sample). The scDNA analyses were performed by Dr. Ferran Nadeu, and we jointly evaluated the results and their integration with WGS.



*Mutation analyzed by scDNA-seq.
Time point 'X': RT.
B: peripheral blood; M: bone marrow.
del: deletion; g: gain; amp: amplificiation; LOH: copy number neutral loss of heterozygousity.
trans: translocaiton; trunc: truncation by SV; focal del: focal deletion by SV.

Figure 100. Clonal dynamics from CLL to RT with scDNA validation. a-d. Subclonal reconstruction and clonal evolution of four cases (19, 3299, 12, and 365) with WGS and scDNA-seq data. The upper fish plot shows the clonal evolution along the course of the disease inferred from WGS analysis. Each time point is indicated with dashed vertical lines, each color represents a different subclone and their height is proportional to their cancer cell fraction (CCF) in each time point. The treatments that the patient received and the elapsed time (in years) between samples are indicated on the top. The tissue is indicated for samples of case 3299 in which different tissues were analyzed by WGS and scDNA-seq in the same time point. The phylogeny of the subclones is depicted together with the main driver alterations [top right]. The lower bar plots show the dynamics of the different subclones according to the scDNA-seq analyses. The total number of cells per sample is shown on the bottom. The mutation tree inferred from scDNA-seq data is shown on the bottom-right part.

Focusing on case 19 with a time lapse of 14.4 years from CLL diagnosis to RT (Figure 100.a), the RT subclone (subclone #5) at transformation (T6) carried *CDKN2A/B* and *TP53* (p.G245D) alterations while the main CLL subclones (#3 and #4) driving the relapse after therapy at T4 and T5 harbored a different *TP53* mutation (p.I195T). The WGS predicted the presence of all these subclones at CLL diagnosis (T1). Using scDNA-seq we confirmed these predictions. Two small populations accounting for 0.1% of the cells carried the *TP53* p.I195T and p.G245D mutations, respectively, at diagnosis (T1) and were also detected at relapse 7.2 years later (T3). The subclone carrying *TP53* p.I195T expanded to dominate the relapse 3.7 years later 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. The scDNA-seq of the 3 additional cases also corroborated the phylogenies and most of the dynamics inferred from WGS (Figure 100.b-d). These results strongly suggest that CLL evolution to RT is characterized by an early driver diversification probably generated before diagnosis. RT may be emerging, at least in a notable fraction of patients, by a selection of pre-existing subclones carrying potent driver mutations rather than a de novo acquisition of leading clones.

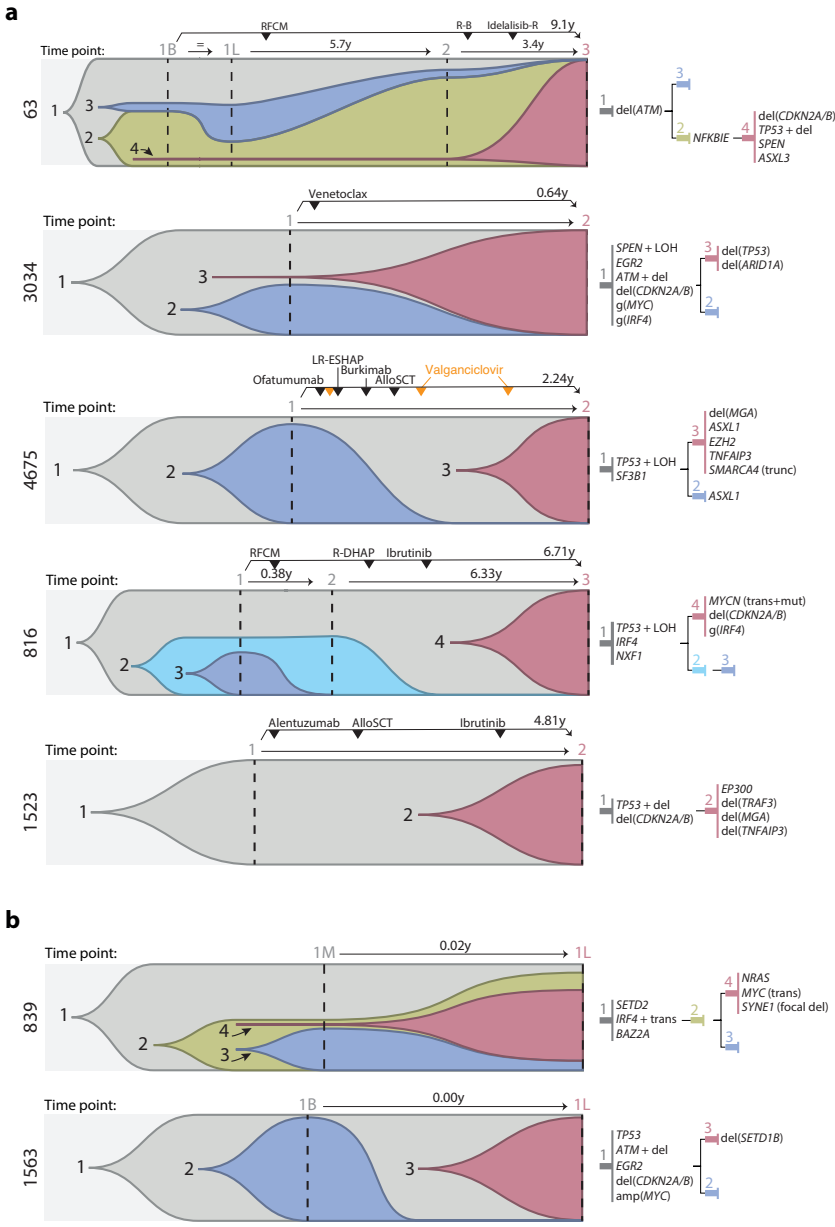


Figure 101. Clonal dynamics from CLL to RT from WGS. a-b. Subclonal architecture and dynamics of six cases with longitudinal samples (a) and two cases with spatial samples (b) analyzed by WGS.

The observed subclonal heterogeneity of CLL before diagnosis is concordant with the massive diversification revealed by single cell DNA methylation (Gaiti et al., 2019) and high-coverage IG (Bagnara et al., 2021; Gemenetzi et al., 2021) analyses. We identified 5 RT carrying specific mutations in the IG genes by WGS which represent a hallmark of their cell of origin. We analyzed if these IG-based RT subclones were already present at the time of CLL diagnosis using high-coverage NGS in cases 12 and 3495. The lack of germline material and cryopreserved cells for case 3495 precluded our complete WGS analysis and we could not reconstruct its subclonal structure and phylogeny. Nonetheless, using a modified variant calling strategy that used the CLL sample as the normal sample, we could identify that the RT occurring after treatment with the BCR inhibitor ibrutinib harbored two novel V(D)J mutations, which generated an unproductive IGH gene.

Using high-coverage NGS (performed by Dr. Ferran Nadeu) we recognized 0.002% sequences carrying the same two mutations at CLL diagnosis 1.72 years before (Figure 102.a). In addition, we also observed the expansion of additional unproductive subclones accounting for 11.8% of all sequences at time of RT, which suggests that BCR-independent subclones may have a proliferative advantage under therapy with BCR inhibitors (Figure 102.a). Similar results were found in case 12 in which the V(D)J sequence of RT carried a novel mutation that could be identified at CLL diagnosis 19.5 years before transformation. This finding was confirmed at both DNA and RNA level (Figure 102.b). It is known that immunogenetic features represent a faithful imprint of the B cell of origin, thus the identification of the same immunogenetic subclone (i.e., with the same IG gene rearrangement) before clinical manifestation of RT provides further evidence for the existence of early seeding.

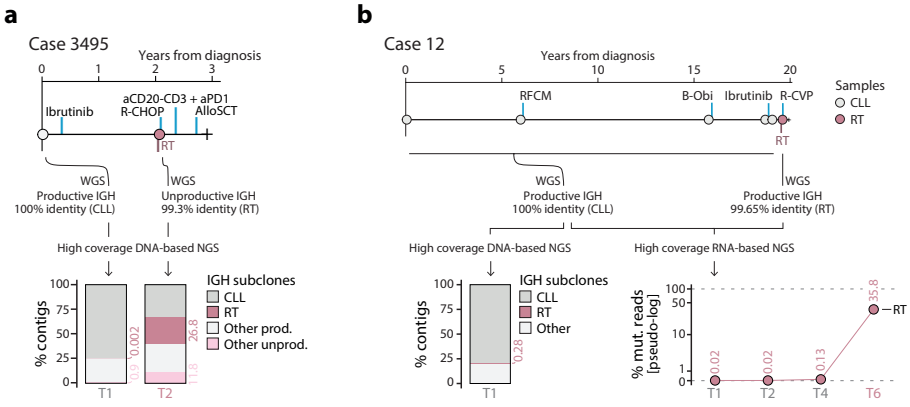


Figure 102. Early seeding of RT based on high-coverage IG. a. Representation of the clinical course and samples analyzed for case 3495 (top) together with the size of the IGH subclones identified using high-coverage NGS analyses (bottom). Subclones with specific RT mutations are indicated in dark pink. Abbreviations for treatment regimens are detailed in Figure 44. b. Clinical course (top) and IGH subclones identified by DNA- and RNA-based NGS (bottom) in case 12.

We finally investigated the transcriptome of RT using scRNA-seq of 19 longitudinal samples from 5 cases (24,800 tumor cells passing filters, mean of 1,305 cells/sample; microenvironment cells were not analyzed due to its low numbers) to verify if dormant RT subclones could be also identified at CLL diagnosis or during the disease course based on their phenotype. These analyses were performed by Ramon Massoni from Dr. Holger Heyn's group at CRG/CNAG.

As expected, RT and CLL cells had remarkably different gene expression profiles (Figure 103a, Figure 104). The transcriptome of CLL cells was characterized by different expression of CXCR4, CD27, and MIR155HG, that may describe the continuous recirculation of CLL cells between peripheral blood and lymph nodes (Calissano et al., 2009, 2011; Cui et al., 2014) and defined three main clusters identified across patients (Figure 103a-b, Figure 104). On the other hand, RT intra-clonal heterogeneity was mainly linked to distinct proliferative capacities. A cluster of cells showed high MKI67 and PCNA expression as well as high S and

G2M cell-cycle phase scores, while the remaining RT clusters showed a more quiescent state and were characterized by the expression of different marker genes among patients including CCND2, MIR155HG, and TP53INP1 (Figure 103a-c, Figure 104).

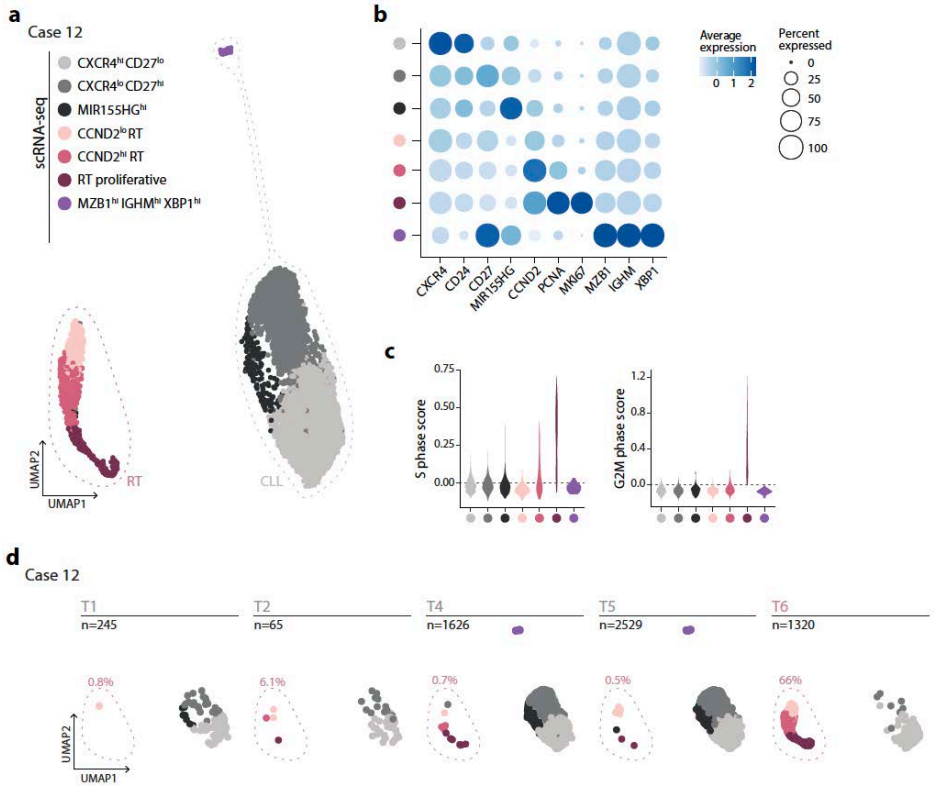


Figure 103. Early seeding of RT based on scRNA-seq for case 12. a. UMAP visualization based on the scRNA-seq data of all time points colored by annotation. RT clusters are colored in pink, while CLL clusters use gray colors. b. Expression of key marker genes in each cluster identified. Color and size represent scaled mean expression and proportion of cells expressing each marker gene, respectively. c. Distribution of cell cycle phase scores for each cluster based on scRNA-seq. d. UMAP plot split by time point, the fraction of RT cells is annotated. 'n', number of cells.

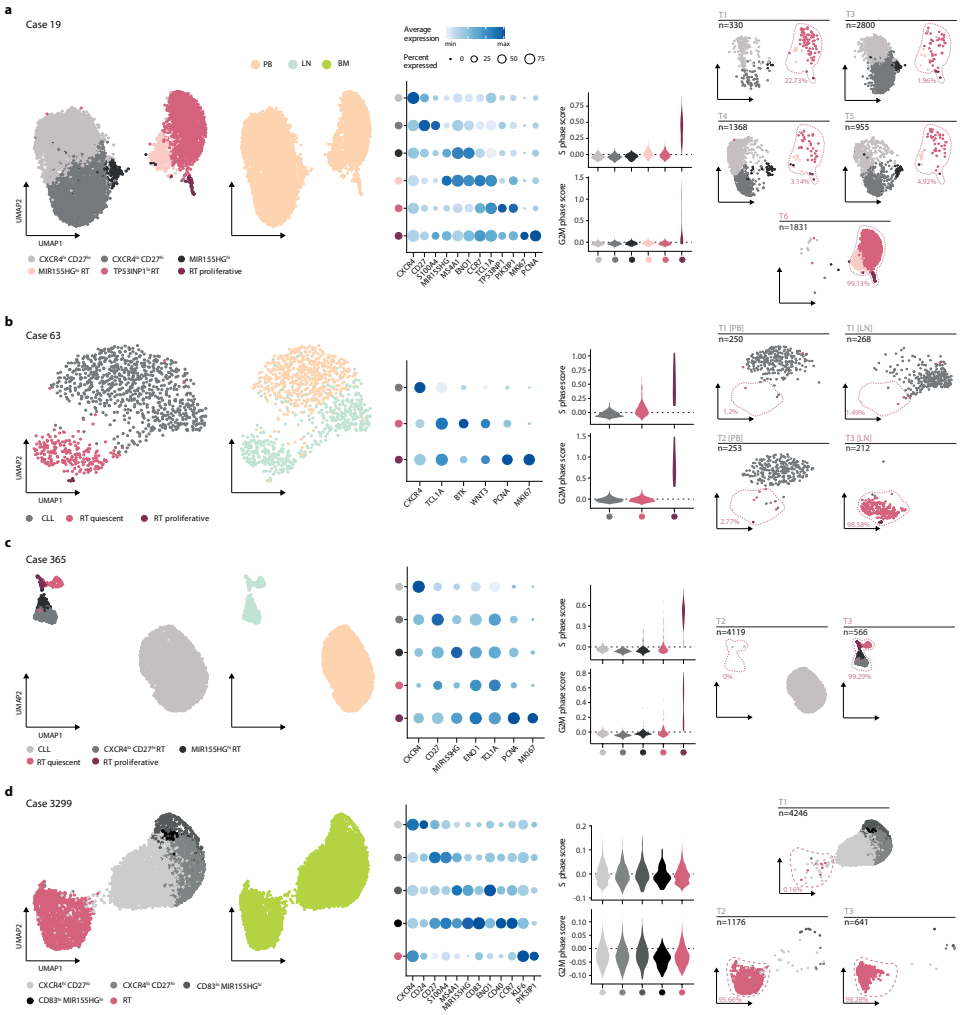


Figure 104. scRNA-seq characterization of CLL and RT. a-d. UMAP visualization of tumor cells from all time points colored by annotation and tissue of origin for cases 19 (a), 63 (b), 365 (c), and 3299 (d). hi, high; lo, low; PB, peripheral blood; LN, lymph node; BM, bone marrow [left]. Dot plot with the expression of key markers in each cluster identified. Color and size represent scaled mean expression and proportion of cells expressing each marker gene, respectively [middle-left]. Violin plots showing the cell cycle phase scores (S and G-to-M) for each cluster of cells [middle-right]. UMAP visualization split by time point [right]. 'n' refers to the total number of cells in that time point, and the percentage refers to the proportion of cells within RT clusters.

When considering each time point separately, we detected RT cells in all CLL samples prior to transformation in case 12, 19, 63, and 3299 but not in case 365 (Figure 103d, Figure 104, Figure 105a-e). To validate these observations, we reanalyzed the longitudinal scRNA-seq dataset from Penter and colleagues (Penter et al., 2021), which consists of 9 CLL patients, one of whom developed RT. In this case, we identified RT cells in the CLL sample collected 1.6 years prior to the expansion and diagnosis of RT (Figure 105f).

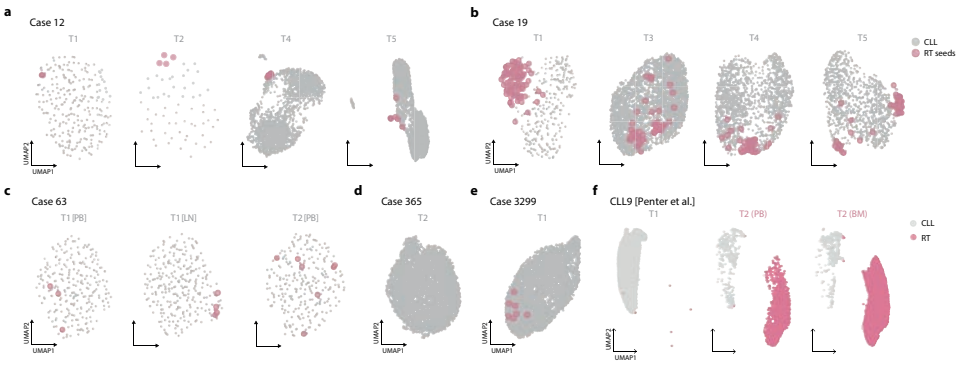


Figure 105. Sample specific visualization of scRNA-seq profiles. a-e. Time point-specific UMAP visualizations for each case. RT seed cells are depicted in rose and with an increased size. f. UMAP visualization of case CLL9 from Penter et al., 2021 split by time point. PB, peripheral blood; BM, bone marrow.

Overall, the presence and dynamics of these RT subclones according to their transcriptomic profile recapitulated the findings obtained by WGS, scDNA-seq, and IG analyses in all 5 cases, suggesting that they captured the same cells. As a proof of concept, we selected genes targeted by chromothripsis (*TNFRS14* and *SPEN*) and simple chromosomal deletions (*TRAF3*) in the RT subclone of case 12 detected by WGS and analyzed their expression along the disease course using scRNA-seq. We observed a low expression of these genes as well as of *CDKN1B*, a hallmark of RT independent of genomic alterations (Cobo et al., 2002), in the dormant RT cells prior to their final expansion (Figure 106).

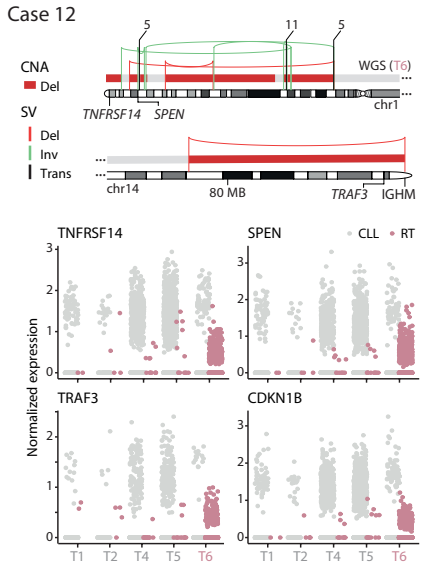


Figure 106. Expression of chromothripsis targeted genes at single cell level. Schematic representation of the chromothripsis (chr1) and deletion of TRAF3 (chr14) identified at RT in case 12 [top]. Expression of the targeted genes and CDKN1B in CLL and RT cells split by time point [bottom].

To further explore the presence of genomic rearrangements in early time points, we used scRNA-seq to identify CNAs involved in simple and complex structural variants found in RT subclones according to WGS. These analyses were performed on all cases with available scRNA-seq, except for case 365, which did not carry any new CNAs compared to the preceding CLL. Although the characterization of the copy number profile of single cells can miss small CNAs due to the small number of genes within the deleted/gained regions, it identified the main RT alterations in dormant RT cells prior to their final expansion up to the time of CLL diagnosis (Figure 107). These results link our genomic and transcriptomic findings and suggest an early acquisition of structural variants, including chromothriptic events, in RT, as reported in other tumor types (Maura, Bolli, et al., 2019).

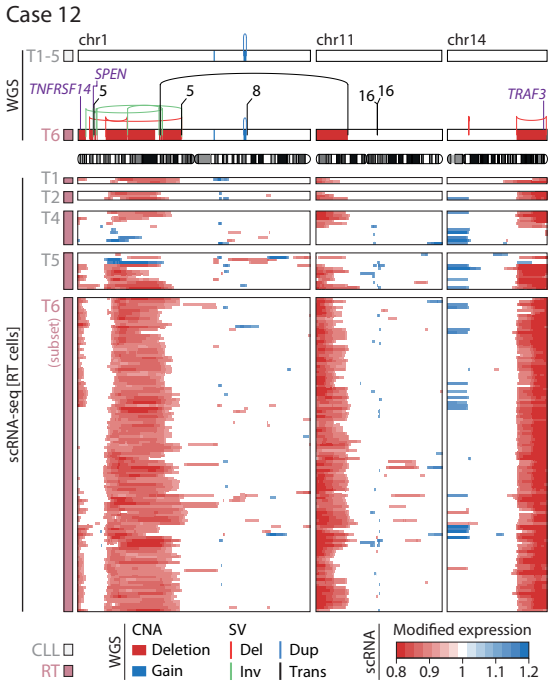


Figure 107. Identification of CNA involved in complex rearrangements in RT cells of case 12 by scRNA-seq. Chromosomal alterations detected by WGS in chromosomes 1, 11, and 14 in CLL and RT samples [top]. Copy number profile of RT cells detected at the different time points according to scRNA-seq. Only a subset of RT cells from time point 6 (time of RT) was included for illustrative purposes [bottom].

Overall, our integrative analyses uncovered a widespread early seeding of RT cells, up to 19 years before their clinical manifestation, which remain virtually dormant until they expand massively and become the dominant population at the time of transformation. This early presence of small subpopulations of the future dominant RT subclones was first inferred from WGS and subsequently confirmed by scDNA-seq, IG deep sequencing, and the gene expression and CNA profiles from scRNA-seq.

The OXPHOS^{high}-BCR^{low} transcriptional axis of RT

To elucidate the transcriptomic evolution and the epigenomic regulation from CLL to RT, we integrated genome-wide profiles of chromatin accessibility (ATAC-seq), chromatin activation (H3K27ac), and DNA methylation (performed by Beatriz Garcia and Dr. Martí Duran from Dr. José I. Martín's group at IDIBAPS) with bulk RNA-seq and scRNA-seq of multiple longitudinal samples from 6 cases, all treated with BCR inhibitors. The DNA methylome of RT mainly reflected the naïve- and memory-like B cell provenance of their CLL counterpart while chromatin activation and accessibility were remarkably different upon transformation (Figure 108).

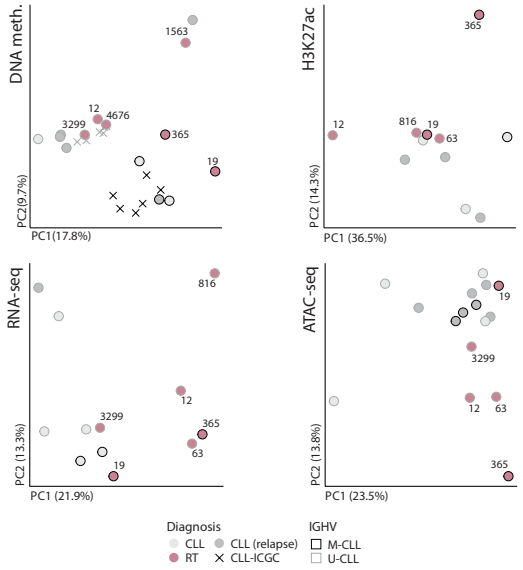


Figure 108. Bulk epigenomic and transcriptomic profile of RT and CLL. Principal components analysis of the bulk epigenetic and transcriptomic layers analyzed. RT samples are indicated in pink, while CLL samples are gray. Mutational status of the immunoglobulin genes is indicated by the border color.

We identified 150 regions with increased H3K27ac and 426 regions that gained accessibility in RT, as compared to normal counterparts and also to prior

CLL or CLL-relapse samples. These *de novo* active regions (i.e., chromatin accessible regions within the RT-specific active chromatin regions) were enriched in transcription factor (TF) families different from those known to modulate the epigenome of CLL (Beekman et al., 2018). Among them, 24 were enriched and upregulated in RT. The top TF was TEAD4, recently described to activate genes involved in oxidative phosphorylation (OXPHOS) through mTOR pathway (Chen et al., 2021) and to co-operate with MYCN in high-risk neuroblastoma (Rajbhandari et al., 2018). Additional TF were related to MYC (MAZ), proliferation/cell cycle (E2F family) or IRF family, among others (Figure 109). Intriguingly, high IRF4 levels seem to attenuate BCR signaling in CLL (Maffei et al., 2021) while they are necessary to induce MYC target genes, OXPHOS and glycolysis in activated normal B cells (D. G. Patterson et al., 2021).

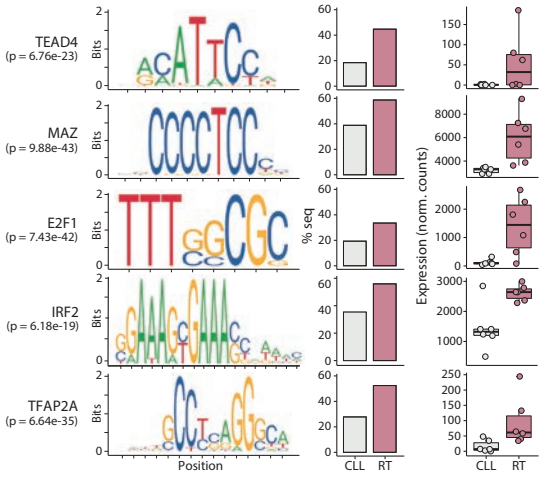


Figure 109. Transcription factors enrichment in RT. TF enriched within the ATAC peaks identified in the regions of increase H3K27ac in RT. The motif (left), percentage of RT-specific active regions and regions with increased H3K27ac in CLL that contained the motif (middle), and TF expression (bulk RNA-seq) in CLL and RT (right) are shown.

The bulk RNA-seq analysis revealed a distinct transcriptomic profile of CLL and RT. Two cases, 19 and 3299 (RT-PLL), had an intermediate expression profile

of RT (Figure 108 and Figure 110), which was also observed by scRNA-seq where they showed a low number of differentially expressed genes with no overlap with the other cases. Excluding these two cases due to their intermediate transcriptomic profile, we identified 2,248 differentially expressed genes (DEG) between RT and CLL (1,439 upregulated, 809 downregulated) (Figure 110). A considerable fraction of upregulated and downregulated genes overlapped with regions with the respective increase and decrease of H3K27ac (20%) and chromatin accessibility (16%) at RT.

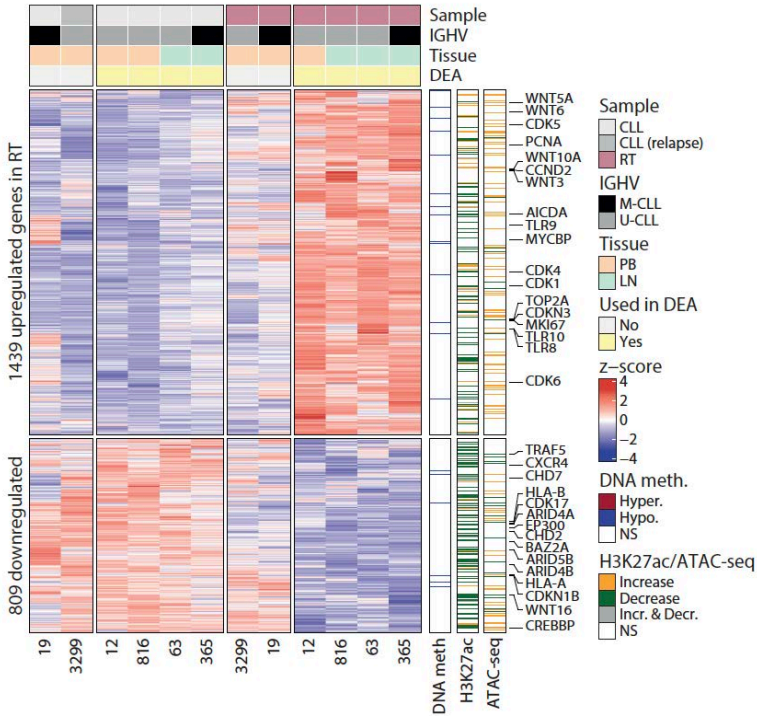


Figure 110. Transcriptome of RT by bulk RNA-seq. a. Heatmap showing the differentially expressed genes (DEG) between CLL and RT. Cases used in the differential expression analysis (DEA) are indicated. The overlap of DEG with DNA methylation changes, H3K27ac, and ATAC peaks is shown on the right. Selected genes are annotated.

Genes upregulated in RT were associated with pathways that seem independent of BCR signaling such as Wnt (WNT5A and others) (Hasan et al., 2021), Toll-like receptors (TLR9 among others) (Ntoufa et al., 2016), and a number of cyclin-dependent kinases. Downregulated genes involved, among others, CXCR4, HLA-A/B, and chromatin remodelers also targeted by genetic alterations (Figure 110 and Figure 111). Gene sets modulated by gene expression in RT were in line with the identified chromatin-based changes and included upregulation of E2F targets, G2M checkpoints, MYC targets, MTORC1, OXPHOS, mitochondrial translation, glycolysis, reactive oxygen species, and DNA repair, among others. In addition, RT showed a downmodulation of the BCR signaling (Figure 111).

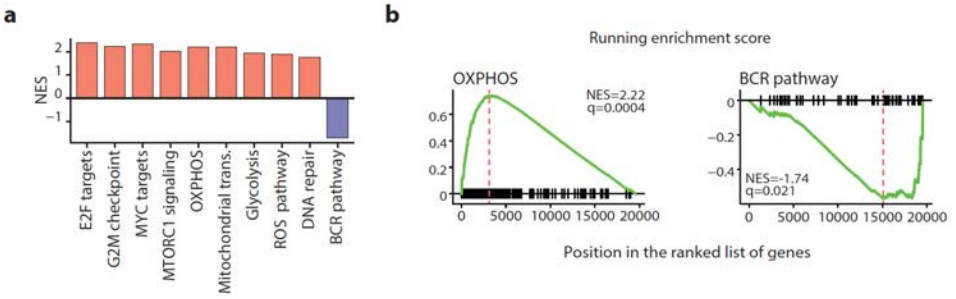


Figure 111. Differentially expressed gene sets in RT based on bulk RNA-seq. a. Summary of the main gene sets modulated in RT. b. Gene set enrichment plot for OXPHOS and BCR signaling.

The scRNA-seq analyses aligned with the bulk RNA-seq results. They confirmed the gene set enrichment analysis and allowed us to further refine the $OXPHOS^{high}$-BCR^{low} pattern recognized by bulk RNA-seq in RT. We recognized that 2/5 cases had $OXPHOS^{high}$-BCR^{low} (12 and 63, although the latter showed some intercluster variability), the 2 M-CLL cases carrying the $IGLV3\text{-}21^{R110}$ 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 (3299) was $OXPHOS^{low}$-BCR^{low}.

In addition, the scRNA-seq analysis showed that the OXPHOS/BCR profiles of RT were already identified in the dormant RT cells detected at diagnosis and at intermediate time points years before their final expansion and clinical manifestation, suggesting that they represent an intrinsic characteristic of RT cells rather than being derived from a de novo epigenetic/transcriptomic modulation upon treatment with BCR inhibitors (Figure 112).

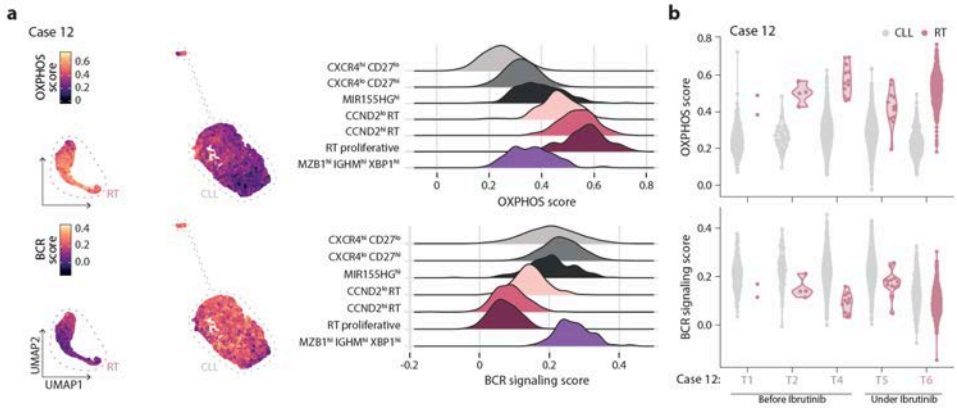


Figure 112. OXPHOS and BCR signaling of selected case 12 by scRNA-seq. a. OXPHOS and BCR signaling scores depicted at single cell level (all time points together). RT and CLL cells are highlighted [left]. Ridge plots show the OXPHOS and BCR score across clusters [right]. b. OXPHOS and BCR signaling scores of CLL and RT cells of case 12 across time points by scRNA-seq.

To expand these observations, we measured the expression of OXPHOS and BCR pathways in the scRNA-seq dataset of CLL and RT from Penter et al., 2021. The RT samples of case CLL9, which developed RT in the absence of any intervening therapy, showed a remarkably higher OXPHOS and slightly lower BCR expression compared to CLL (Figure 113).

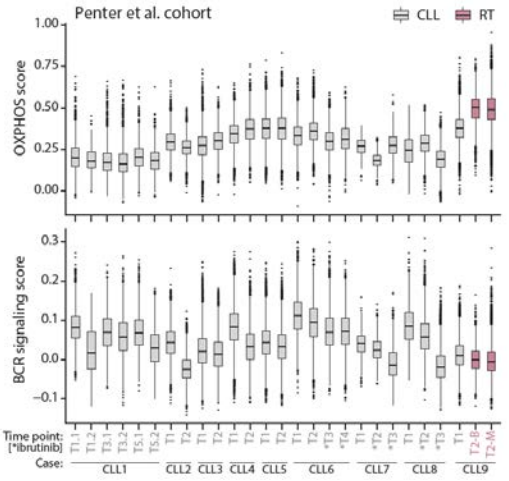


Figure 113. Confirmation analysis of the longitudinal scRNA-seq dataset from the study Penter et al., 2021. *Distribution of OXPHOS and BCR signaling scores at single cell level across the different time points of the nine cases included in the study of Penter et al. B, peripheral blood; M, bone marrow; *sample collected under treatment with ibrutinib.*

Overall, the epigenome and transcriptome of RT converge to an $OXPHOS^{high}$-BCR^{low} axis that is reminiscent of that observed in the de novo DLBCL subtype characterized by high OXPHOS and mitochondrial translation (DLBCL-OXPHOS), and insensitive to BCR inhibition (Caro et al., 2012; Monti, 2005; Norberg et al., 2017). Therefore, this axis might explain the selection and rapid expansion of small RT subclones under therapy with BCR inhibitors.

OXPHOS and BCR activity in RT

Next, we sought to validate the OXPHOS and BCR activity in RT previously described. These experiments were performed by Heribert Playa from Dr. Dolors Colomer's group at IDIBAPS and Pablo M. Garcia-Roves at IDIBELL. Although it was challenging to obtain suitable cryopreserved cells from paired CLL-RT samples, we could obtain enough material for three cases of the CLL-RT cohort (cases 19, 63, and 12) at different time points and with multiple technical replicates.

Independent cases were also considered, but the cryopreserved cells' quality and/or quantity did not meet the requirements for the experiments.

For the available cases, we measured the respiratory capacity and BCR signaling in CLL and RT to confirm our previous findings obtained by transcriptomic analyses and showed that the high OXPHOS levels identified in RT could potentially be used as a target for therapeutic intervention in the future (Figure 114).

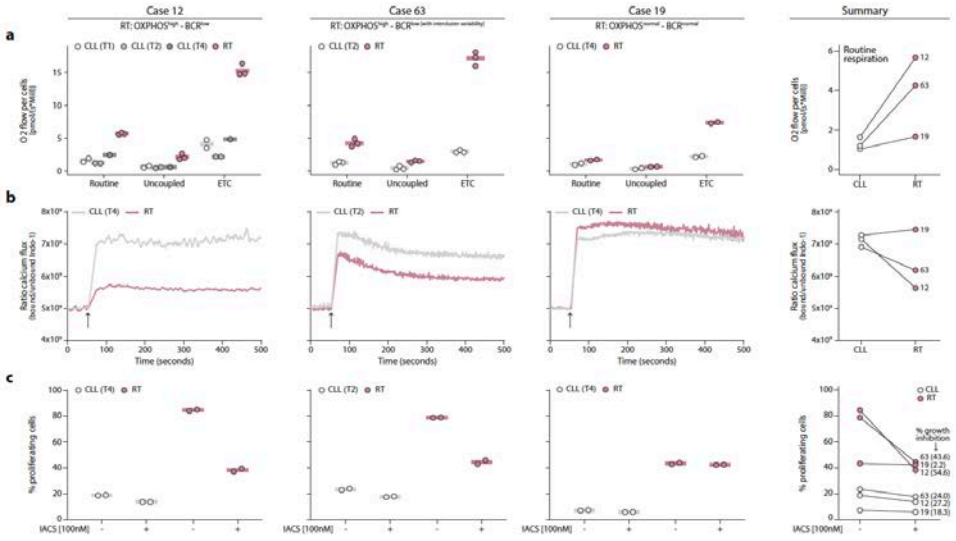


Figure 114. Cellular respiration, BCR signaling, and OXPHOS inhibition in RT cells. a. Oxygen (O2) consumption of intact CLL and RT cells of cases 12, 63, and 19 at routine respiration (routine), oligomycin-inhibited Leak respiration (uncoupled), and uncoupler-stimulated electron transfer system capacity (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 cases collapsed [right]. Richter samples are indicated in pink. b. Calcium kinetics of tumoral cells (CD19+, CD5+) upon stimulation with 4-hydroxytamoxifen (4-OHT) and anti-BCR antibody (black arrow). Basal calcium was adjusted at 5x109 Indo-1 ratio for 60 seconds prior cell stimulation with F(ab')2 anti-human IgM + H2O2 at 37ºC. Then, Ca2+ flux was recorded up to 500 seconds [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-hour incubation with or without IACS-

010759 (IACS) at 100nM. Percentage of proliferating cells was determined by 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-010759 (IACS) treatment after 72 hours. The normalized percentage of growth inhibition is indicated [right].

We first performed in vitro experiments using paired CLL and RT cells from cases 12, 19, and 63 to functionally validate the $OXPHOS^{high}$-BCR^{low} axis identified in RT, including functional evaluations of the oxygen consumption and BCR signaling in CLL and RT. Respirometry assays confirmed that $OXPHOS^{high}$ RT cells (cases 12 and 63) had a 3.5-fold higher oxygen consumption at routine respiration and 5-fold higher electron transfer system capacity (ETC) compared to CLL. In addition, $OXPHOS^{normal}$ RT (case 19) showed a routine oxygen consumption similar to CLL, although also had a relatively higher ETC than its CLL counterpart (Figure 114.a).

Using intracellular calcium flux analysis by flow cytometry, BCR signaling was measured by Ca^{2+} mobilization upon BCR stimulation with anti-BCR antibody (IgM) and showed that BCR^{low} RT cells (cases 12 and 63, the latter showing intercluster variability by scRNA-seq) had a lower Ca^{2+} flux compared to their respective CLL cells, which contrasted with the higher flux observed in BCR^{normal} RT cells (case 19) (Figure 114.b). Note that case 19 carried the $IGLV3\text{-}21^{R110}$ mutation, known to trigger autonomous BCR signaling (Minici et al., 2017), both in the CLL and RT .

To provide a functional validation that our findings could be therapeutically relevant, we focused on OXPHOS inhibition in primary CLL and RT cells. To determine the biological effect of $OXPHOS^{high}$ in RT, we performed in vitro cell growth assays using paired CLL and RT cells from cases 12, 19, and 63 with and without treatment with IACS-010759 (100nM), an OXPHOS inhibitor that targets mitochondrial complex I (Figure 114.c). These experiments showed that $OXPHOS^{high}$ RT (cases 12 and 63) had a higher proliferation at 72 hours compared

to OXPHOS^{normal} RT (case 19), and all of them higher than their respective CLL (Figure 114.c). OXPHOS inhibition in OXPHOS^{high} RT resulted in marked growth inhibition with an average of 49.1% decrease in cell proliferation, which contrasted with that observed in OXPHOS^{normal} RT (2.2% decrease) and CLL (23.2% decrease) (Figure 114.c). Note that we also confirmed that treatment with IACS-010759 (100nM) inhibited the cellular respiration of CLL and RT cells of the two cases analyzed (case 12 and 63; OXPHOS^{high} RT).

Overall, these results confirm the role of OXPHOS^{high} phenotype in the high proliferation of RT and suggest its potential therapeutic value in RT as has been proposed in other neoplasms (Caro et al., 2012; Molina et al., 2018; Norberg et al., 2017; Vangapandu et al., 2018; L. Zhang et al., 2019).

4.3.3 Study 5: Case report of a young adult with CLL harboring *ATM* germline variants

In this study we investigated a case report of a young adult with CLL. I performed and interpreted all WGS analyses, interpreted scDNA-seq analyses, and wrote the manuscript.

The first subsection is an introduction of what is known in the field in relation to our case report. Next, the results subsection includes the clinical course of the patient, which was written by Dra. Laura Magnano, and the 8-year genomic evolution under the influence of different therapies.

This work has been published in Blood Cancer Journal and is attached in the Appendix.

4.3.3.1 Introduction

Chronic lymphocytic leukemia (CLL) is the most common leukemia of adults in western countries, commonly diagnosed in the elderly with a median age of approximately 70 years. However, CLL can also be detected in adolescent and young adults (AYA). According to different studies, 0.85-3.7% of patients with CLL are diagnosed in AYA and 3% of these patients had a first-degree relative with CLL (Cherng et al., 2021). Families with multiple individuals affected with CLL and other related B-cell tumors have been described with contradictory findings regarding their potential early age at diagnosis (Goldin et al., 2009). Despite these observations, our knowledge about the molecular profile and predisposing factors in AYA CLL is scarce (Luskin et al., 2014; Nassereddine & Dunleavy, 2019).

The understanding of the biology of CLL has evolved significantly in recent decades. Comprehensive studies have dissected the genomic, epigenomic, and transcriptomic landscape of CLL (Puente et al., 2015). Approximately 9-18% of patients with CLL harbor del(11q) or Ataxia telangiectasia gene *(ATM)* disruption, which occurs in younger patients with bulky disease and poor survival (Döhner et al., 2000; Nadeu et al., 2016; Wierda et al., 2011). These deletions are frequently associated with germline and acquired mutations of *ATM* (Skowronska et al., 2012). *ATM* codes for the ATM protein kinase that participates in cell cycle, DNA repair, and apoptosis. Patients with the inherited disorder ataxia telangiectasia have biallelic alterations of the *ATM* gene and increased susceptibility to lymphoid malignancies (Reiman et al., 2011). Rare, protein-coding germline *ATM* variants are associated with CLL in adults (Tiao et al., 2017). However, *ATM* mutations are uncommon in familial CLL (Yuille et al., 2002).

Here, we describe an 18-year-old woman diagnosed with CLL whose family history included a younger brother with B-cell acute lymphoblastic leukemia (B-

ALL) and other family members carrying germline *ATM* mutations. A combination of whole genome and single cell characterization of this CLL at diagnosis and at additional time points during the course of the disease provided an opportunity to understand the genomic profile of AYA CLL and the sequence of events driving its evolution.

4.3.3.2 Results

Clinical course

An 18-year-old female was diagnosed with CLL, Binet-Rai stage AI, in the study of a lymphocytosis detected in a routine blood test at another institution. She had a past medical history of anxiety-depressive syndrome during childhood and chronic headache, but no neurological symptoms were reported. The patient had a younger brother diagnosed with B-ALL when he was 3 years old, who was in complete remission 13 years later, and an older sister with epilepsy. Her parents were both healthy. At the time of CLL diagnosis, the patient was asymptomatic with a normal physical exam. Her white blood cell count (WBC) was $9.08x10^9$/L, with 75% lymphocytes. Hemoglobin and platelet count were normal. Peripheral blood smear showed small atypical lymphocytes consistent with CLL, whose phenotype was $CD5^+$, $CD23^+$, $CD43^+$, $CD200^+$, $CD10^-$, CD20, and CD22 weakly positive with weak kappa light chain restriction. The fluorescence in situ hybridization (FISH) analysis for *ATM* (11q22), *D12Z3* (cen 12), *DLEU* (13q14.3), *LAMP1* (13q34), and *TP53* (17p13) were normal. One year after diagnosis, the patient received two cycles of rituximab plus fludarabine and cyclophosphamide (FCR) due to progressive disease, achieving a complete remission.

The patient was then referred to Hospital Clínic de Barcelona. Physical examination was normal without evidence of lymphadenopathy or splenomegaly. WBC count was $2x10^9$/L with 10% lymphocytes, hemoglobin 117 g/L, and normal

platelet count. Watchful waiting was recommended. Five years later, the CLL progressed with increased lymphocytosis, inguinal, axillary, and laterocervical lymphadenopathy (2-3 cm), and splenomegaly of 4 cm below the costal margin. At that time, the karyotype was 46,XX,del(13)(q12q21)[6]/46,XX[10] and a heterozygous del(13q14.3) was detected by FISH in 92% of nuclei. FISH for *ATM*, *D12Z3*, and *TP53* were normal and no *TP53* mutations were observed. The sequence of the IGHV genes showed a clonal rearrangement of the IGHV3-21 with 100% homology to the germline, not belonging to any major stereotype subset. Due to CLL progression, ibrutinib 420 mg per day was started and the patient achieved a partial response. However, after 20 months, ibrutinib had to be discontinued due to the severe diarrhea and acalabrutinib 100 mg every 12 hours was started. Progression of CLL was observed after 13 months of treatment and rituximab and venetoclax were initiated (Figure 45).

8 years of genomic evolution

The patient was included in the CLL program of the International Cancer Genome Consortium and the whole genomes of the normal and tumor sample at diagnosis were sequenced (Puente et al., 2015). No somatically-acquired driver alterations were detected but three germline *ATM* mutations were identified, including a pathogenic 28-base frameshift deletion (p.N3003Dfs*6) and two missense single nucleotide variants (p.K2204M and p.Y1961C). Although the p.K2204M missense variant has not been identified in previous studies, the p.Y1961C has been reported in a CLL patient and its modeling showed reduced ATM kinase activity (Barone et al., 2009). Based on this result, we studied the segregation of these mutations in the family members by Sanger sequencing. The mother harbored the frameshift deletion, while the father and the sister carried the two missense variants. Both the patient and her brother with B-ALL inherited all three variants (Figure 115). A milder ataxia telangiectasia phenotype, where

the disease progresses at a slower pace, has been observed in patients with reduced levels of ATM kinase activity (Stewart et al., 2001). At time of last follow-up the two siblings (28 and 16 years old) had not developed neurological symptoms.

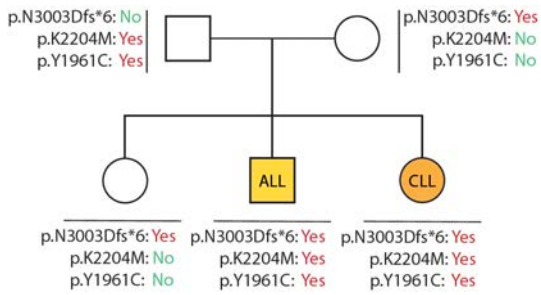


Figure 115. Pedigree tree of germline variants in ATM. The two missense variants carried by the mother and the frameshift variant from the father were inherited by the chronic lymphocytic leukemia (CLL) case studied and her brother that developed acute lymphoblastic leukemia (ALL)

To better unfold the contribution of somatic alterations during the evolution of the disease, whole-genome sequencing (WGS) was performed at 3 additional time points over a period of 8 years and complemented with single-cell DNA-sequencing (Figure 45 and see Methods - section 3.2.3). Using a longitudinal sample-aware mutation calling pipeline that increases sensitivity, we identified 689 genome-wide mutations, including 7 non-synonymous variants, in the WGS at diagnosis, which increased up to 1779 genome-wide mutations, including 18 non-synonymous, at the latest sample analyzed. Among them, four mutations were found in CLL driver genes over the course of the disease: *XPO1* (p.E571K), *SF3B1* (p.G742D), *MGA* (p.C1238G), and *POT1* (p.C44S). The mutations in *XPO1* and *SF3B1* were already present at diagnosis but were missed in our previous study (Puente et al., 2015) due to their very low frequencies. After 4 years (time point 2), their clonal size expanded, and the remaining two driver mutations in

MGA and *POT1* were detected. Regarding structural alterations, only del(13q) was clonally detected at the second time point and onwards (Figure 116).

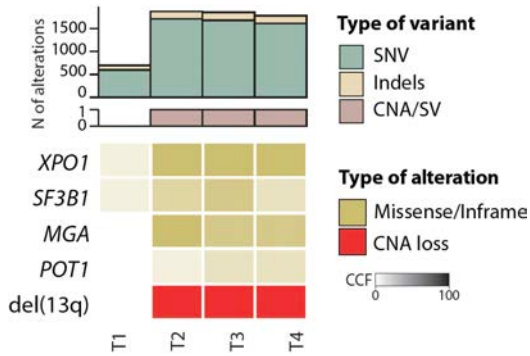


Figure 116. Somatic mutations identified during the disease course. The upper barplots show the number of mutations [single nucleotide variants (SNVs) and short insertions and deletions (indels)] and copy number alterations (CNAs) or structural variants (SVs) at each time point. The lower oncoprint shows the driver alterations; the transparency of the color is proportional to their cancer cell fraction (CCF).

Somatic driver alterations were present at different allele frequencies through the disease course, suggesting an ongoing clonal evolution driving the pre- and post-treatment progression of the disease. To dissect the underlying clonal evolution, we reconstructed the subclonal evolution and explored the mutational processes active during the CLL course (Figure 117). This analysis revealed a branching pattern of evolution in which the founding CLL clone did not carry any recognized driver alteration beyond the *ATM* germline variants. Additionally, two minor subclones were already present at diagnosis: subclone #3 carrying del(13q), *XPO1,* and *MGA*, and subclone #4 which originated from subclone #3 and acquired the *SF3B1* mutation (Figure 117.a). These lineage trajectories are in line with previous literature in which *ATM* loss preceded del(13q) in a familial CLL study (Kostopoulos et al., 2015) and with a recently described combinatorial effect of *ATM* loss and *SF3B1* mutation (Yin et al., 2019).

Intriguingly, these small subclones at diagnosis expanded after treatment with FCR, that, on the other hand, reduced or eliminated the initial subclones #1 and #2, with no additional CLL drivers, suggesting that decreased competition allowed the expansion of subclones carrying potent drivers. Of note, subclone #4 carrying the *SF3B1* mutation represented the largest subpopulation of cells at relapse post-treatment with FCR (T2), in line with the poor prognosis of *SF3B1* mutated cases under FCR therapy (Stilgenbauer et al., 2014). Nonetheless, this subclone slightly diminished at time point 3 and was virtually eradicated at time point 4 after treatment with ibrutinib, which is in line with the higher sensitivity of *SF3B1* mutated CLL cells to BCR inhibition in vitro (Yin et al., 2019). Additional diversification was observed in subclone #3 at T2 which led to the emergence of subclone #6 harboring the *POT1* mutation. This subclone expanded under ibrutinib treatment and accounted for 54% at the last time point analyzed 3 years after its detection (Figure 117.a).

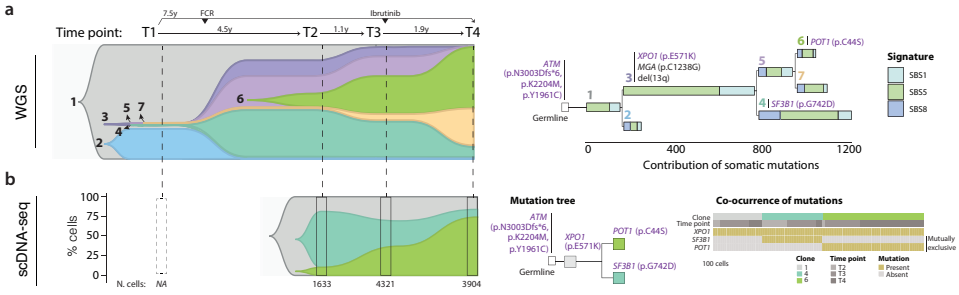


Figure 117. Evolutionary trajectories along the course of the disease. a. The fishplot [left] depicts the subclonal architecture and clonal dynamics inferred from WGS. Each vertical line represents a time point analyzed. Each subclone is painted in a different color, and its height is proportional to the CCF at each time point. The upper-right tree shows the phylogeny of the tumor cell subpopulations, the length of the branches is proportional to the number of acquired SNVs, and they are colored by contribution of mutational signatures identified in CLL [right]. The clock-like signatures SBS1 and SBS5 contributed most of the mutations acquired. b. The fishplot (left) shows the clonal dynamics measured by single cell analysis. For each available time point, the integrated barplot shows the proportion of cells harboring each specific combination of alterations in the driver genes illustrated

on the "Mutation tree" (middle). The total number of analyzed cells at each analyzed sample is shown at the bottom. The "Co-occurrence of mutations" plot (right) indicates the presence or absence of mutations in each cell. For illustrative purposes, cells have been merged in bins of 100.

To confirm these evolutionary trajectories, we performed single-cell DNA-sequencing of 32 CLL driver genes and identified the reported mutations in *XPO1*, *SF3B1,* and *POT1* [note that *MGA* was not included in the commercial gene panel used]. This single-cell analysis confirmed the timing of acquisition of these driver mutations and the clonal dynamics inferred from WGS (Figure 117.b).

Finally, we explored the mutational processes active during the CLL course. We roughly observed the same mutational signatures in all subclones and identified signatures SBS1 and SBS5, which are related to cell division and found in all cancer types as well as non-tumor cells (Alexandrov et al., 2015), as the responsible for most of the mutations acquired (Figure 117.a).

Here we have reported the 8-year genomic evolution of a CLL diagnosed in a young patient that inherited three *ATM* variants, two of them previously reported to inactivate or reduce ATM activity (Barone et al., 2009), which represented the only recognized driver events in the founding CLL clone. This might suggest that ATM inactivation might be a genomic factor contributing to CLL initiation, while tumor evolution and disease progression was dictated by the acquisition of secondary driver alterations, which could be detected in small subclones years before their expansion, and by different types of treatment that influenced subsequent clonal dynamics. Altogether, the lack of somatically-acquired genetic driver alterations in the founding CLL clone of this patient emphasizes the need to study the germline as well as non-genetic aspects of the tumors to further understand the mechanisms leading to CLL.

5 Discussion

Over the past decades, next-generation sequencing has boosted biomedical research at an unprecedented scale. It has changed the way we tackle biological questions and led to remarkable scientific discoveries. Not only have NGS technologies become faster and more affordable but also have advanced to produce longer DNA reads and reached higher resolution up to the level of single cells. Along the way, numerous challenges and open questions have come up: 1. The computational needs to sustain these studies have grown together with the accelerated production of data. Besides, sophisticated bioinformatics methodologies are needed to perform complex and demanding analyses. 2. Upon mastery and understanding of the technologies, the point of interest is moved to the exploitation of the data and to the extraction of biologically and clinically meaningful information that can expand our knowledge. 3. Finally, to obtain real benefit from these insights, they must reach the lives of citizens to improve them with better disease prevention and management, and by contributing to the promise of personalized medicine. During this thesis we have dealt with these topics at different levels. Starting from the infrastructure level, going through the methodological aspects, and finally applying them to biomedical studies, while seeing and handling constraints and challenges that hamper current cancer research first-hand.

Large-scale genomic analyses depend upon infrastructures that go towards distributed cloud-based solutions

Within the PCAWG project (2014-2020), we reanalyzed more than 2800 normal-tumor pairs of whole genomes which was the largest set of whole genomes at the time. Thus, the first need of the project was the coordination of multiple computational institutions that could provide the resources to perform such analyses, as well as the storage to manage vast amounts of data adding up to 1PB. As in any other project that aims to integrate datasets from different

sources, data had to be organized and harmonized. In our case, each data-contributing project had to reformat their sequencing data with standardized metadata and submit it to their regional data center(s) that would later conduct the execution of the project's workflows. Overall, the PCAWG initiative utilized 14 computing clouds and HPC centers over 2.5 years.

During this long period of time, data was being submitted, pipelines were being developed, and, at the same time, analyses were being performed. This overlap in time of development and execution carried some hindrances and adjustments had to be made after many already completed executions (e.g., applying filtering strategies that were not included in the first version of the variant calling workflows). It also caused delays on the analyses as workflows were not yet developed or properly tested. Nonetheless, the project advanced and coordinated a massive and distributed analysis, where virtualization was used as a means of reproducible and portable software.

There are numerous efforts towards reproducible research, which includes, of course, the software used within the studies. Complex pipelines that contain multiple tools with dependencies among them are not easy to reproduce and the community is increasingly promoting the use of packaging solutions that can be easily run in any environment. The development of very well defined, standardized, and containerized state-of-the art portable pipelines is burdened by the rapidly evolving field of bioinformatics, where new tools are often published, and new versions are released within months or even weeks. So probably a trade-off between the effort of building such a pipeline and the utility of the product should also be considered.

Within the PCAWG project, we made a huge effort towards this direction. However, there were many steps applied after the basic variant calling, which also

included manual filtering and revisions, some of the tools could not be packaged into standard docker containers and, overall, it would be hardly impossible to reproduce all these analyses as they were done during the project. Moreover, the reference genome version that was used has long been replaced by the newest version. So, nowadays, if one were to perform a similar analysis, it would be more convenient to use newer versions of the programs and migrate to the newest version of the reference genome. Indeed, this is what we are doing within the ICGC-ARGO project, where pipelines are being designed from scratch, although on top of all the technological advances that started within the PCAWG project and of the tremendous experience that everyone gained from this endeavor.

Going back to the development of the project, its implementation in the BSC's HPC premises was carried out in Study 1. In particular, in our center, we dealt with the requirements of large-scale executions and found ways to accommodate these mainly cloud-based solutions into our HPC infrastructures. At the beginning of the project, technologies like Docker and the usage of cloud environments were growing in popularity and were starting to be introduced in the genomics area. Cancer genomics, and genomics in general, produces tons of data, and can benefit from portable workflows that can be executed right where the data is located, rather than having to download terabytes or, potentially, petabytes of data. This solution entails a convenient bypass for the bottleneck of data transferring and/or the lack of storage capacity in some centers. The project was conceived in this way and cloud infrastructures with residing data were made available to the working groups who carried out the downstream analyses.

During the development phase, each computational center performed the assigned executions and synchronized the results with the other data centers. The BSC was part of this process, but we had additional hiccups as our pure HPC center at the time did not allow most of the used technologies right away. After study

with the Operations team at the BSC, we could break some barriers and implement exceptional ways to carry out these non-HPC executions. This project was a special case and the BSC allowed us privileged operations that are usually not granted. While our HPC center started to realize the need of virtualization, especially within the Life Sciences department, among others, the technologies also evolved to be more HPC-friendly and more secure containerization systems were starting to evolve, such as Singularity. Now, we can freely use Singularity at the BSC which does not set us aside from the advances and common practices in Life Sciences.

On top of these technological efforts, researchers exploited the data generated during the project and over 20 papers were published in *Nature* and affiliated journals. Despite these scientific discoveries, which took advantage of the integrative PCAWG dataset and very well described the genomics of all included cancer types in terms of driver alterations, mutational signatures, or tumor heterogeneity, among others, associations with clinical data could not be thoroughly exploited. The project included a minimal set of clinical data, but this restricted information does not have enough power to explore the full potential of the PCAWG dataset. For instance, associations of biomarkers with particular characteristics of patients only present in their whole clinical history, which was not available, could not be assessed. To fully profit from such large-scale datasets and bring translational research closer to real clinical applications, phenotypical and clinical information needs to come with molecular data, as it is the only way we can associate genetic alterations with the clinical outcomes of patients. Opportunely, this is the main idea behind the ICGC-ARGO initiative, where a resource of more than 100,000 cancer patients will integrate molecular and high-quality clinical annotations to find answers to key clinical questions.

The identification of variants still remains a challenge

Most of the genomes included in PCAWG had already been analyzed by the groups who generated the data. However, each dataset was subjected to different workflows to identify somatic mutations and their integration without any biases reflecting the tools used was not possible. Variant calling is still a challenge and, although there is a vast variety of programs, their results are heterogeneous, which makes them not comparable and complicates their integration. Thus, harmonization of the results from independent workflows that can overcome methodological differences and faithfully distinguish biological signals from noise and technical biases is not straightforward.

The answer to this heterogeneity has often been the reanalysis of all data using homogeneous pipelines, which requires additional and redundant computational resources (Campbell et al., 2020; Ellrott et al., 2018). Besides this expensive solution, one could also think of a way to harmonize variant calling results from different programs by assessing their performance on selected benchmarking datasets and evaluating if they can really be integrated and how. In line with this, within the EUCANCan project, we are currently working on protocols to assess the performance of variant calling programs and pipelines as well as the degree of compatibility among them. The assessment of integrability of different datasets analyzed using distinct pipelines would spare the need of many computational resources.

Discordance among different variant callers not only complicates data integration from different studies, but also the selection of tools for the design of a pipeline. There is a need for bioinformatics expertise to correctly interpret the results and it is also important to know the biology behind the data. For instance, in the case of CLL, as it is a type of blood cancer, when the germline samples are

extracted from blood, although they are purified, they might have remains of tumor contamination. This particularity requires adjustments to the variant calling settings because many programs do not allow any reads supporting somatic variants in the normal sample by default.

The lack of well characterized benchmarking datasets for somatic variants, and the demanding characteristics of some tumor samples, such as low purity, FFPE archival material, lack of matched-normal sample, impedes the selection of one single strategy that can be blindly trusted. Synthetic sequence data have been used to benchmark variant calling algorithms, but it is very difficult, if not impossible, to perfectly simulate the sequencing protocols and especially the complexity and heterogeneity of a tumor sample. Next to the pure simulated data, where reads are simulated from scratch, there was another approach that actually used real sequenced data to introduce somatic alterations (Ewing et al., 2015; Lee et al., 2018). This combining strategy is closer to reality, as it uses real reads, in contrast to simulating them artificially. In any case, simulated mutations might not follow the true molecular processes that generate real somatic mutations in the first place, which is another caveat for synthetic data. Cell lines have also been used with the intent to provide good benchmarking datasets that can be openly available (Espejo Valle-Inclan et al., 2022; Shand et al., 2020; Tai Fang et al., 2021). Finally, orthogonal validations from real data can provide the most realistic settings but have other limitations. Not only are we restricted to the number of validated variants, but it is also very difficult to be 100% certain of the validation of some variants. However, identifying a variant in two different experiments can indeed increase our confidence that it will most probably be a “true” variant. Similarly, although not seeing the variant in the second experiment might not always mean that the variant is not there, if we can confirm the good quality of the validation, we could consider it a “false” variant for variant calling evaluation.

Taking all this into account, we have carried out various efforts on procedures for variant calling interpretation throughout this thesis, which can serve as guidelines for future evaluations and for assessing performance on other datasets (Study 2 and Study 3). To compensate for the lack of benchmarking datasets for somatic variant calling, we also used orthogonal validation, which is a resourceful approach, although it is limited to the genomic loci used in the other technique. Similarly, previously validated variants can also be used but might be biased to regions of interest like driver genes, or to the programs that were used at the time. On top of this, most of the benchmarking efforts are done on good quality material, while some clinical samples are not of best quality. From archival FFPE blocks to cases without germline samples, these usually unique samples are not fit to the general variant calling strategies. However, the particularities of some patients or some cancer types and their scarcity make these samples very precious, and worth trying.

We have recognized that quality control provides good guidance on the reliability of the prospective variant calling results, particularly in these borderline samples. Together with quality control, the bioinformatics expertise and knowledge of the tools gained during this thesis has been an asset as it can give hints on how good your data and results might be. Familiarity with both the methods and the biology of the samples analyzed has aided us to grasp the validity of the results at first glance and spot potential caveats. Numerous checkpoints can be used to reaffirm or invalidate the variant calling results: the expected number of mutations, the mutational profile in terms of mutational signatures, the allele frequency of the variants, and the detection of known drivers, among others.

There is a large variety of complementary variant calling tools. Usually, they are designed for one or two specific types of variants. For instance, SNVs and

indels are commonly identified together by the same program, but identification of CNAs or SVs requires a totally different strategy and is implemented by other programs. Analysis of whole genomes usually involves the selection of several methods to cover all types of variants: SNVs, indels, SVs, and CNAs. And not only that, even within one same type of variant, it is well known that methods produce discrepant results (Alioto et al., 2015). Each algorithm can have its own strengths and weaknesses: one method might be very good at detecting low frequency variants, but at the same time it might call a high number of false positives, while other methods are more conservative and have a lower false positive rate, at the price of missing the most challenging variants. To find a balance between them, it is a common approach to use several methods and create a consensus among them (Campbell et al., 2020; Ellrott et al., 2018), which basically follows the wisdom of crowd principle (Costello & Stolovitzky, 2013).

As previously reported (Alioto et al., 2015; S. Y. Kim et al., 2014; M. Wang et al., 2020), we have also observed that this strategy can improve the overall performance of individual variant callers. Moreover, it might have less variability due to peculiarities of the tumor samples. We have experienced individual methods that can go awry in lower quality samples, while the consensus among different tools might aid to counter this effect and might achieve decent results. Overall, the numerous evaluations on cancer genome and exome sequencing data that we performed during the thesis, allowed us to understand the differences and common limitations of variant callers and prepared me for their application and interpretation of their results in real studies.

Biological research translation into the clinics and data sharing challenges

Despite technical and methodological challenges, NGS has certainly led to unsurpassed discoveries, especially in cancer research (Shyr & Liu, 2013). The next

step is to put this knowledge in the hands of clinicians and health care systems so that it can really benefit our society. Our growing understanding of the cancer genome, including driver events, mutational processes, intratumor heterogeneity, and evolutionary trajectories, can have an impact on drug development, prognostication, treatment selection, and improved patient management. Precision medicine is supported and guided by our understanding of the cancer genomic landscape to provide better diagnosis, treatments, and aspires to get better outcomes for each individual patient. The introduction of genomic analysis in clinical settings is having the highest impact in rare diseases and oncology (Schilsky, 2014). In the latter, clinicians have long known that each patient's cancer is unique. Thus, tumor molecular profiling has the potential to place personalized cancer care over conventional non-personalized approaches (Schwaederle et al., 2015). Next-generation sequencing is increasingly used in the clinical setting, motivated by the growth of molecular target therapies, which rely on the genetic variants identified in the patient samples.

The promise of personalized medicine builds upon the integration of genomic information, clinical data, and patient preferences to provide the best treatment possible for each patient's unique cancer. It has been seen that patients who have actionable molecular alterations (i.e., with strong evidence of benefit from a specific therapy) and who receive a matched therapy had significantly longer median overall survival (Pishvaian et al., 2020). However, molecularly guided treatments are not always possible. Not all patients harbor actionable variants, and the overall success of these therapies depend on the knowledge of actionable mutations in cancer. Moreover, even patients with sequencing-matched therapies might not have access to them and struggle with restrictions to access clinical trials, due to previous treatments or comorbidities, and, in some countries, insurance coverage limitations (Morash et al., 2018).

The MedPerCan project (Study 2) envisioned the potential of genomic analysis and its translation into the clinics. It had two very well-defined goals aimed to guide the implementation of personalized medicine in oncology in Catalonia. First, a multidisciplinary and multi-institutional circuit was to be established including the main actors: hospitals, sequencing centers and computational infrastructures for data analysis. This platform would integrate the outcomes of genomic analyses, potentially run at large scale, and relevant clinical information. Next, the impact of clinical decision-making based on this knowledge base was to be evaluated. Although this initiative was opportune in time and scope, it never fulfilled its ultimate goal. One of its main weaknesses was the reduced shareability of clinical data, which was limited to very few basic metadata fields, which were insufficient to fully test the procedures. This is a recurrent issue faced by many similar projects, as we are not yet culturally prepared to share private clinical data even if it is among accredited partners and in protected infrastructures.

Limited data sharing, especially in relation to clinical information, obstructs data reuse and scientific advances as genomic data by itself cannot be used to identify and correlate genetic alterations to clinical responses. Reusability, together with findability, accessibility, and interoperability, the so-called FAIR principles, are all key topics addressed by a mass of projects worldwide. They aspire to provide the bases and guidelines, in terms of infrastructure, software applications, and data sharing, to enable access and exploitation of data and to foster research and its translation into the clinics. However, despite these efforts, even if we are well-advanced technologically, and even if we are pushing for the sake of patients and well-being of citizens, social and legal issues are not resolved, and it is a long hard road.

Large-scale initiatives have proved the value of assembling and integrating large biological datasets leading to new insights into the molecular basis of cancer (Campbell et al., 2020a; Weinstein et al., 2013). This wealth of information presents the opportunity to strengthen our findings and most importantly its translation into clinical care. Using these data to its full extent will require cross-national cooperation, including procedures to pass over the ethical and legal barriers across borders and harmonization that allows interoperability among datasets from different sources and data types.

Many projects are working towards this direction. At the European level, among others, the 1+ Million Genomes initiative aims to enable genomic and clinical data secure access across Europe, and EOSC4Cancer has the goal to foster the exploitation of cancer data providing the infrastructure, protocols, and guidelines to organize and integrate cancer-related data resources to foster cancer research and its subsequent benefits for EU citizens. Altogether, these initiatives work in parallel to bring together technologies and data pieces that have the potential to enhance scientific discoveries and, most importantly, optimize our health systems with personalized medicine protocols that can improve patient management and outcome.

Meanwhile, local studies at smaller scales are pushing to expand our understanding of the molecular basis of cancer origin, progression, and response to treatment. In this direction, we have worked together with Dr. Elías Campo from IDIBAPS/Hospital Clínic to further elucidate the molecular basis of CLL and its evolution during the disease course and to its most aggressive form, the Richter transformation (Study 4 and Study 5). As in other cancer types, NGS has greatly contributed to sequencing studies that taught us the molecular attributes of CLL. Early studies that collectively identified recurrent genetic alterations,

transcriptional alterations, and epigenetic changes allowed us to see the tremendous heterogeneity of this neoplasm (Gruber & Wu, 2014).

The mechanisms behind Richter transformation in CLL are elusive and have unmet clinical needs

We now know that CLL comprises subpopulations of tumor cells, or subclones, rather than a monolithic population. This intra-tumor heterogeneity becomes quite problematic in the presence of treatment, because a small subpopulation can become the dominant clone upon relapse, leading to a more aggressive form of the disease. In this regard, cancer evolution is a central obstacle to curative therapy and, what is worse, the selective pressures of our treatments might accelerate this evolutionary process. More often than not, what becomes the clonal population after treatment can be traced back to a small subpopulation present (long) before treatment initiation (Burger et al., 2016; Guièze et al., 2019; Landau et al., 2015, 2017). This inherent capacity for evolution of the pre-treatment diversity is one of the major bottlenecks of treatment success and has been seen in virtually all cancer types. In the case of CLL, the most aggressive form of evolution occurs in a small percentage of patients, whose CLL transforms into a high-grade lymphoma. This complication, called Richter transformation (RT), is associated with dismal clinical outcomes and with unmet clinical needs.

We have used this paradigmatic form of cancer evolution as a model to study tumor evolution, which is hampered by the analysis of bulk tumor samples at low coverages and single or limited number of spatial or sequential samples (Gerstung et al., 2020; Griffith et al., 2015; Vendramin et al., 2021). A better knowledge of these evolutionary trajectories can contribute to our understanding of the role of therapy as a driver of clonal diversification and selection, and might translate into

more effective clinical protocols and anticipation-based treatment strategies (Ferrando & López-Otín, 2017).

Previous studies have identified recurrent alterations and risk factors of RT (Chigrinova et al., 2013; Fabbri et al., 2013; Klintman et al., 2021; Rossi et al., 2011; Scandurra et al., 2010), but the mechanisms underlying this transformation are not fully understood, and available material to study this neoplasm is scarce. In Study 3, we could gather up to 19 cases with available fresh frozen material to conduct WGS and further analyses, including transcriptomic techniques and epigenetic experiments, in a subset of cases. Furthermore, we have investigated our results at both bulk and single cell resolution. Overall, the project involved the work of many people from different research groups and, altogether, we could reveal the spectrum of mutational processes, genomic and epigenomic alterations, subclonal composition, and temporal dynamics of this transformation under different treatments.

We have found that the genome of RT is characterized by a remarkable structural complexity, often including single-hit catastrophic events like chromothripsis or chromoplexy that can target multiple driver genes. As previously described, we identified driver alterations in cell cycle, MYC, NOTCH, and NF-κB pathways (Chigrinova et al., 2013; Fabbri et al., 2013; Rossi et al., 2011). We have also recognized that it carries the imprint of early-in-time, treatment-related mutational processes, such as the novel SBS-RT potentially associated with bendamustine and chlorambucil exposure. The detection of previous treatment footprints at the time of transformation supports the model of single cell expansion, where a pre-existing cell exposed to a mutagenic therapy can carry its imprint, which is only detectable after its expansion by bulk sequencing (Pich et al., 2019; Rustad et al., 2020).

Indeed, the reconstruction of the subclonal composition and dynamics from WGS identified the presence of minute RT subclones up to the time of CLL diagnosis. At first, we were not confident enough of these striking results that seemed to point to the existence of potent RT seeds dormant for up to 19 years before they were triggered, somehow leading to the overt RT manifestation. As a curiosity, at the beginning of the study, we actually applied more stringent criteria that did not identify these RT seeds so early in time. However, as the project advanced, we applied other techniques that further confirmed our initial finding and that supported the idea of a very early diversification of CLL leading to fully-assembled RT-cells in terms of genomic, immunogenetic, and transcriptomic profiles, already at CLL diagnosis before the clonal expansion associated with the clinical transformation 6-19 years later.

At the genomic level, we identified the driver alterations in each subclone and validated their composition and evolution by single-cell DNA sequencing using a gene panel of 32 CLL driver genes. This first technique already confirmed the evolution inferred from bulk WGS and recognized the presence of a small percentage of cells already at the time of CLL diagnosis carrying the RT driver alterations. Immunogenetic analyses by deep sequencing of the immunoglobulin (IG) genes also identified the early presence of IG RT-specific mutations, which were previously detected in the WGS samples at the time of transformation. This early seeding of RT subclones is aligned with previous studies finding the presence of resistant subclones before treatment initiation and mathematical models timing the acquisition of driver events years before diagnosis (Gerstung et al., 2020; Landau et al., 2017; Sentís et al., 2020). Likewise, an early immunogenetic diversification after the leukemic transformation has also been described (Bagnara et al., 2021; Gemenetzi et al., 2021).

Beyond the mutational landscape, we also asked ourselves if these early seeds of RT could have other RT-like features and explored the DNA methylation, chromatic accessibility and activation, and the transcriptional profile of RT. There were remarkable changes in chromatin configuration and transcriptional programming. We observed overexpression of cell cycle regulators, Toll-like receptors, Wnt, MYC, MTORC1, and OXPHOS related transcripts, and downregulation of the B-cell receptor signaling pathway that might be compensated by the activation of Toll-like, MYC, and MAPK pathways (Chakraborty et al., 2021; Dadashian et al., 2019; Ntoufa et al., 2016; Varano et al., 2017). Of note, the upregulation of OXPHOS and downregulation of BCR pathways defined an $OXPHOS^{high}$-BCR^{low} axis characteristic of RT which reminded the de novo DLBCL-OXPHOS subset, which is insensitive to inhibitors of BCR signaling (Caro et al., 2012). The rapid expansion of RT subclones under BKT inhibitor treatments is in line with its low BCR signaling, except for the cases carrying the IGLV3-21^{R110} mutation leading to autonomous BCR activation (Minici et al., 2017), the increased number of subclones carrying unproductive IG genes, and the development of RT with plasmablastic differentiation, a cell type independent of BCR signaling (Chan et al., 2017).

Using scRNA sequencing, we identified this reprogrammed transcriptomic profile in small RT subclones years before their clinical manifestation and up to the time of CLL diagnosis, confirming our WGS findings. Furthermore, the link between the previous genomic landscape and the transcriptomic program is further confirmed by the presence of RT structural changes identified in single cells and inferred from scRNA-seq.

Altogether, we confirmed the presence of fully assembled RT-cells that can be dormant for many years although carrying potent driving forces. The very early emergence of these subclones driving the late stages of cancer evolution might

set the basis for future single-cell-based predictive strategies able to identify these lethal seeds before their final expansion. Finally, we also uncovered a potential vulnerability of RT that could be exploited therapeutically. OXPHOS inhibition revealed a remarkable cell growth inhibition of RT cells in vitro (Molina et al., 2018; L. Zhang et al., 2019), a finding worth exploring in future treatment options.

Despite numerous novel targeted therapies currently available, RT remains the biggest therapeutic challenge in CLL. The combined discovery of RT early seeds and RT-specific therapeutic targets might provide an opportunity for early intervention to eradicate dormant RT subclones and prevent their future expansion leading to this lethal transformation of CLL.

The limited number of patients of this study might be one of its weakest points. Especially in heterogeneous cancers like CLL or RT, many cases are needed to identify recurrence and commonalities among patients. Due to the rarity of RT and the low number of samples with good quality that are available, only a handful of our cases had the complete set of omics analyses. Nonetheless, we were able to unveil novel genomic drivers and epigenomic and transcriptomic reconfigurations, very early emergence of RT seeds, and potential treatment options targeting the OXPHOS pathway that can be further explored in other cases and more studies to come. Indeed, studies like this would greatly benefit from data sharing, where larger cohorts could be gathered. Usually, genomic data is shared with minimal clinical metadata, if any, but it is obvious that we need more than that, as precise and extended clinical information is essential for comprehensive translational research. For instance, in our study, we would not have been able to establish the relation between SBS-RT and bendamustine or chlorambucil if we hadn't had the whole clinical history of the patients, including all the treatments during the course of the disease. If partial information was

given, i.e., sharing only the last treatment prior to RT, it would not have been enough.

Complex interactions between genomic and epigenomic alterations, the microenvironment, and treatment pressures can determine the disease course. In the RT study (Study 3), we have seen how the integration of different layers of omics data, including genomics, epigenomics, and transcriptomics information, can give a broader view of the molecular processes underlying this transformation. Together with this, the complete clinical history allowed us to recognize that the novel mutational signature SBS-RT might be related to treatments that the patients received during the CLL stage, and how BTK inhibitors might favor the selection of subclones that do not rely on the BCR signaling, as shown by the expansion of subclones carrying unproductive BCR. Altogether, we found that the dynamics of these tumors seem to be driven by the selection of subclones from the pre-existing subclonal diversity, rather than the emergence of new subclones.

Genomic characterization spanning 8 years of disease course of a young adult with CLL

In line with this, in the case report of a young adult with CLL (Study 4), where we analyzed 4 time points along 8 years of the disease course, we identified minor subclones at the time of CLL diagnosis, indicating again a pre-existing diversification that can dictate later clonal dynamics. Upon treatment pressures, we recognized how these subclones expanded or shrank. In particular, the subclone carrying a *SF3B1* mutation, which confers poor prognosis under FCR therapy (Stilgenbauer et al., 2014), represented the largest subclone at relapse post-treatment with FCR therapy, and slightly diminished after ibrutinib treatment, in line with the higher sensitivity of *SF3B1* mutated CLL cells to BCR

inhibition in vitro (Yin et al., 2019). However, despite the somatic mutations identified during this evolution, we did not recognize any somatic driver alteration in the CLL founding clone, which highlights the importance of exploring germline variation, as well as other non-genomic aspects. Indeed, our case carried three *ATM* germline variants, two of them reported to inactivate or reduce *ATM* activity (Barone et al., 2009). These mutations could have a driving role in CLL initiation, as they are the only driver alterations identified in this patient in the CLL founding clone, and they are also carried by the younger brother who also developed another neoplasm when he was 3 years old.

The relevance of technological and methodological aspects in biomedical studies and their clinical application

These studies allowed me not only to put the previously established strategies for tumor genome analyses into use, but also to expand my contribution beyond the computational counterpart and interpret the biological meaning of the results to find answers to biomedical questions and tackle important clinical needs.

The methodological work within the first part of the thesis has served as the basis for scientific discoveries giving insights into the mechanisms driving CLL evolution and its lethal Richter transformation with potential clinical value. Both aspects have given me a comprehensive view of modern genomics in Biomedicine and of current cancer genomics initiatives and their needs for a vast exploitation of the continuously generated data towards the translation and implementation of personalized medicine strategies within our health care systems.

Overall, the technological and methodological work carried out during the thesis served as the basis to engage in real biomedical studies. In the same way, it highlighted the importance of a proper infrastructure and well-defined

methodology for the implementation of large-scale, as well as small-scale, studies. These procedures are even more fundamental for the application of genomic analysis into the clinics, where the obtained results will guide clinical decisions that can directly affect the patients. Interdisciplinary efforts between clinicians, technicians, and bioinformaticians are necessary to understand the needs and contributions of each field. Rather than working separately, focusing only on their respective areas, multidisciplinary dialogues are the best way to respond to the real clinical needs.

6 Conclusions

1. We contributed to the large-scale genomics initiative Pancancer Analysis of Whole Genomes and made possible its execution in the BSC's HPC infrastructure. We adapted the project's computational solutions to the specific requirements of a traditional HPC environment, while maintaining the homogeneity of the analysis with the other data centers.

2. We have implemented and evaluated variant calling strategies. The analysis of the results we obtained pointed out that filtering and consensus strategies can improve their performance, and that there is a need for expert intervention to accurately interpret the results of variant calling as well as downstream analyses.

3. Global consideration of the analysis of Richter transformation in CLL indicates that it introduces a higher genomic, epigenomic, and transcriptomic complexity than CLL.

4. We have identified that the mutational profile of RT can be shaped by the imprint of previous mutagenic therapies, suggesting the prior existence of a cell that takes in and carries all these mutations until it expands at the time of transformation.

5. Using longitudinal whole-genome sequencing, we have unveiled the early presence of minute RT subclones up to the time of CLL diagnosis. Posterior external validations confirmed the presence of these early seeds that capture RT-specific features.

6. RT shows a distinct gene expression profile than its CLL counterpart. Aligned with the single-cell RNA sequencing and functional analyses performed by collaborators, we identified alterations in metabolism-related pathways (i.e., OXPHOS) that could be used as a therapeutic vulnerability.

7. From the analysis of a case of CLL in a young adult, we found that the CLL founding clone was solely defined by three *ATM* germline variants, suggesting their potential driver role in the initial development of CLL. As expected, somatically-acquired alterations under the selective pressure of treatments influenced the clonal evolution of the disease.

8. Taken together, these studies unveiled the heterogeneity that exists at the time of CLL diagnosis, often carrying the seeds that can potentially drive progression, relapse, and transformation.

7 References

Agathangelidis, A., Darzentas, N., Hadzidimitriou, A., … Stamatopoulos, K. (2012). Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood*, *119*(19), 4467–4475. https://doi.org/10.1182/BLOOD-2011-11-393694

Ahmed, A. E., Allen, J. M., Bhat, T., … Mainzer, L. S. (2021). Design considerations for workflow management systems use in production genomics research and the clinic. *Scientific Reports 2021 11:1*, *11*(1), 1–18. https://doi.org/10.1038/s41598-021-99288-8

Ahn, I. E., Underbayev, C., Albitar, A., … Wiestner, A. (2017). Clonal evolution leading to ibrutinib resistance in chronic lymphocytic leukemia. *Blood*, *129*(11), 1469–1479. https://doi.org/10.1182/blood-2016-06-719294

Alexandrov, L. B., Jones, P. H., Wedge, D. C., … Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, *47*(12), 1402–1407. https://doi.org/10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., … Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alioto, T. S., Buchhalter, I., Derdak, S., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*(1), 1–13. https://doi.org/10.1038/ncomms10001

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Amstutz, P., Crusoe, M. R., Tijanić, N., … Stojanovic, L. (2016). Common Workflow Language, v1.0. *Figshare*. https://doi.org/10.6084/M9.FIGSHARE.3115156

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/BIOINFORMATICS/BTU638

Anderson, M. A., Tam, C., Lew, T. E., … Roberts, A. W. (2017). Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood*, *129*(25), 3362–3370. https://doi.org/10.1182/blood-2017-01-763003

Andersson, E. R., Sandberg, R., & Lendahl, U. (2011). Notch signaling: simplicity in design, versatility in function. *Development*, *138*(17), 3593–3612. https://doi.org/10.1242/DEV.063610

Arthur, S. E., Jiang, A., Grande, B. M., … Morin, R. D. (2018). Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nature Communications*, *9*(1), 4001. https://doi.org/10.1038/s41467-018-06354-3

Auton, A., Abecasis, G. R., Altshuler, D. M., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Baca, S. C., Prandi, D., Lawrence, M. S., … Garraway, L. A. (2013). Punctuated Evolution of Prostate Cancer Genomes. *Cell*, *153*(3), 666–677. https://doi.org/10.1016/j.cell.2013.03.021

Bagnara, D., Tang, C., Brown, J. R., … Chiorazzi, N. (2021). Post-Transformation IGHV-IGHD-IGHJ Mutations in Chronic Lymphocytic Leukemia B Cells: Implications for Mutational Mechanisms and Impact on Clinical Course. *Frontiers in Oncology*, *11*, 1769. https://doi.org/10.3389/fonc.2021.640731

Bailey, M. H., Tokheim, C., Porta-Pardo, E., … Mariamidze, A. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Bamford, S., Dawson, E., Forbes, S., … Wooster, R. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, *91*(2), 355. https://doi.org/10.1038/SJ.BJC.6601894

Barone, G., Groom, A., Reiman, A., … Taylor, A. M. R. (2009). Modeling ATM mutant proteins from missense changes confirms retained kinase activity. *Human Mutation*, *30*(8), 1222–1230. https://doi.org/10.1002/humu.21034

Beà, S., López-Guillermo, A., Ribas, M., … Campo, E. (2002). Genetic Imbalances in Progressed B-Cell Chronic Lymphocytic Leukemia and Transformed Large-Cell Lymphoma (Richter's Syndrome). *The American Journal of Pathology*, *161*(3), 957–968. https://doi.org/10.1016/S0002-9440(10)64256-3

Beà, S., Valdés-Mas, R., Navarro, A., … Campo, E. (2013). Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proceedings of the National Academy of Sciences*, *110*(45), 18250–18255. https://doi.org/10.1073/pnas.1314608110

Beekman, R., Chapaprieta, V., Russiñol, N., … Martin-Subero, J. I. (2018). The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nature Medicine*, *24*(6), 868–880. https://doi.org/10.1038/s41591-018-0028-4

Benatti, S., Atene, C. G., Fiorcari, S., … Maffei, R. (2021). IRF4 L116R mutation promotes proliferation of chronic lymphocytic leukemia B cells inducing MYC. *Hematological Oncology*. https://doi.org/10.1002/hon.2915

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 2008 456:7218*, *456*(7218), 53–59. https://doi.org/10.1038/NATURE07517

Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, *15*(6), 353–365. https://doi.org/10.1038/s41571-018-0002-6

Bergmann, E. A., Chen, B. J., Arora, K., … Zody, M. C. (2016). Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics*, *32*(20), 3196–3198. https://doi.org/10.1093/BIOINFORMATICS/BTW389

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolli, N., Avet-Loiseau, H., Wedge, D. C., … Munshi, N. C. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*, *5*(1), 2997. https://doi.org/10.1038/ncomms3997

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., … Van Duijn, C. M. (2014). The Genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics*, *22*(2), 221–227. https://doi.org/10.1038/ejhg.2013.118

Bortnick, A., & Allman, D. (2013). What is and what should always have been: long-lived plasma cells induced by T cell-independent antigens. *Journal of Immunology (Baltimore, Md. : 1950)*, *190*(12), 5913–5918. https://doi.org/10.4049/JIMMUNOL.1300161

Brady, S. W., Gout, A. M., & Zhang, J. (2021). Therapeutic and prognostic insights from the analysis of cancer mutational signatures. *Trends in Genetics*, *0*(0). https://doi.org/10.1016/J.TIG.2021.08.007

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Bretones, G., Álvarez, M. G., Arango, J. R., … López-Otín, C. (2018). Altered patterns of global protein synthesis and translational fidelity in RPS15-mutated chronic lymphocytic leukemia. *Blood*, *132*(22), 2375–2388. https://doi.org/10.1182/BLOOD-2017-09-804401

Brochet, X., Lefranc, M.-P., & Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Research*, *36*(Web Server), W503–W508. https://doi.org/10.1093/nar/gkn316

Brown, J. R., Hanna, M., Tesar, B., … Freedman, A. S. (2012). Germline copy number variation associated with Mendelian inheritance of CLL in two families. *Leukemia 2012 26:7*, *26*(7), 1710–1713. https://doi.org/10.1038/leu.2012.33

Brown, J. R., Hillmen, P., O'Brien, S., … Byrd, J. C. (2017). Extended follow-up and impact of high-risk prognostic factors from the phase 3 RESONATE study in patients with previously treated CLL/SLL. *Leukemia 2018 32:1*, *32*(1), 83–91. https://doi.org/10.1038/LEU.2017.175

Burger, J. A., & Buggy, J. J. (2013). Bruton tyrosine kinase inhibitor ibrutinib (PCI-32765). *Leukemia & Lymphoma*, *54*(11), 2385–2391. https://doi.org/10.3109/10428194.2013.777837

Burger, J. A., Landau, D. A., Taylor-Weiner, A., … Wu, C. J. (2016). Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11589

Byrd, J. C., Furman, R. R., Coutre, S. E., … O'Brien, S. (2013). Targeting BTK with Ibrutinib in Relapsed Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *369*(1), 32–42. https://doi.org/10.1056/NEJMoa1215637

Bystry, V., Agathangelidis, A., Bikos, V., … Darzentas, N. (2015). ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics*, *31*(23), btv456. https://doi.org/10.1093/bioinformatics/btv456

Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature*, *255*(5505), 197–200. https://doi.org/10.1038/255197a0

Calabrese, C., Davidson, N. R., Demircioglu, D., … Brooks, A. N. (2020). Genomic basis for RNA alterations in cancer. *Nature 2020 578:7793*, *578*(7793), 129–136. https://doi.org/10.1038/S41586-020-1970-0

Calissano, C., Damle, R. N., Hayes, G., … Chiorazzi, N. (2009). In vivo intraclonal and interclonal kinetic heterogeneity in B-cell chronic lymphocytic leukemia. *Blood*, *114*(23), 4832–4842. https://doi.org/10.1182/blood-2009-05-219634

Calissano, C., Damle, R. N., Marsilio, S., … Chiorazzi, N. (2011). Intraclonal Complexity in Chronic Lymphocytic Leukemia: Fractions Enriched in Recently Born/Divided and Older/Quiescent Cells. *Molecular Medicine*, *17*(11–12), 1374–1382. https://doi.org/10.2119/molmed.2011.00360

Campbell, P. J., Getz, G., Korbel, J. O., … ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Campo, E., Cymbalista, F., Ghia, P., … Stilgenbauer, S. (2018). TP53 aberrations in chronic lymphocytic leukemia: an overview of the clinical implications of improved diagnostics. *Haematologica*, *103*(12), 1956–1968. https://doi.org/10.3324/HAEMATOL.2018.187583

Caravagna, G., Heide, T., Williams, M. J., … Sottoriva, A. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics 2020 52:9*, *52*(9), 898–907. https://doi.org/10.1038/S41588-020-0675-5

Caro, P., Kishan, A. U., Norberg, E., … Danial, N. N. (2012). Metabolic Signatures Uncover Distinct Targets in Molecular Subsets of Diffuse Large B Cell Lymphoma. *Cancer Cell*, *22*(4), 547–560.

https://doi.org/10.1016/j.ccr.2012.08.014
Cerhan, J. R., & Slager, S. L. (2015). Familial predisposition and genetic risk factors for lymphoma. *Blood*, *126*(20), 2265–2273. https://doi.org/10.1182/BLOOD-2015-04-537498
Chakraborty, S., Martines, C., Porro, F., … Efremov, D. G. (2021). B-cell receptor signaling and genetic lesions in TP53 and CDKN2A/CDKN2B cooperate in Richter transformation. *Blood*, *138*(12), 1053–1066. https://doi.org/10.1182/blood.2020008276
Chakravarty, D., Gao, J., Phillips, S., … Schultz, N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, *1*, 1–16. https://doi.org/10.1200/PO.17.00011
Chan, K.-L., Blombery, P., Jones, K., … Tam, C. S. (2017). Plasmablastic Richter transformation as a resistance mechanism for chronic lymphocytic leukaemia treated with BCR signalling inhibitors. *British Journal of Haematology*, *177*(2), 324–328. https://doi.org/10.1111/bjh.14062
Chapuy, B., Stewart, C., Dunford, A. J., … Shipp, M. A. (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*, *24*(5), 679–690. https://doi.org/10.1038/s41591-018-0016-8
Chen, C.-L., Hsu, S.-C., Chung, T.-Y., … Kung, H.-J. (2021). Arginine is an epigenetic regulator targeting TEAD4 to modulate OXPHOS in prostate cancer cells. *Nature Communications*, *12*(1), 2398. https://doi.org/10.1038/s41467-021-22652-9
Cherng, H. J., Jammal, N., Paul, S., … Jain, N. (2021). Clinical and molecular characteristics and treatment patterns of adolescent and young adult patients with chronic lymphocytic leukaemia. *British Journal of Haematology*, *194*(1), 61–68. https://doi.org/10.1111/bjh.17498
Chigrinova, E., Rinaldi, A., Kwee, I., … Bertoni, F. (2013). Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood*, *122*(15), 2673–2682. https://doi.org/10.1182/blood-2013-03-489518
Chitalia, A., Swoboda, D. M., McCutcheon, J. N., … Cheson, B. D. (2019). Descriptive analysis of genetic aberrations and cell of origin in Richter transformation. *Leukemia & Lymphoma*, *60*(4), 971–979. https://doi.org/10.1080/10428194.2018.1516878
Cingolani, P., Patel, V. M., Coon, M., … Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*, 35. https://doi.org/10.3389/fgene.2012.00035
Cingolani, P., Platts, A., Wang, L. L., … Ruden, D. M. (2012). 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*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695
Close, V., Close, W., Kugler, S. J., … Mertens, D. (2019). FBXW7 mutations reduce binding of NOTCH1, leading to cleaved NOTCH1 accumulation and target gene activation in CLL. *Blood*, *133*(8), 830–839. https://doi.org/10.1182/BLOOD-2018-09-874529
Cobo, F., Martínez, A., Pinyol, M., … Campo, E. (2002). Multiple cell cycle regulator alterations in Richter's transformation of chronic lymphocytic leukemia. *Leukemia*, *16*(6), 1028–1034. https://doi.org/10.1038/sj.leu.2402529
Cortés-Ciriano, I., Lee, J. J.-K., Xi, R., … PCAWG Consortium. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, *52*(3), 331–341. https://doi.org/10.1038/s41588-019-0576-7
Costello, J. C., & Stolovitzky, G. (2013). Seeking the Wisdom of Crowds Through Challenge-Based Competitions in Biomedical Research. *Clinical Pharmacology & Therapeutics*, *93*(5), 396–398. https://doi.org/10.1038/CLPT.2013.36
Craig Venter, J., Adams, M. D., Myers, E. W., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040
Cui, B., Chen, L., Zhang, S., … Kipps, T. J. (2014). MicroRNA-155 influences B-cell receptor signaling and associates with aggressive disease in chronic lymphocytic leukemia. *Blood*, *124*(4), 546–554. https://doi.org/10.1182/blood-2014-03-559690
Dadashian, E. L., McAuley, E. M., Liu, D., … Herman, S. E. M. (2019). TLR Signaling Is Activated in Lymph Node–Resident CLL Cells and Is Only Partially Inhibited by Ibrutinib. *Cancer Research*, *79*(2), 360–371. https://doi.org/10.1158/0008-5472.CAN-18-0781
Damle, R. N., Wasil, T., Fais, F., … Chiorazzi, N. (1999). Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic LeukemiaPresented in part at the 40th Annual Meeting of The American Society of Hematology, held in Miami Beach, FL, December 4-8, 1998. *Blood*, *94*(6), 1840–1847. https://doi.org/10.1182/BLOOD.V94.6.1840
Damm, F., Mylonas, E., Cosson, A., … Bernard, O. A. (2014). Acquired Initiating Mutations in Early Hematopoietic Cells of CLL Patients. *Cancer Discovery*, *4*(9), 1088–1101. https://doi.org/10.1158/2159-8290.CD-14-0104
Danecek, P., Bonfield, J. K., Liddle, J., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008
Davis, A., Gao, R., & Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1867*(2), 151–161. https://doi.org/10.1016/J.BBCAN.2017.01.003
Davis, C. A., Hitz, B. C., Sloan, C. A., … Cherry, J. M. (2018). The Encyclopedia of

DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

de Kanter, J. K., Peci, F., Bertrums, E., … van Boxtel, R. (2021). Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell*, *28*(10), 1726-1739.e6. https://doi.org/10.1016/j.stem.2021.07.012

De Paoli, L., Cerri, M., Monti, S., … Rossi, D. (2013). MGA, a suppressor of MYC, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leukemia & Lymphoma*, *54*(5), 1087–1090. https://doi.org/10.3109/10428194.2012.723706

Dentro, S. C., Leshchiner, I., Haase, K., … Loo, P. Van. (2021). Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, *184*(8), 2239-2254.e39. https://doi.org/10.1016/J.CELL.2021.03.009

Dentro, S. C., Wedge, D. C., & Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine*, *7*(8), a026625. https://doi.org/10.1101/cshperspect.a026625

DI Tommaso, P., Chatzou, M., Floden, E. W., … Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology 2017 35:4*, *35*(4), 316–319. https://doi.org/10.1038/NBT.3820

Ding, L., Bailey, M. H., Porta-Pardo, E., … Mariamidze, A. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, *173*(2), 305-320.e10. https://doi.org/10.1016/j.cell.2018.03.033

Ding, W. (2018). Richter transformation in the era of novel agents. *Hematology*, *2018*(1), 256–263. https://doi.org/10.1182/asheducation-2018.1.256

Dobin, A., Davis, C. A., Schlesinger, F., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. https://doi.org/10.1093/NAR/GKN425

Döhner, H., Stilgenbauer, S., Benner, A., … Lichter, P. (2000). Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *343*(26), 1910–1916. https://doi.org/10.1056/NEJM200012283432602

Dunham, I., Kundaje, A., Aldred, S. F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

E. Ezekwudo, D., Ifabiyi, T., Usim, A., & Jaiyesimi, I. (2019). "Double Hit" in Chronic Lymphocytic Leukemia: Therapeutic Strategies for Patients with 17p Deletion and TP53 Mutation. *Journal of Cancer Science and Clinical Therapeutics*, *03*(02), 54–69. https://doi.org/10.26502/jcsct.5079019

Eberle, M. A., Fritzilas, E., Krusche, P., … Bentley, D. R. (2017). A reference data set

of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, *27*(1), 157–164. https://doi.org/10.1101/GR.210500.116

Eberwine, J., Sul, J. Y., Bartfai, T., & Kim, J. (2013). The promise of single-cell sequencing. *Nature Methods 2014 11:1*, *11*(1), 25–27. https://doi.org/10.1038/nmeth.2769

Eggermont, A. M. M., Apolone, G., Baumann, M., … Calvo, F. (2019). Cancer Core Europe: A translational research infrastructure for a European mission on cancer. *Molecular Oncology*, *13*(3), 521–527. https://doi.org/10.1002/1878-0261.12447

Ellrott, K., Bailey, M. H., Saksena, G., … Mariamidze, A. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems*, *6*(3), 271-281.e7. https://doi.org/10.1016/j.cels.2018.03.002

Espejo Valle-Inclan, J., Besselink, N. J. M., de Bruijn, E., … Cuppen, E. (2022). A multi-platform reference for somatic structural variation detection. *Cell Genomics*, *2*(6), 100139. https://doi.org/10.1016/J.XGEN.2022.100139

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

Ewing, A. D., Houlahan, K. E., Hu, Y., … Boutros, P. C. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods 2015 12:7*, *12*(7), 623–630. https://doi.org/10.1038/nmeth.3407

Fabbri, G., Holmes, A. B., Viganotti, M., … Dalla-Favera, R. (2017). Common nonmutational NOTCH1 activation in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(14), E2911–E2919. https://doi.org/10.1073/PNAS.1702564114

Fabbri, G., Khiabanian, H., Holmes, A. B., … Dalla-Favera, R. (2013). Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *Journal of Experimental Medicine*, *210*(11), 2273–2288. https://doi.org/10.1084/jem.20131448

Fan, Y., Xi, L., Hughes, D. S. T., … Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, *17*(1), 178. https://doi.org/10.1186/s13059-016-1029-6

Ferrando, A. A., & López-Otín, C. (2017). Clonal evolution in leukemia. *Nature Medicine*, *23*(10), 1135–1145. https://doi.org/10.1038/nm.4410

Flinn, I. W., O'Brien, S., Kahl, B., … Horwitz, S. (2018). Duvelisib, a novel oral dual inhibitor of PI3K-δ,γ, is clinically active in advanced hematologic

malignancies. *Blood*, *131*(8), 877–887. https://doi.org/10.1182/BLOOD-2017-05-786566

Furman, R. R., Cheng, S., Lu, P., … Wang, Y. L. (2014). Ibrutinib resistance in chronic lymphocytic leukemia. In *New England Journal of Medicine* (Vol. 370, Issue 24, pp. 2352–2354). Massachussetts Medical Society. https://doi.org/10.1056/NEJMc1402716

Furman, R. R., Sharman, J. P., Coutre, S. E., … O'Brien, S. M. (2014). Idelalisib and Rituximab in Relapsed Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *370*(11), 997–1007. https://doi.org/10.1056/NEJMoa1315226

Gahn, B., Schäfer, C., Neef, J., … Wörmann, B. (1997). Detection of Trisomy 12 and Rb-Deletion in CD34+ Cells of Patients With B-Cell Chronic Lymphocytic Leukemia. *Blood*, *89*(12), 4275–4281. https://doi.org/10.1182/BLOOD.V89.12.4275

Gaiti, F., Chaligne, R., Gu, H., … Landau, D. A. (2019). Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, *569*(7757), 576–580. https://doi.org/10.1038/s41586-019-1198-z

Gao, J., Aksoy, B. A., Dogrusoz, U., … Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, *6*(269). https://doi.org/10.1126/SCISIGNAL.2004088

Gemenetzi, K., Psomopoulos, F., Carriles, A. A., … Chatzidimitriou, A. (2021). Higher-order immunoglobulin repertoire restrictions in CLL: the illustrative case of stereotyped subsets 2 and 169. *Blood*, *137*(14), 1895–1904. https://doi.org/10.1182/blood.2020005216

Gerstung, M., Jolly, C., Leshchiner, I., … Van Loo, P. (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128. https://doi.org/10.1038/s41586-019-1907-7

Gohil, S. H., & Wu, C. J. (2019). Dissecting CLL through high-dimensional single-cell technologies. *Blood*, *133*(13), 1446. https://doi.org/10.1182/BLOOD-2018-09-835389

Goldin, L. R., Bjorkholm, M., Kristinsson, S. Y., … Landgren, O. (2009). Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. *Haematologica*, *94*(5), 647–653. https://doi.org/10.3324/haematol.2008.003632

Goldin, L. R., McMaster, M. L., Rotunno, M., … Caporaso, N. E. (2016). Whole exome sequencing in families with CLL detects a variant in Integrin β 2 associated with disease susceptibility. *Blood*, *128*(18), 2261–2263. https://doi.org/10.1182/BLOOD-2016-02-697771

Goldin, L. R., Pfeiffer, R. M., Li, X., & Hemminki, K. (2004). Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database. *Blood*, *104*(6), 1850–1854. https://doi.org/10.1182/BLOOD-2004-01-0341

Good, B. M., Ainscough, B. J., McMichael, J. F., … Griffith, O. L. (2014). Organizing knowledge to enable personalization of medicine in cancer. *Genome Biology*, *15*(8), 438. https://doi.org/10.1186/s13059-014-0438-7

Greaves, M. (2015). Evolutionary Determinants of Cancer. *Cancer Discovery*, *5*(8), 806–820. https://doi.org/10.1158/2159-8290.CD-15-0439

Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, *481*(7381), 306–313. https://doi.org/10.1038/nature10762

Griffith, M., Miller, C. A., Griffith, O. L., … Wilson, R. K. (2015). Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems*, *1*(3), 210–223. https://doi.org/10.1016/j.cels.2015.08.015

Griffith, M., Spies, N. C., Krysiak, K., … Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. In *Nature Genetics* (Vol. 49, Issue 2, pp. 170–174). Nature Publishing Group. https://doi.org/10.1038/ng.3774

Gruber, M., & Wu, C. J. (2014). Evolving Understanding of the CLL Genome. *Seminars in Hematology*, *51*(3), 177–187. https://doi.org/10.1053/J.SEMINHEMATOL.2014.05.004

Guièze, R., Liu, V. M., Rosebrock, D., … Wu, C. J. (2019). Mitochondrial Reprogramming Underlies Resistance to BCL-2 Inhibition in Lymphoid Malignancies. *Cancer Cell*, *36*(4), 369-384.e13. https://doi.org/10.1016/j.ccell.2019.08.005

Hallek, M. (2019). Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. *American Journal of Hematology*, *94*(11), 1266–1287. https://doi.org/10.1002/AJH.25595

Hallek, M., Cheson, B. D., Catovsky, D., … Kipps, T. J. (2018). iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. *Blood*, *131*(25), 2745–2760. https://doi.org/10.1182/BLOOD-2017-09-806398

Hamblin, T. J., Davis, Z., Gardiner, A., … Stevenson, F. K. (1999). Unmutated Ig VH Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. *Blood*, *94*(6), 1848–1854. https://doi.org/10.1182/BLOOD.V94.6.1848

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Haradhvala, N. J., Kim, J., Maruvka, Y. E., … Getz, G. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications 2018 9:1*, *9*(1), 1–9. https://doi.org/10.1038/s41467-018-04002-4

Haradhvala, N. J., Polak, P., Stojanov, P., … Getz, G. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and

Repair. *Cell*, *164*(3), 538–549. https://doi.org/10.1016/j.cell.2015.12.050

Hasan, M. K., Ghia, E. M., Rassenti, L. Z., … Kipps, T. J. (2021). Wnt5a enhances proliferation of chronic lymphocytic leukemia and ERK1/2 phosphorylation via a ROR1/DOCK2-dependent mechanism. *Leukemia*, *35*(6), 1621–1630. https://doi.org/10.1038/s41375-020-01055-7

Herling, C. D., Abedpour, N., Weiss, J., … Peifer, M. (2018). Clonal dynamics towards the development of venetoclax resistance in chronic lymphocytic leukemia. *Nature Communications*, *9*(1), 727. https://doi.org/10.1038/s41467-018-03170-7

Hoadley, K. A., Yau, C., Hinoue, T., … Mariamidze, A. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*(2), 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022

Hudson, T. J., Anderson, W., Aretz, A., … Wainwright, B. J. (2010). International network of cancer genome projects. *Nature 2010 464:7291*, *464*(7291), 993–998. https://doi.org/10.1038/NATURE08987

Innocenti, I., Rossi, D., Trapè, G., … Laurenti, L. (2018). Clinical, pathological, and biological characterization of Richter syndrome developing after ibrutinib treatment for relapsed chronic lymphocytic leukemia. *Hematological Oncology*, *36*(3), 600–603. https://doi.org/10.1002/HON.2502

Itchaki, G., Tiao, G., Improgo, M. R. D., … Brown, J. R. (2017). Rare Germline Variant in NFATC4 Associated with Familial Chronic Lymphocytic Leukemia. *Blood*, *130*(Supplement 1), 2993–2993. https://doi.org/https://doi.org/10.1182/blood.V130.Suppl_1.2993.2993

Jain, P., Keating, M., Wierda, W., … O'Brien, S. (2015). Outcomes of patients with chronic lymphocytic leukemia after discontinuing ibrutinib. *Blood*, *125*(13), 2062–2067. https://doi.org/10.1182/blood-2014-09-603670

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., … Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research 2017 6:876*, *6*, 876. https://doi.org/10.12688/f1000research.11407.1

Jones, D., Raine, K. M., Davies, H., … Campbell, P. J. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Current Protocols in Bioinformatics*, *56*(1), 15.10.1-15.10.18. https://doi.org/10.1002/cpbi.20

Kadri, S., Lee, J., Fitzpatrick, C., … Wang, Y. L. (2017). Clonal evolution underlying leukemia progression and Richter transformation in patients with ibrutinib-relapsed CLL. *Blood Advances*, *1*(12), 715–727. https://doi.org/10.1182/BLOODADVANCES.2016003632

Karczewski, K. J., Francioli, L. C., Tiao, G., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karube, K., Enjuanes, A., Dlouhy, I., … Campo, E. (2018). Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia*, *32*(3), 675–684. https://doi.org/10.1038/leu.2017.251

Kasar, S., Kim, J., Improgo, R., … Brown, J. R. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, *6*(1), 8866. https://doi.org/10.1038/ncomms9866

Kikushige, Y., Ishikawa, F., Miyamoto, T., … Akashi, K. (2011). Self-Renewing Hematopoietic Stem Cell Is the Primary Target in Pathogenesis of Human Chronic Lymphocytic Leukemia. *Cancer Cell*, *20*(2), 246–259. https://doi.org/10.1016/J.CCR.2011.06.029

Kim, S., Scheffler, K., Halpern, A. L., … Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods 2018 15:8*, *15*(8), 591–594. https://doi.org/10.1038/S41592-018-0051-X

Kim, S. Y., Jacob, L., & Speed, T. P. (2014). Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics*, *15*(1), 154. https://doi.org/10.1186/1471-2105-15-154

Kipps, T. J., Stevenson, F. K., Wu, C. J., … Rai, K. (2017). Chronic lymphocytic leukaemia. *Nature Reviews Disease Primers 2017 3:1*, *3*(1), 1–22. https://doi.org/10.1038/NRDP.2016.96

Klein, U., & Dalla-Favera, R. (2008). Germinal centres: role in B-cell physiology and malignancy. *Nature Reviews Immunology 2008 8:1*, *8*(1), 22–33. https://doi.org/10.1038/NRI2217

Klintman, J., Appleby, N., Stamatopoulos, B., … Schuh, A. (2021). Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood*, *137*(20), 2800–2816. https://doi.org/10.1182/blood.2020005650

Koboldt, D. C., Zhang, Q., Larson, D. E., … Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Koh, G., Zou, X., & Nik-Zainal, S. (2020). Mutational signatures: experimental design and analytical framework. *Genome Biology*, *21*(1), 37. https://doi.org/10.1186/s13059-020-1951-5

Kohlhaas, V., Blakemore, S. J., Al-Maarri, M., … Wunderlich, F. T. (2021). Active Akt signaling triggers CLL toward Richter transformation via overactivation of Notch1. *Blood*, *137*(5), 646–660. https://doi.org/10.1182/BLOOD.2020005734

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Korbel, J. O., & Campbell, P. J. (2013). Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell*, *152*(6), 1226–1236. https://doi.org/10.1016/j.cell.2013.02.023

Koren, A., Polak, P., Nemesh, J., … McCarroll, S. A. (2012). Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *The American Journal of Human Genetics*, *91*(6), 1033–1040. https://doi.org/10.1016/j.ajhg.2012.10.018

Kostopoulos, I. V., Tsakiridou, A. A., Pavlidis, D., … Papadhimitriou, S. I. (2015). Familial chronic lymphocytic leukemia in two siblings with ATM/13q14 deletion and a similar pattern of clonal evolution. *Blood Cancer Journal*, *5*(7), e322–e322. https://doi.org/10.1038/bcj.2015.50

Krusche, P., Trigg, L., Boutros, P. C., … Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560. https://doi.org/10.1038/S41587-019-0054-X

Kucab, J. E., Zou, X., Morganella, S., … Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell*, *177*(4), 821-836.e16. https://doi.org/10.1016/j.cell.2019.03.001

Kulis, M., Heath, S., Bibikova, M., … Martín-Subero, J. I. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature Genetics*, *44*(11), 1236–1242. https://doi.org/10.1038/ng.2443

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Lai, Z., Markovets, A., Ahdesmaki, M., … Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw227

Lampson, B. L., Kasar, S. N., Matos, T. R., … Brown, J. R. (2016). Idelalisib given front-line for treatment of chronic lymphocytic leukemia causes frequent immune-mediated hepatotoxicity. *Blood*, *128*(2), 195–203. https://doi.org/10.1182/BLOOD-2016-03-707133

Landau, D. A., Carter, S. L., Stojanov, P., … Wu, C. J. (2013). Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, *152*(4), 714–726. https://doi.org/10.1016/j.cell.2013.01.019

Landau, D. A., Sun, C., Rosebrock, D., … Wu, C. J. (2017). The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nature Communications*, *8*(1), 1–12. https://doi.org/10.1038/s41467-017-02329-y

Landau, D. A., Tausch, E., Taylor-Weiner, A. N., … Wu, C. J. (2015). Mutations driving CLL and their evolution in progression and relapse. *Nature*, *526*(7574), 525–530. https://doi.org/10.1038/nature15395

Lander, E. S., Linton, L. M., Birren, B., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Law, P. J., Berndt, S. I., Speedy, H. E., … Slager, S. (2017). Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nature Communications 2017 8:1*, *8*(1), 1–12. https://doi.org/10.1038/NCOMMS14175

Lee-Six, H., Olafsson, S., Ellis, P., … Stratton, M. R. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, *574*(7779), 532–537. https://doi.org/10.1038/s41586-019-1672-7

Lee, A. Y., Ewing, A. D., Ellrott, K., … Ye, K. (2018). Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biology*, *19*(1), 1–15. https://doi.org/10.1186/S13059-018-1539-5/FIGURES/4

Lek, M., Karczewski, K. J., Minikel, E. V., … Williams, A. L. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature 2016 536:7616*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Letai, A. (2017). Functional precision cancer medicine—moving beyond pure genomics. *Nature Medicine 2017 23:9*, *23*(9), 1028–1035. https://doi.org/10.1038/NM.4389

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y. I., Knowles, D. A., Humphrey, J., … Pritchard, J. K. (2017). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics 2017 50:1*, *50*(1), 151–158. https://doi.org/10.1038/s41588-017-0004-9

Li, Y., Roberts, N. D., Wala, J. A., … Campbell, P. J. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature*, *578*(7793), 112–121. https://doi.org/10.1038/s41586-019-1913-9

Lin, W. Y., Fordham, S. E., Sunter, N., … Allan, J. M. (2021). Genome-wide association study identifies risk loci for progressive chronic lymphocytic leukemia. *Nature Communications 2021 12:1*, *12*(1), 1–8. https://doi.org/10.1038/s41467-020-20822-9

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241.

https://doi.org/10.1002/humu.22932
Ljungström, V., Cortese, D., Young, E., … Rosenquist, R. (2016). Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. *Blood*, *127*(8), 1007–1016. https://doi.org/10.1182/BLOOD-2015-10-674572
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
Luskin, M., Wertheim, G., Morrissette, J., … Loren, A. (2014). CLL/SLL diagnosed in an adolescent. *Pediatric Blood & Cancer*, *61*(6), 1107–1110. https://doi.org/10.1002/pbc.24884
Maciejowski, J., & De Lange, T. (2017). Telomeres in cancer: tumour suppression and genome instability. *Nature Reviews Molecular Cell Biology 2017 18:3*, *18*(3), 175–186. https://doi.org/10.1038/nrm.2016.171
Maddocks, K. J., Ruppert, A. S., Lozanski, G., … Woyach, J. A. (2015). Etiology of Ibrutinib Therapy Discontinuation and Outcomes in Patients With Chronic Lymphocytic Leukemia. *JAMA Oncology*, *1*(1), 80. https://doi.org/10.1001/jamaoncol.2014.218
Maffei, R., Fiorcari, S., Benatti, S., … Marasca, R. (2021). IRF4 modulates the response to BCR activation in chronic lymphocytic leukemia regulating IKAROS and SYK. *Leukemia*, *35*(5), 1330–1343. https://doi.org/10.1038/s41375-021-01178-5
Mansouri, L., Grabowski, P., Degerman, S., … Rosenquist, R. (2013). Short telomere length is associated with NOTCH1/SF3B1/TP53 aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. *American Journal of Hematology*, *88*(8), 647–651. https://doi.org/10.1002/AJH.23466
Mansouri, L., Papakonstantinou, N., Ntoufa, S., … Rosenquist, R. (2016). NF-κB activation in chronic lymphocytic leukemia: A point of convergence of external triggers and intrinsic lesions. *Seminars in Cancer Biology*, *39*, 40–48. https://doi.org/10.1016/J.SEMCANCER.2016.07.005
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200
Martincorena, I., Raine, K. M., Gerstung, M., … Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042
Martínez-Jiménez, F., Muiños, F., Sentís, I., … Lopez-Bigas, N. (2020). A compendium of mutational cancer driver genes. *Nature Reviews Cancer 2020 20:10*, *20*(10), 555–572. https://doi.org/10.1038/S41568-020-0290-X
Maura, F., Bolli, N., Angelopoulos, N., … Campbell, P. J. (2019). Genomic landscape and chronological reconstruction of driver events in multiple myeloma.

Nature Communications, *10*(1), 3835. https://doi.org/10.1038/s41467-019-11680-1

Maura, F., Degasperi, A., Nadeu, F., … Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-11037-8

Maura, F., Weinhold, N., Diamond, B., … Landgren, O. (2021). The mutagenic impact of melphalan in multiple myeloma. *Leukemia 2021*, 1–6. https://doi.org/10.1038/S41375-021-01293-3

Mayakonda, A., Lin, D.-C., Assenov, Y., … Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

McGranahan, N., Favero, F., de Bruin, E. C., … Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, *7*(283). https://doi.org/10.1126/scitranslmed.aaa1408

McKenna, A., Hanna, M., Banks, E., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mercer, T. R., Xu, J., Mason, C. E., & Tong, W. (2021). The Sequencing Quality Control 2 study: establishing community standards for sequencing in precision medicine. *Genome Biology*, *22*(1), 306. https://doi.org/10.1186/s13059-021-02528-3

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*.

Method of the Year 2013. (2013). *Nature Methods 2014 11:1*, *11*(1), 1–1. https://doi.org/10.1038/nmeth.2801

Minici, C., Gounari, M., Übelhart, R., … Degano, M. (2017). Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukaemia. *Nature Communications*, *8*(1), 15746. https://doi.org/10.1038/ncomms15746

Molina, J. R., Sun, Y., Protopopova, M., … Marszalek, J. R. (2018). An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nature Medicine*, *24*(7), 1036–1046. https://doi.org/10.1038/s41591-018-0052-4

Moncunill, V., Gonzalez, S., Beà, S., … Torrents, D. (2014). Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, *32*(11), 1106–1112. https://doi.org/10.1038/nbt.3027

Monti, S. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, *105*(5), 1851–1861. https://doi.org/10.1182/blood-2004-07-2947

Morash, M., Mitchell, H., Beltran, H., … Pathak, J. (2018). The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *Journal of Personalized Medicine*, *8*(3), 30. https://doi.org/10.3390/jpm8030030

Morganti, S., Tarantino, P., Ferraro, E., … Curigliano, G. (2020). Role of Next-Generation Sequencing Technologies in Personalized Medicine. In *P5 eHealth: An Agenda for the Health Technologies of the Future* (pp. 125–154). Springer International Publishing. https://doi.org/10.1007/978-3-030-27994-3_8

Muller, E., Goardon, N., Brault, B., … Castera, L. (2016). OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget*, *7*(48), 79485–79493. https://doi.org/10.18632/oncotarget.13103

Nadeu, F., Clot, G., Delgado, J., … Campo, E. (2018). Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia*, *32*(3), 645–653. https://doi.org/10.1038/leu.2017.291

Nadeu, F., Delgado, J., Royo, C., … Campo, E. (2016). Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood*, *127*(17), 2122–2130. https://doi.org/10.1182/blood-2015-07-659144

Nadeu, F., Diaz-Navarro, A., Delgado, J., … Campo, E. (2020). Genomic and Epigenomic Alterations in Chronic Lymphocytic Leukemia. *Annual Review of Pathology: Mechanisms of Disease*, *15*(1), 149–177. https://doi.org/10.1146/annurev-pathmechdis-012419-032810

Nadeu, F., Martin-Garcia, D., Clot, G., … Campo, E. (2020). Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. *Blood*, *136*(12), 1419–1432. https://doi.org/10.1182/blood.2020005289

Nadeu, F., Mas-de-les-Valls, R., Navarro, A., … Campo, E. (2020). IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nature Communications*, *11*(1), 3390. https://doi.org/10.1038/s41467-020-17095-7

Narzisi, G., Corvelo, A., Arora, K., … Zody, M. C. (2018). Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Communications Biology 2018 1:1*, *1*(1), 1–9. https://doi.org/10.1038/s42003-018-0023-9

Nassereddine, S., & Dunleavy, K. (2019). A Case of Chronic Lymphocytic Leukemia in an AYA Patient. *Clinical Lymphoma Myeloma and Leukemia*, *19*, S280. https://doi.org/10.1016/j.clml.2019.07.216

Ng, A. W. T., Poon, S. L., Huang, M. N., … Rozen, S. G. (2017). Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Science Translational Medicine*, *9*(412), eaan6446.

https://doi.org/10.1126/scitranslmed.aan6446
Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., … Stratton, M. R. (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, *149*(5), 979–993. https://doi.org/10.1016/j.cell.2012.04.024
Nik-Zainal, S., Van Loo, P., Wedge, D. C., … Breast Cancer Working Group of the International Cancer Genome Consortium. (2012). The life history of 21 breast cancers. *Cell*, *149*(5), 994–1007. https://doi.org/10.1016/j.cell.2012.04.023
Norberg, E., Lako, A., Chen, P.-H., … Danial, N. N. (2017). Differential contribution of the mitochondrial translation pathway to the survival of diffuse large B-cell lymphoma subsets. *Cell Death & Differentiation*, *24*(2), 251–262. https://doi.org/10.1038/cdd.2016.116
Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations. *Science*, *194*(4260), 23–28. https://doi.org/10.1126/science.959840
Ntoufa, S., Vilia, M. G., Stamatopoulos, K., … Muzio, M. (2016). Toll-like receptors signaling: A complex network for NF-κB activation in B-cell lymphoid malignancies. *Seminars in Cancer Biology*, *39*, 15–25. https://doi.org/10.1016/j.semcancer.2016.07.001
O'Connor, B. D., Merriman, B., & Nelson, S. F. (2010). SeqWare Query Engine: storing and searching sequence data in the cloud. *BMC Bioinformatics*, *11*(S12), S2. https://doi.org/10.1186/1471-2105-11-S12-S2
Oakes, C. C., Seifert, M., Assenov, Y., … Plass, C. (2016). DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nature Genetics 2015 48:3*, *48*(3), 253–264. https://doi.org/10.1038/NG.3488
Ouillette, P., Fossum, S., Parkin, B., … Malek, S. N. (2010). Aggressive Chronic Lymphocytic Leukemia with Elevated Genomic Complexity Is Associated with Multiple Gene Defects in the Response to DNA Double-Strand Breaks. *Clinical Cancer Research*, *16*(3), 835–847. https://doi.org/10.1158/1078-0432.CCR-09-2534
Parker, H., Rose-Zerilli, M. J. J., Larrayoz, M., … Strefford, J. C. (2016). Genomic disruption of the histone methyltransferase SETD2 in chronic lymphocytic leukaemia. *Leukemia 2016 30:11*, *30*(11), 2179–2186. https://doi.org/10.1038/LEU.2016.134
Patel, K., & Pagel, J. M. (2021). Current and future treatment strategies in chronic lymphocytic leukemia. *Journal of Hematology & Oncology 2021 14:1*, *14*(1), 1–20. https://doi.org/10.1186/S13045-021-01054-W
Patterson, D. G., Kania, A. K., Price, M. J., … Boss, J. M. (2021). An IRF4–MYC–mTORC1 Integrated Pathway Controls Cell Growth and the Proliferative Capacity of Activated B Cells during B Cell Differentiation In Vivo. *The Journal of Immunology*, *207*(7), 1798–1811. https://doi.org/10.4049/jimmunol.2100440

Patterson, S. E., Liu, R., Statz, C. M., … Mockus, S. M. (2016). The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human Genomics*, *10*(1). https://doi.org/10.1186/S40246-016-0061-7

Penter, L., Gohil, S. H., Lareau, C., … Wu, C. J. (2021). Longitudinal Single-Cell Dynamics of Chromatin Accessibility and Mitochondrial Mutations in Chronic Lymphocytic Leukemia Mirror Disease History. *Cancer Discovery*, *11*(12), 3048–3063. https://doi.org/10.1158/2159-8290.CD-21-0276

Petrackova, A., Turcsanyi, P., Papajik, T., & Kriegova, E. (2021). Revisiting Richter transformation in the era of novel CLL agents. *Blood Reviews*, *49*, 100824. https://doi.org/10.1016/J.BLRE.2021.100824

Pich, O., Cortes-Bullich, A., Muiños, F., … Lopez-Bigas, N. (2021). The evolution of hematopoietic cells under cancer therapy. *Nature Communications*, *12*(1), 4803. https://doi.org/10.1038/s41467-021-24858-3

Pich, O., Muiños, F., Lolkema, M. P., … Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. *Nature Genetics 2019 51:12*, *51*(12), 1732–1740. https://doi.org/10.1038/s41588-019-0525-5

Pieper, K., Grimbacher, B., & Eibel, H. (2013). B-cell biology and development. *Journal of Allergy and Clinical Immunology*, *131*(4), 959–971. https://doi.org/10.1016/J.JACI.2013.01.046

Pishvaian, M. J., Blais, E. M., Brody, J. R., … Petricoin, E. F. (2020). Overall survival in patients with pancreatic cancer receiving matched therapies following molecular profiling: a retrospective analysis of the Know Your Tumor registry trial. *The Lancet Oncology*, *21*(4), 508–518. https://doi.org/10.1016/S1470-2045(20)30074-7

Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature*, *569*(7756), 445–447. https://doi.org/10.1038/D41586-019-01506-X

Puente, X. S., Beà, S., Valdés-Mas, R., … Campo, E. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, *526*(7574), 519–524. https://doi.org/10.1038/nature14666

Puente, X. S., Pinyol, M., Quesada, V., … Campo, E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, *475*(7354), 101–105. https://doi.org/10.1038/nature10113

Queirós, A. C., Villamor, N., Clot, G., … Martín-Subero, J. I. (2014). A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia 2015 29:3*, *29*(3), 598–605. https://doi.org/10.1038/LEU.2014.252

Quesada, V., Conde, L., Villamor, N., … López-Otín, C. (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics*, *44*(1), 47–52. https://doi.org/10.1038/ng.1032

Raine, K. M., Hinton, J., Butler, A. P., … Campbell, P. J. (2015). cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Current Protocols in Bioinformatics*, *52*, 15.7.1-12. https://doi.org/10.1002/0471250953.bi1507s52

Raine, K. M., Van Loo, P., Wedge, D. C., … Campbell, P. J. (2016). ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current Protocols in Bioinformatics*, *56*(1), 15.9.1-15.9.17. https://doi.org/10.1002/cpbi.17

Rajbhandari, P., Lopez, G., Capdevila, C., … Califano, A. (2018). Cross-Cohort Analysis Identifies a TEAD4–MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. *Cancer Discovery*, *8*(5), 582–599. https://doi.org/10.1158/2159-8290.CD-16-0861

Ramsay, A. J., Quesada, V., Foronda, M., … López-Otín, C. (2013). POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nature Genetics 2013 45:5*, *45*(5), 526–530. https://doi.org/10.1038/NG.2584

Rausch, T., Zichner, T., Schlattl, A., … Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Reddy, E. P., Reynolds, R. K., Santos, E., & Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature 1982 300:5888*, *300*(5888), 149–152. https://doi.org/10.1038/300149A0

Rehm, H. L., Page, A. J. H., Smith, L., … Birney, E. (2021). GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*, *1*(2), 100029. https://doi.org/10.1016/J.XGEN.2021.100029

Reiman, A., Srinivasan, V., Barone, G., … Taylor, A. M. (2011). Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. *British Journal of Cancer*, *105*(4), 586–591. https://doi.org/10.1038/bjc.2011.266

Rheinbay, E., Nielsen, M. M., Abascal, F., … Zamora, J. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature 2020 578:7793*, *578*(7793), 102–111. https://doi.org/10.1038/S41586-020-1965-X

Richter, M. N. (1928). Generalized Reticular Cell Sarcoma of Lymph Nodes Associated with Lymphatic Leukemia. *The American Journal of Pathology*, *4*(4), 285-292.7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2006994/

Rimmer, A., Phan, H., Mathieson, I., … Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Rivas-Delgado, A., Nadeu, F., Enjuanes, A., … López-Guillermo, A. (2021).

Mutational Landscape and Tumor Burden Assessed by Cell-free DNA in Diffuse Large B-Cell Lymphoma in a Population-Based Study. *Clinical Cancer Research*, *27*(2), 513–521. https://doi.org/10.1158/1078-0432.CCR-20-2558

Roberts, A. W., Davids, M. S., Pagel, J. M., … Seymour, J. F. (2016). Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *374*(4), 311–322. https://doi.org/10.1056/NEJMoa1513257

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., … Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rodríguez, D., Bretones, G., Quesada, V., … López-Otín, C. (2015). Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood*, *126*(2), 195–202. https://doi.org/10.1182/BLOOD-2014-10-604959

Rosati, E., Baldoni, S., De Falco, F., … Sportoletti, P. (2018). NOTCH1 aberrations in Chronic lymphocytic leukemia. *Frontiers in Oncology*, *8*(JUN), 229. https://doi.org/10.3389/FONC.2018.00229/FULL

Rosenquist, R., Ghia, P., Hadzidimitriou, A., … Stamatopoulos, K. (2017). Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. *Leukemia 2017 31:7*, *31*(7), 1477–1481. https://doi.org/10.1038/leu.2017.125

Rossi, D., Rasi, S., Spina, V., … Gaidano, G. (2012). Different impact of NOTCH1 and SF3B1 mutations on the risk of chronic lymphocytic leukemia transformation to Richter syndrome. *British Journal of Haematology*, *158*(3), 426–429. https://doi.org/10.1111/J.1365-2141.2012.09155.X

Rossi, D., Spina, V., Cerri, M., … Gaidano, G. (2009). Stereotyped B-Cell Receptor Is an Independent Risk Factor of Chronic Lymphocytic Leukemia Transformation to Richter Syndrome. *Clinical Cancer Research*, *15*(13), 4415–4422. https://doi.org/10.1158/1078-0432.CCR-08-3266

Rossi, D., Spina, V., Deambrogi, C., … Gaidano, G. (2011). The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood*, *117*(12), 3391–3401. https://doi.org/10.1182/blood-2010-09-302174

Rustad, E. H., Yellapantula, V., Leongamornlert, D., … Maura, F. (2020). Timing the initiation of multiple myeloma. *Nature Communications*, *11*(1), 1917. https://doi.org/10.1038/s41467-020-15740-9

Sanchez-Vega, F., Mina, M., Armenia, J., … Mariamidze, A. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, *173*(2), 321-337.e10. https://doi.org/10.1016/j.cell.2018.03.035

Scandurra, M., Rossi, D., Deambrogi, C., … Bertoni, F. (2010). Genomic profiling of Richter's syndrome: recurrent lesions and differences with de novo diffuse

large B-cell lymphomas. *Hematological Oncology*, *28*(2), 62–67. https://doi.org/10.1002/hon.932

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Molecular Systems Biology*, *8*(1), 612. https://doi.org/10.1038/msb.2012.47

Scheffold, A., & Stilgenbauer, S. (2020). Revolution of Chronic Lymphocytic Leukemia Therapy: the Chemo-Free Treatment Paradigm. *Current Oncology Reports*, *22*(2), 1–10. https://doi.org/10.1007/S11912-020-0881-4/FIGURES/2

Schilsky, R. L. (2014). Implementing personalized cancer care. *Nature Reviews Clinical Oncology 2014 11:7*, *11*(7), 432–438. https://doi.org/10.1038/nrclinonc.2014.54

Schwaederlé, M., Ghia, E., Rassenti, L. Z., … Kipps, T. J. (2013). Subclonal evolution involving SF3B1 mutations in chronic lymphocytic leukemia. *Leukemia 2013 27:5*, *27*(5), 1214–1217. https://doi.org/10.1038/LEU.2013.22

Schwaederle, M., Zhao, M., Lee, J. J., … Kurzrock, R. (2015). Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *Journal of Clinical Oncology*, *33*(32), 3817–3825. https://doi.org/10.1200/JCO.2015.61.5997

Seifert, M., Sellmann, L., Bloehdorn, J., … Küppers, R. (2012). Cellular origin and pathophysiology of chronic lymphocytic leukemia. *Journal of Experimental Medicine*, *209*(12), 2183–2198. https://doi.org/10.1084/JEM.20120833

Sellick, G. S., Catovsky, D., & Houlston, R. S. (2006). Familial Chronic Lymphocytic Leukemia. *Seminars in Oncology*, *33*(2), 195–201. https://doi.org/10.1053/J.SEMINONCOL.2006.01.013

Sentís, I., Gonzalez, S., Genescà, E., … Lopez-Bigas, N. (2020). The evolution of relapse of adult T cell acute lymphoblastic leukemia. *Genome Biology*, *21*(1), 284. https://doi.org/10.1186/s13059-020-02192-z

Shand, M., Soto, J., Lichtenstein, L., … Banks, E. (2020). A validated lineage-derived somatic truth data set enables benchmarking in cancer genome analysis. *Communications Biology 2020 3:1*, *3*(1), 1–8. https://doi.org/10.1038/s42003-020-01460-9

Shen, R., & Seshan, V. E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, *44*(16), e131–e131. https://doi.org/10.1093/NAR/GKW520

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shuai, S., Suzuki, H., Diaz-Navarro, A., … Stein, L. D. (2019). The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature 2019 574:7780*, *574*(7780), 712–716. https://doi.org/10.1038/S41586-019-1651-Z

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biological Procedures Online*, *15*(1), 4. https://doi.org/10.1186/1480-9222-15-4

Skowronska, A., Austen, B., Powell, J. E., … Stankovic, T. (2012). ATM germline heterozygosity does not play a role in chronic lymphocytic leukemia initiation but influences rapid disease progression through loss of the remaining ATM allele. *Haematologica*, *97*(1), 142–146. https://doi.org/10.3324/HAEMATOL.2011.048827

Soneson, C., Love, M. I., Robinson, M. D., & Floor, S. N. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research 2015 4:1521*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.1

Speedy, H. E., Beekman, R., Chapaprieta, V., … Martín-Subero, J. I. (2019). Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nature Communications 2019 10:1*, *10*(1), 1–9. https://doi.org/10.1038/S41467-019-11582-2

Speedy, H. E., Kinnersley, B., Chubb, D., … Houlston, R. S. (2016). Germ line mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia. *Blood*, *128*(19), 2319–2326. https://doi.org/10.1182/BLOOD-2016-01-695692

Sportoletti, P., Baldoni, S., Cavalli, L., … Falzetti, F. (2010). NOTCH1 PEST domain mutation is an adverse prognostic factor in B-CLL. *British Journal of Haematology*, *151*(4), 404–406. https://doi.org/10.1111/J.1365-2141.2010.08368.X

Stamatopoulos, K., Agathangelidis, A., Rosenquist, R., & Ghia, P. (2016). Antigen receptor stereotypy in chronic lymphocytic leukemia. *Leukemia 2017 31:2*, *31*(2), 282–291. https://doi.org/10.1038/leu.2016.322

Stankovic, T., & Skowronska, A. (2014). The role of ATM mutations and 11q deletions in disease progression in chronic lymphocytic leukemia. *Leukemia & Lymphoma*, *55*(6), 1227–1239. https://doi.org/10.3109/10428194.2013.829919

Stephens, P. J., Greenman, C. D., Fu, B., … Campbell, P. J. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, *144*(1), 27–40. https://doi.org/10.1016/j.cell.2010.11.055

Stevenson, F. K., Krysov, S., Davies, A. J., … Packham, G. (2011). B-cell receptor signaling in chronic lymphocytic leukemia. *Blood*, *118*(16), 4313–4320. https://doi.org/10.1182/BLOOD-2011-06-338855

Stewart, G. S., Last, J. I. K., Stankovic, T., … Taylor, A. M. R. (2001). Residual Ataxia Telangiectasia Mutated Protein Function in Cells from Ataxia Telangiectasia Patients, with 5762ins137 and 7271T→G Mutations, Showing a Less Severe Phenotype. *Journal of Biological Chemistry*, *276*(32), 30133–30141. https://doi.org/10.1074/jbc.M103160200

Stilgenbauer, S., Schnaiter, A., Paschka, P., … Döhner, H. (2014). Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the

CLL8 trial. *Blood*, *123*(21), 3247–3254. https://doi.org/10.1182/blood-2014-01-546150

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature 2009 458:7239*, *458*(7239), 719–724. https://doi.org/10.1038/nature07943

Sudmant, P. H., Rausch, T., Gardner, E. J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Supek, F., Miñana, B., Valcárcel, J., … Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, *156*(6), 1324–1335. https://doi.org/10.1016/j.cell.2014.01.051

Tai Fang, L., Zhu, B., Zhao, Y., … Zhang, C. (2021). Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nature Biotechnology 2021 39:9*, *39*(9), 1151–1160. https://doi.org/10.1038/S41587-021-00993-6

Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, *12*(4), e1004873. https://doi.org/10.1371/journal.pcbi.1004873

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., … Lopez-Bigas, N. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, *10*(1), 25. https://doi.org/10.1186/s13073-018-0531-8

Tarabichi, M., Martincorena, I., Gerstung, M., … Lingjærde, O. C. (2018). Neutral tumor evolution? *Nature Genetics*, *50*(12), 1630. https://doi.org/10.1038/S41588-018-0258-X

Tate, J. G., Bamford, S., Jubb, H. C., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Tiao, G., Improgo, M. R., Kasar, S., … Brown, J. R. (2017). Rare germline variants in ATM are associated with chronic lymphocytic leukemia. *Leukemia*, *31*(10), 2244–2247. https://doi.org/10.1038/leu.2017.201

Uffelmann, E., Huang, Q. Q., Munung, N. S., … Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers 2021 1:1*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

Vangapandu, H. V., Alston, B., Morse, J., … Gandhi, V. (2018). Biological and metabolic effects of IACS-010759, an OxPhos inhibitor, on chronic lymphocytic leukemia cells. *Oncotarget*, *9*(38), 24980–24991. https://doi.org/10.18632/oncotarget.25166

Varano, G., Raffel, S., Sormani, M., … Casola, S. (2017). The B-cell receptor controls fitness of MYC-driven lymphoma cells via GSK3β inhibition. *Nature*, *546*(7657), 302–306. https://doi.org/10.1038/nature22353

Vendramin, R., Litchfield, K., & Swanton, C. (2021). Cancer evolution: Darwin and

beyond. *The EMBO Journal*, *40*(18), e108389. https://doi.org/10.15252/EMBJ.2021108389

Venkatesan, S., & Swanton, C. (2016). Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. *Https://Doi.Org/10.1200/EDBK_158930*, *36*, e141–e149. https://doi.org/10.1200/EDBK_158930

Villamor, N., Conde, L., Martínez-Trillos, A., … López-Guillermo, A. (2012). NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia 2013 27:5*, *27*(5), 1100–1106. https://doi.org/10.1038/leu.2012.357

Volkova, N. V., Meier, B., González-Huici, V., … Gerstung, M. (2020). Mutational signatures are jointly shaped by DNA damage and repair. *Nature Communications 2020 11:1*, *11*(1), 1–15. https://doi.org/10.1038/s41467-020-15912-7

Voss, K., Auwera, G. Van der, Gentry, J., … Gentry, J. (2017). Full-stack genomics pipelining with GATK4 + WDL + Cromwell . *F1000Research*, *6*. https://doi.org/10.7490/F1000RESEARCH.1114634.1

Wagle, N., Emery, C., Berger, M. F., … Garraway, L. A. (2011). Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of Clinical Oncology*, *29*(22), 3085–3096. https://doi.org/10.1200/JCO.2010.33.2312

Wala, J. A., Bandopadhayay, P., Greenwald, N. F., … Beroukhim, R. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Research*, *28*(4), 581–591. https://doi.org/10.1101/gr.221028.117

Walter, K., Min, J. L., Huang, J., … Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–89. https://doi.org/10.1038/nature14962

Wang, L., Brooks, A. N., Fan, J., … Wu, C. J. (2016). Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell*, *30*(5), 750–763. https://doi.org/10.1016/J.CCELL.2016.10.005

Wang, M., Luo, W., Jones, K., … Zhu, B. (2020). SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Scientific Reports*, *10*(1), 12898. https://doi.org/10.1038/s41598-020-69772-8

Wierda, W. G., O'Brien, S., Wang, X., … Keating, M. J. (2011). Multivariable Model for Time to First Treatment in Patients With Chronic Lymphocytic Leukemia. *Journal of Clinical Oncology*, *29*(31), 4088–4095. https://doi.org/10.1200/JCO.2010.33.9002

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/SDATA.2016.18

Williams, M. J., Werner, B., Barnes, C. P., … Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature Genetics 2015 48:3*, *48*(3), 238–244. https://doi.org/10.1038/ng.3489

Wilm, A., Aw, P. P. K., Bertrand, D., … Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. https://doi.org/10.1093/nar/gks918

Worthey, E. A., Mayer, A. N., Syverson, G. D., … Dimmock, D. P. (2011). Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, *13*(3), 255–262. https://doi.org/10.1097/GIM.0B013E3182088158

Woyach, J. A., Furman, R. R., Liu, T.-M., … Byrd, J. C. (2014). Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib. *New England Journal of Medicine*, *370*(24), 2286–2294. https://doi.org/10.1056/NEJMoa1400029

Xu, J., Yang, P., Xue, S., … Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. In *Human Genetics* (Vol. 138, Issue 2, pp. 109–124). Springer Verlag. https://doi.org/10.1007/s00439-019-01970-5

Yang, F., Brady, S. W., Tang, C., … Zhou, B.-B. S. (2021). Chemotherapy and mismatch repair deficiency cooperate to fuel TP53 mutagenesis and ALL relapse. *Nature Cancer*, *2*(8), 819–834. https://doi.org/10.1038/s43018-021-00230-8

Yi, K., & Ju, Y. S. (2018). Patterns and mechanisms of structural variations in human cancer. *Experimental & Molecular Medicine 2018 50:8*, *50*(8), 1–11. https://doi.org/10.1038/s12276-018-0112-3

Yin, S., Gambe, R. G., Sun, J., … Wang, L. (2019). A Murine Model of Chronic Lymphocytic Leukemia Based on B Cell-Restricted Expression of Sf3b1 Mutation and Atm Deletion. *Cancer Cell*, *35*(2), 283-296.e5. https://doi.org/10.1016/j.ccell.2018.12.013

Young, E., Noerenberg, D., Mansouri, L., … Damm, F. (2016). EGR2 mutations define a new clinically aggressive subgroup of chronic lymphocytic leukemia. *Leukemia 2017 31:7*, *31*(7), 1547–1554. https://doi.org/10.1038/LEU.2016.359

Yuille, M. R., Condie, A., Hudson, C. D., … Houlston, R. S. (2002). ATM mutations are rare in familial chronic lymphocytic leukemia. *Blood*, *100*(2), 603–609. https://doi.org/10.1182/BLOOD.V100.2.603

Zenz, T., Eichhorst, B., Busch, R., … Stilgenbauer, S. (2010). TP53 mutation and survival in chronic lymphocytic leukemia. *Journal of Clinical Oncology*, *28*(29), 4473–4479. https://doi.org/10.1200/JCO.2009.27.8762

Zhang, C. Z., Spektor, A., Cornils, H., … Pellman, D. (2015). Chromothripsis from

DNA damage in micronuclei. *Nature 2015 522:7555*, *522*(7555), 179–184. https://doi.org/10.1038/nature14493

Zhang, L., Yao, Y., Zhang, S., … Wang, M. (2019). Metabolic reprogramming toward oxidative phosphorylation identifies a therapeutic target for mantle cell lymphoma. *Science Translational Medicine*, *11*(491), eaau1167. https://doi.org/10.1126/scitranslmed.aau1167

Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, *35*(12), 2084–2092. https://doi.org/10.1093/bioinformatics/bty895

Some figures have been designed using resources from Flaticon.com and Freepik, and the cover contains Human Vectors by Vecteezy.

8 Appendix

8.1 List of co-author publications

This section contains the list of additional publications where I contributed, by virtue of the expertise I acquired, and the application of the methodologies produced during this thesis.

A practical guide for mutational signature analysis in hematological malignancies. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, **Royo R**, Ziccheddu B, Puente XS, Avet-Loiseau H, Campbell PJ, Nik-Zainal S, Campo E, Munshi N, Bolli N. *Nat Commun. 2019* Jul 5;10(1):2969. doi: 10.1038/s41467-019-11037-8.

Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. Nadeu F, Martin-Garcia D, Clot G, Díaz-Navarro A, Duran-Ferrer M, Navarro A, Vilarrasa-Blasi R, Kulis M, **Royo R**, Gutiérrez-Abril J, Valdés-Mas R, López C, Chapaprieta V, Puiggros M, Castellano G, Costa D, Aymerich M, Jares P, Espinet B, Muntañola A, Ribera-Cortada I, Siebert R, Colomer D, Torrents D, Gine E, López-Guillermo A, Küppers R, Martin-Subero JI, Puente XS, Beà S, Campo E. *Blood. 2020* Sep 17;136(12):1419-1432. doi: 10.1182/blood.2020005289.

Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis. Nadeu F, **Royo R**, Maura F, Dawson KJ, Dueso-Barroso A, Aymerich M, Pinyol M, Beà S, López-Guillermo A, Delgado J, Puente XS, Campo E. *Leukemia. 2020* Jul;34(7):1929-1933. doi: 10.1038/s41375-020-0730-3.

The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Duran-Ferrer M, Clot G, Nadeu F, Beekman R, Baumann T, Nordlund J, Marincevic-Zuniga Y, Lönnerholm G, Rivas-Delgado A, Martín S,

Ordoñez R, Castellano G, Kulis M, Queirós AC, Lee ST, Wiemels J, Royo R, Puiggrós M, Lu J, Giné E, Beà S, Jares P, Agirre X, Prosper F, López-Otín C, Puente XS, Oakes CC, Zenz T, Delgado J, López-Guillermo A, Campo E, Martín-Subero JI. *Nat Cancer. 2020* Nov;1(11):1066-1081. doi: 10.1038/s43018-020-00131-2.

IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. Nadeu F, Mas-de-Les-Valls R, Navarro A, Royo R, Martín S, Villamor N, Suárez-Cisneros H, Mares R, Lu J, Enjuanes A, Rivas-Delgado A, Aymerich M, Baumann T, Colomer D, Delgado J, Morin RD, Zenz T, Puente XS, Campbell PJ, Beà S, Maura F, Campo E. *Nat Commun. 2020* Jul 7;11(1):3390. doi: 10.1038/s41467-020-17095-7.

IGLV3-21R110 identifies an aggressive biological subtype of chronic lymphocytic leukemia with intermediate epigenetics. Nadeu F, Royo R, Clot G, Duran-Ferrer M, Navarro A, Martín S, Lu J, Zenz T, Baumann T, Jares P, Puente XS, Martín-Subero JI, Delgado J, Campo E. *Blood. 2021* May 27;137(21):2935-2946. doi: 10.1182/blood.2020008311.

Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. Reijns MAM, Parry DA, Williams TC, Nadeu F, Hindshaw RL, Rios Szwed DO, Nicholson MD, Carroll P, Boyle S, Royo R, Cornish AJ, Xiang H, Ridout K; Genomics England Research Consortium; Colorectal Cancer Domain UK 100,000 Genomes Project, Schuh A, Aden K, Palles C, Campo E, Stankovic T, Taylor MS, Jackson AP. *Nature. 2022* Feb;602(7898):623-631. doi: 10.1038/s41586-022-04403-y.

Disease-specific U1 spliceosomal RNA mutations in mature B-cell neoplasms. Nadeu F*, Shuai S*, Clot G*, Hilton L, Diaz-Navarro A, Martín S, Royo R, Baumann T, Kulis M, López-Oreja I, Cossio M, Lu J, Ljungström V, Young E, Plevova K, Knisbacher B, Lin Z, Hahn C, Bousquets P, Alcoceba M, González M, Colado E, Payer A, Aymerich M, Terol M, Rivas-Delgado A, Enjuanes A, Ruiz-Gaspà S,

Chatzikonstantinou T, Hägerstrand D, Jylhä C, Skaftason A, Mansouri L, Doubek M, J. van Gastel-Mol E, Davis Z, Walewska R, Scarfò L, Trentin L, Visentin A, Parikh S, Rabe K, Moia R, Armand M, Rossi D, Davi F, Gaidano G, Kay N, Shanafelt T, Ghia P, Oscier D, Langerak A, Beà S, López-Guillermo A, Neuberg D, Wu C, Getz G, Pospisilova S, Stamatopoulos K, Rosenquist R, Huber W, Zenz T, Colomer D, Martín-Subero I, Delgado J, Morin R, Stein L, Puente XS, Campo E. *Under review.*

Germline genetic variants that affect Wnt signaling as cause of serrated and adenomatous polyposis. Isabel Quintana, Mariona Terradas, Pilar Mur, Iris te Paske, Claudia Maestro, David Torrents, Montserrat Puiggròs, **Romina Royo**, Raul Tonda, Genís Parra, Davide Piscia, Sergi Beltrán, Matilde Navarro, Virginia Piñol, Joan Brunet, Noemi Gonzalez-Abuin, Gemma Aiza, Sophia Peters, Verena Steinke-Lange, Anna Sommer, Isabel Spier, Yasmijn van Herwaarden Galuh Astuti, Elke Hollinski, Nicoline Hoogerbrugge, Richarda de Voer, Stefan Aretz, Gabriel Capellá, Laura Valle. *Under review.*

8.2 Publications included in the Thesis

Pan-cancer analysis of whole genomes. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (includes **Royo R, 39 of 1341**). *Nature. 2020 Feb;*578(7793):82-93. doi: 10.1038/s41586-020-1969-6

Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. Nadeu F*, **Royo R***, Massoni-Badosa R*, Playa-Albinyana H*, Garcia-Torre B*, Duran-Ferrer M, Dawson KJ, Kulis M, Diaz-Navarro A, Villamor N, Melero JL, Chapaprieta V, Dueso-Barroso A, Delgado J, Moia R, Ruiz-Gil S, Marchese D, Giró A, Verdaguer-Dot N, Romo M, Clot G, Rozman M, Frigola G, Rivas-Delgado A, Baumann T, Alcoceba M, González M, Climent F, Abrisqueta P, Castellví J, Bosch F, Aymerich M, Enjuanes A, Ruiz-Gaspà S, López-Guillermo A, Jares P, Beà S, Capella-Gutierrez S, Gelpí JL, López-Bigas N, Torrents D, Campbell PJ, Gut I, Rossi D, Gaidano G, Puente XS, Garcia-Roves PM, Colomer D, Heyn H, Maura F, Martín-Subero JI, Campo E. *Nat Med. 2022 Aug;*28(8):1662-1671. doi: 10.1038/s41591-022-01927-8

* Equal contribution

ATM germline variants in a young adult with chronic lymphocytic leukemia: 8years of genomic evolution. Royo R*, Magnano L*, Delgado J, Ruiz-Gil S, Gelpí JL, Heyn H, Taylor MA, Stankovic T, Puente XS, Nadeu F, Campo E. *Blood Cancer J. 2022 Jun 7;*12(6):90. doi: 10.1038/s41408-022-00686-6

* Equal contribution

Article

Pan-cancer analysis of whole genomes

https://doi.org/10.1038/s41586-020-1969-6

Received: 29 July 2018

Accepted: 11 December 2019

Published online: 5 February 2020

Open access

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Cancer is driven by genetic change, and the advent of massively parallel sequencing has enabled systematic documentation of this variation at the whole-genome scale[1–3]. Here we report the integrative analysis of 2,658 whole-cancer genomes and their matching normal tissues across 38 tumour types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). We describe the generation of the PCAWG resource, facilitated by international data sharing using compute clouds. On average, cancer genomes contained 4–5 driver mutations when combining coding and non-coding genomic elements; however, in around 5% of cases no drivers were identified, suggesting that cancer driver discovery is not yet complete. Chromothripsis, in which many clustered structural variants arise in a single catastrophic event, is frequently an early event in tumour evolution; in acral melanoma, for example, these events precede most somatic point mutations and affect several cancer-associated genes simultaneously. Cancers with abnormal telomere maintenance often originate from tissues with low replicative activity and show several mechanisms of preventing telomere attrition to critical levels. Common and rare germline variants affect patterns of somatic mutation, including point mutations, structural variants and somatic retrotransposition. A collection of papers from the PCAWG Consortium describes non-coding mutations that drive cancer beyond those in the *TERT* promoter[4]; identifies new signatures of mutational processes that cause base substitutions, small insertions and deletions and structural variation[5,6]; analyses timings and patterns of tumour evolution[7]; describes the diverse transcriptional consequences of somatic mutation on splicing, expression levels, fusion genes and promoter activity[8,9]; and evaluates a range of more-specialized features of cancer genomes[8,10–18].

Cancer is the second most-frequent cause of death worldwide, killing more than 8 million people every year; the incidence of cancer is expected to increase by more than 50% over the coming decades[19,20]. 'Cancer' is a catch-all term used to denote a set of diseases characterized by autonomous expansion and spread of a somatic clone. To achieve this behaviour, the cancer clone must co-opt multiple cellular pathways that enable it to disregard the normal constraints on cell growth, modify the local microenvironment to favour its own proliferation, invade through tissue barriers, spread to other organs and evade immune surveillance[21]. No single cellular program directs these behaviours. Rather, there is a large pool of potential pathogenic abnormalities from which individual cancers draw their own combinations: the commonalities of macroscopic features across tumours belie a vastly heterogeneous landscape of cellular abnormalities.

This heterogeneity arises from the stochastic nature of Darwinian evolution. There are three preconditions for Darwinian evolution: characteristics must vary within a population; this variation must be heritable from parent to offspring; and there must be competition for survival within the population. In the context of somatic cells, heritable variation arises from mutations acquired stochastically throughout life, notwithstanding additional contributions from germline and epigenetic variation. A subset of these mutations alter the cellular phenotype, and a small subset of those variants confer an advantage on clones during the competition to escape the tight physiological controls wired into somatic cells. Mutations that provide a selective advantage to the clone are termed driver mutations, as opposed to selectively neutral passenger mutations.

Initial studies using massively parallel sequencing demonstrated the feasibility of identifying every somatic point mutation, copy-number change and structural variant (SV) in a given cancer[1–3]. In 2008, recognizing the opportunity that this advance in technology provided, the global cancer genomics community established the ICGC with the goal of systematically documenting the somatic mutations that drive common tumour types[22].

The pan-cancer analysis of whole genomes

The expansion of whole-genome sequencing studies from individual ICGC and TCGA working groups presented the opportunity to undertake a meta-analysis of genomic features across tumour types. To achieve this, the PCAWG Consortium was established. A Technical Working Group implemented the informatics analyses by aggregating the raw sequencing data from different working groups that studied individual tumour types, aligning the sequences to the human genome and delivering a set of high-quality somatic mutation calls for downstream analysis (Extended Data Fig. 1). Given the recent meta-analysis

A list of members and their affiliations appears in the online version of the paper and lists of working groups appear in the Supplementary Information.

Box 1

Online resources for data access, visualization and analysis

The PCAWG landing page (http://docs.icgc.org/pcawg) provides links to several data resources for interactive online browsing, analysis and download of PCAWG data and results (Supplementary Table 4).

Direct download of PCAWG data

Aligned PCAWG read data in BAM format are also available at the European Genome Phenome Archive (EGA; https://www.ebi.ac.uk/ega/search/site/pcawg under accession number EGAS00001001692). In addition, all open-tier PCAWG genomics data, as well as reference datasets used for analysis, can be downloaded from the ICGC Data Portal at http://docs.icgc.org/pcawg/data/. Controlled-tier genomic data, including SNVs and indels that originated from TCGA projects (in VCF format) and aligned reads (in BAM format) can be downloaded using the Score (https://www.overture.bio/) software package, which has accelerated and secure file transfer, as well as BAM slicing facilities to selectively download defined regions of genomic alignments.

PCAWG computational pipelines

The core alignment, somatic variant-calling, quality-control and variant consensus-generation pipelines used by PCAWG have each been packaged into portable cross-platform images using the Dockstore system[84] and released under an Open Source licence that enables unrestricted use and redistribution. All PCAWG Dockstore images are available to the public at https://dockstore.org/organizations/PCAWG/collections/PCAWG.

ICGC Data Portal

The ICGC Data Portal[85] (https://dcc.icgc.org) serves as the main entry point for accessing PCAWG datasets with a single uniform web interface and a high-performance data-download client. This uniform interface provides users with easy access to the myriad of PCAWG sequencing data and variant calls that reside in many repositories and compute clouds worldwide. Streaming technology[86] provides users with high-level visualizations in real time of BAM and VCF files stored remotely on the Cancer Genome Collaboratory.

UCSC Xena

UCSC Xena[87] (https://pcawg.xenahubs.net) visualizes all PCAWG primary results, including copy-number, gene-expression, gene-fusion and promoter-usage alterations, simple somatic mutations, large somatic structural variations, mutational signatures and phenotypic data. These open-access data are available through a public Xena hub, and consensus simple somatic mutations can be loaded to the local computer of a user via a private Xena hub. Kaplan–Meier plots, histograms, box plots, scatter plots and transcript-specific views offer additional visualization options and statistical analyses.

The Expression Atlas

The Expression Atlas (https://www.ebi.ac.uk/gxa/home) contains RNA-sequencing and expression microarray data for querying gene expression across tissues, cell types, developmental stages and/or experimental conditions[88]. Two different views of the data are provided: summarized expression levels for each tumour type and gene expression at the level of individual samples, including reference-gene expression datasets for matching normal tissues.

PCAWG Scout

PCAWG Scout (http://pcawgscout.bsc.es/) provides a framework for -omics workflow and website templating to generate on-demand, in-depth analyses of the PCAWG data that are openly available to the whole research community. Views of protected data are available that still safeguard sensitive data. Through the PCAWG Scout web interface, users can access an array of reports and visualizations that leverage on-demand bioinformatic computing infrastructure to produce results in real time, allowing users to discover trends as well as form and test hypotheses.

Chromothripsis Explorer

Chromothripsis Explorer (http://compbio.med.harvard.edu/chromothripsis/) is a portal that allows structural variation in the PCAWG dataset to be explored on an individual patient basis through the use of circos plots. Patterns of chromothripsis can also be explored in aggregated formats.

of exome data from the TCGA Pan-Cancer Atlas[23–25], scientific working groups concentrated their efforts on analyses best-informed by whole-genome sequencing data.

We collected genome data from 2,834 donors (Extended Data Table 1), of which 176 were excluded after quality assurance. A further 75 had minor issues that could affect some of the analyses (grey-listed donors) and 2,583 had data of optimal quality (white-listed donors) (Supplementary Table 1). Across the 2,658 white- and grey-listed donors, whole-genome sequencing data were available from 2,605 primary tumours and 173 metastases or local recurrences. Mean read coverage was 39× for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38× and 60× (Supplementary Fig. 1). RNA-sequencing data were available for 1,222 donors. The final cohort comprised 1,469 men (55%) and 1,189 women (45%), with a mean age of 56 years (range, 1–90 years) across 38 tumour types (Extended Data Table 1 and Supplementary Table 1).

To identify somatic mutations, we analysed all 6,835 samples using a uniform set of algorithms for alignment, variant calling and quality control (Extended Data Fig. 1, Supplementary Fig. 2 and Supplementary Methods 2). We used three established pipelines to call somatic single-nucleotide variations (SNVs), small insertions and deletions (indels), copy-number alterations (CNAs) and SVs. Somatic retrotransposition events, mitochondrial DNA mutations and telomere lengths were also called by bespoke algorithms. RNA-sequencing data were uniformly processed to call transcriptomic alterations. Germline variants identified by the three separate pipelines included single-nucleotide polymorphisms, indels, SVs and mobile-element insertions (Supplementary Table 2).

The requirement to uniformly realign and call variants on approximately 5,800 whole genomes presented considerable computational challenges, and raised ethical issues owing to the use of data from different jurisdictions (Extended Data Table 2). We used cloud computing[26,27] to distribute alignment and variant calling across 13 data centres on 3 continents (Supplementary Table 3). Core pipelines were packaged into Docker containers[28] as reproducible, stand-alone packages, which we have made available for download. Data repositories for raw and derived datasets, together with portals for data visualization and exploration, have also been created (Box 1 and Supplementary Table 4).

Benchmarking of genetic variant calls

To benchmark mutation calling, we ran the 3 core pipelines, together with 10 additional pipelines, on 63 representative tumour–normal genome pairs (Supplementary Note 1). For 50 of these cases, we performed validation by hybridization of tumour and matched normal DNA to a custom bait set with deep sequencing[29]. The 3 core somatic variant-calling pipelines had individual estimates of sensitivity of 80–90% to detect a true somatic SNV called by any of the 13 pipelines; more

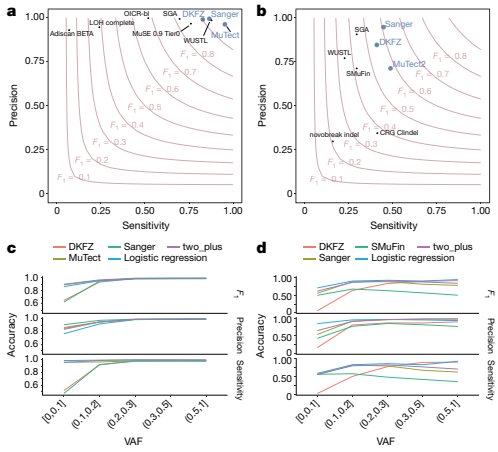


Fig. 1 | Validation of variant-calling pipelines in PCAWG. a, Scatter plot of estimated sensitivity and precision for somatic SNVs across individual algorithms assessed in the validation exercise across $n = 63$ PCAWG samples. Core algorithms included in the final PCAWG call set are shown in blue. **b**, Sensitivity and precision estimates across individual algorithms for somatic indels. **c**, Accuracy (precision, sensitivity and F_1 score, defined as $2 \times \text{sensitivity} \times \text{precision}/(\text{sensitivity} + \text{precision})$) of somatic SNV calls across variant allele fractions (VAFs) for the core algorithms. The accuracy of two methods of combining variant calls (two-plus, which was used in the final dataset, and logistic regression) is also shown. **d**, Accuracy of indel calls across variant allele fractions.

than 95% of SNV calls made by each of the core pipelines were genuine somatic variants (Fig. 1a). For indels—a more-challenging class of variants to identify with short-read sequencing—the 3 core algorithms had individual sensitivity estimates in the range of 40–50%, with precision of 70–95% (Fig. 1b). For individual SV algorithms, we estimated precision to be in the range 80–95% for samples in the 63-sample pilot dataset.

Next, we defined a strategy to merge results from the three pipelines into one final call-set to be used for downstream scientific analyses (Methods and Supplementary Note 2). Sensitivity and precision of consensus somatic variant calls were 95% (90% confidence interval, 88–98%) and 95% (90% confidence interval, 71–99%), respectively, for SNVs (Extended Data Fig. 2). For somatic indels, sensitivity and precision were 60% (34–72%) and 91% (73–96%), respectively (Extended Data Fig. 2). Regarding somatic SVs, we estimate the sensitivity of merged calls to be 90% for true calls generated by any one pipeline; precision was estimated as 97.5%. The improvement in calling accuracy from combining different pipelines was most noticeable in variants with low variant allele fractions, which probably originate from tumour subclones (Fig. 1c, d). Germline variant calls, phased using a haplotype-reference panel, displayed a precision of more than 99% and a sensitivity of 92–98% (Supplementary Note 2).

Analysis of PCAWG data

The uniformly generated, high-quality set of variant calls across more than 2,500 donors provided the springboard for a series of scientific working groups to explore the biology of cancer. A comprehensive suite of companion papers that describe the analyses and discoveries across these thematic areas is copublished with this paper[4–18] (Extended Data Table 3).

Pan-cancer burden of somatic mutations

Across the 2,583 white-listed PCAWG donors, we called 43,778,859 somatic SNVs, 410,123 somatic multinucleotide variants, 2,418,247 somatic indels, 288,416 somatic SVs, 19,166 somatic retrotransposition events and 8,185 de novo mitochondrial DNA mutations (Supplementary Table 1). There was considerable heterogeneity in the burden of somatic mutations across patients and tumour types, with a broad correlation in mutation burden among different classes of somatic variation (Extended Data Fig. 3). Analysed at a per-patient level, this correlation held, even when considering tumours with similar purity and ploidy (Supplementary Fig. 3). Why such correlation should apply on a pan-cancer basis is unclear. It is likely that age has some role, as we observe a correlation between most classes of somatic mutation and age at diagnosis (around 190 SNVs per year, $P = 0.02$; about 22 indels per year, $P = 5 \times 10^{-5}$; 1.5 SVs per year, $P < 2 \times 10^{-16}$; linear regression with likelihood ratio tests; Supplementary Fig. 4). Other factors are also likely to contribute to the correlations among classes of somatic mutation, as there is evidence that some DNA-repair defects can cause multiple types of somatic mutation[30], and a single carcinogen can cause a range of DNA lesions[31].

Panorama of driver mutations in cancer

We extracted the subset of somatic mutations in PCAWG tumours that have high confidence to be driver events on the basis of current knowledge. One challenge to pinpointing the specific driver mutations in an individual tumour is that not all point mutations in recurrently mutated cancer-associated genes are drivers[32]. For genomic elements significantly mutated in PCAWG data, we developed a 'rank-and-cut' approach to identify the probable drivers (Supplementary Methods 8.1). This approach works by ranking the observed mutations in a given genomic element based on recurrence, estimated functional consequence and expected pattern of drivers in that element. We then estimate the excess burden of somatic mutations in that genomic element above that expected for the background mutation rate, and cut the ranked mutations at this level. Mutations in each element with the highest driver ranking were then assigned as probable drivers; those below the threshold will probably have arisen through chance and were assigned as probable passengers. Improvements to features that are used to rank the mutations and the methods used to measure them will contribute to further development of the rank-and-cut approach.

We also needed to account for the fact that some bona fide cancer genomic elements were not rediscovered in PCAWG data because of low statistical power. We therefore added previously known cancer-associated genes to the discovery set, creating a 'compendium of mutational driver elements' (Supplementary Methods 8.2). Then, using stringent rules to nominate driver point mutations that affect these genomic elements on the basis of prior knowledge[33], we separated probable driver from passenger point mutations. To cover all classes of variant, we also created a compendium of known driver SVs, using analogous rules to identify which somatic CNAs and SVs are most likely to act as drivers in each tumour. For probable pathogenic germline variants, we identified all truncating germline point mutations and SVs that affect high-penetrance germline cancer-associated genes.

This analysis defined a set of mutations that we could confidently assert, based on current knowledge, drove tumorigenesis in the more than 2,500 tumours of PCAWG. We found that 91% of tumours had at least one identified driver mutation, with an average of 4.6 drivers per tumour identified, showing extensive variation across cancer types (Fig. 2a). For coding point mutations, the average was 2.6 drivers per tumour, similar to numbers estimated in known cancer-associated genes in tumours in the TCGA using analogous approaches[32].

To address the frequency of non-coding driver point mutations, we combined promoters and enhancers that are known targets of

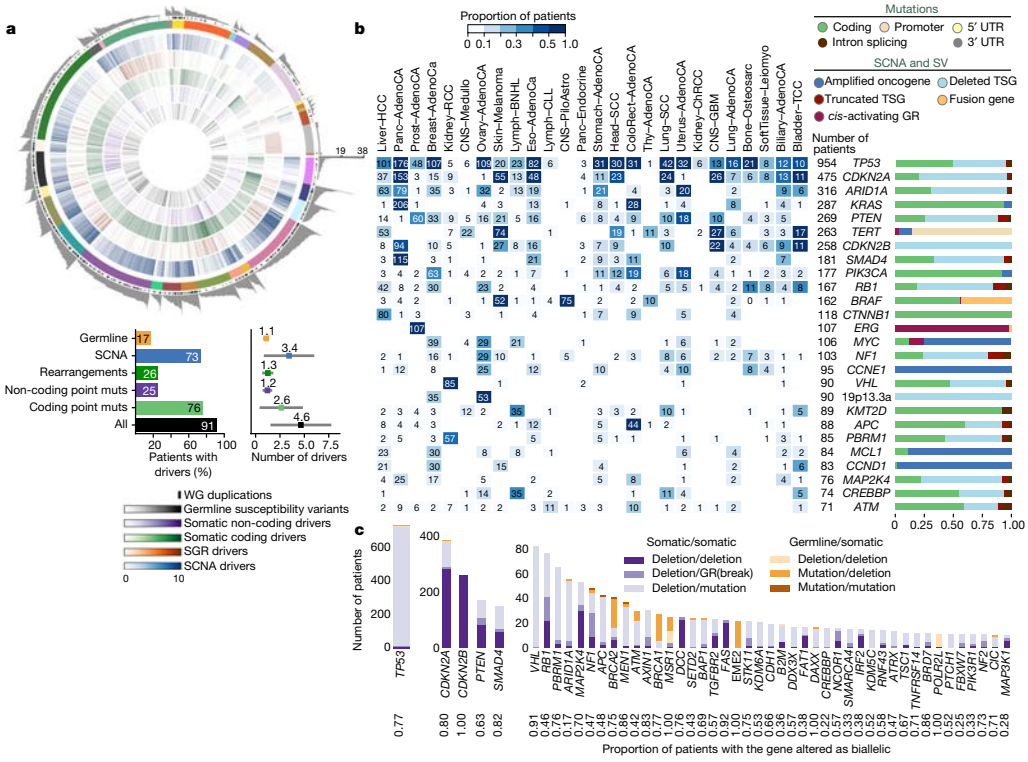


Fig. 2 | Panorama of driver mutations in PCAWG. a, Top, putative driver mutations in PCAWG, represented as a circos plot. Each sector represents a tumour in the cohort. From the periphery to the centre of the plot the concentric rings represent: (1) the total number of driver alterations; (2) the presence of whole-genome (WG) duplication; (3) the tumour type; (4) the number of driver CNAs; (5) the number of driver genomic rearrangements; (6) driver coding point mutations; (7) driver non-coding point mutations; and (8) pathogenic germline variants. Bottom, snapshots of the panorama of driver mutations. The horizontal bar plot (left) represents the proportion of patients with different types of drivers. The dot plot (right) represents the mean number of each type of driver mutation across tumours with at least one event (the square dot) and the standard deviation (grey whiskers), based on $n = 2,583$ patients. **b**, Genomic elements targeted by different types of mutations in the cohort altered in more than 65 tumours. Both germline and somatic variants are included. Left, the heat map shows the recurrence of alterations across cancer types. The colour indicates the proportion of mutated tumours and the number indicates the absolute count of mutated tumours. Right, the proportion of each type of alteration that affects each genomic element. **c**, Tumour-suppressor genes with biallelic inactivation in 10 or more patients. The values included under the gene labels represent the proportions of patients who have biallelic mutations in the gene out of all patients with a somatic mutation in that gene. GR, genomic rearrangement; SCNA, somatic copy-number alteration; SGR, somatic genome rearrangement; TSG, tumour suppressor gene; UTR, untranslated region.

non-coding drivers[34–37] with those newly discovered in PCAWG data; this is reported in a companion paper[4]. Using this approach, only 13% (785 out of 5,913) of driver point mutations were non-coding in PCAWG. Nonetheless, 25% of PCAWG tumours bear at least one putative non-coding driver point mutation, and one third (237 out of 785) affected the *TERT* promoter (9% of PCAWG tumours). Overall, non-coding driver point mutations are less frequent than coding driver mutations. With the exception of the *TERT* promoter, individual enhancers and promoters are only infrequent targets of driver mutations[4].

Across tumour types, SVs and point mutations have different relative contributions to tumorigenesis. Driver SVs are more prevalent in breast adenocarcinomas (6.4 ± 3.7 SVs (mean ± s.d.) compared with 2.2 ± 1.3 point mutations; $P < 1 \times 10^{-16}$, Mann–Whitney *U*-test) and ovary adenocarcinomas (5.8 ± 2.6 SVs compared with 1.9 ± 1.0 point mutations; $P < 1 \times 10^{-16}$), whereas driver point mutations have a larger contribution in colorectal adenocarcinomas (2.4 ± 1.4 SVs compared with 7.4 ± 7.0 point mutations; $P = 4 \times 10^{-10}$) and mature B cell lymphomas (2.2 ± 1.3 SVs compared with 6 ± 3.8 point mutations; $P < 1 \times 10^{-16}$), as previously shown[38]. Across tumour types, there are differences in which classes of mutation affect a given genomic element (Fig. 2b).

We confirmed that many driver mutations that affect tumour-suppressor genes are two-hit inactivation events (Fig. 2c). For example, of the 954 tumours in the cohort with driver mutations in *TP53*, 736 (77%) had both alleles mutated, 96% of which (707 out of 736) combined a somatic point mutation that affected one allele with somatic deletion of the other allele. Overall, 17% of patients had rare germline protein-truncating variants (PTVs) in cancer-predisposition genes[39], DNA-damage response genes[40] and somatic driver genes. Biallelic inactivation due to somatic alteration on top of a germline PTV was observed in 4.5% of patients overall, with 81% of

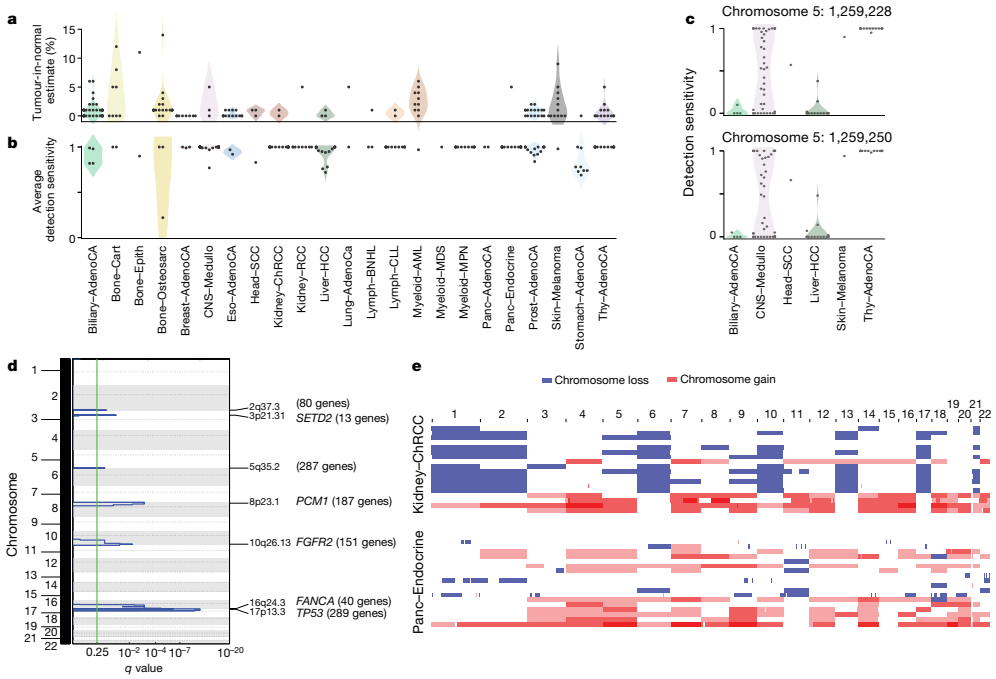


Fig. 3 | Analysis of patients with no detected driver mutations. a, Individual estimates of the percentage of tumour-in-normal contamination across patients with no driver mutations in PCAWG ($n = 181$). No data were available for myelodysplastic syndromes and acute myeloid leukaemia. Points represent estimates for individual patients, and the coloured areas are estimated density distributions (violin plots). Abbreviations of the tumour types are defined in Extended Data Table 1. **b**, Average detection sensitivity by tumour type for tumours without known drivers ($n = 181$). Each dot represents a given sample and is the average sensitivity of detecting clonal substitutions across the genome, taking into account purity and ploidy. Coloured areas are estimated density distributions, shown for cohorts with at least five cases. **c**, Detection sensitivity for *TERT* promoter hotspots in tumour types in which *TERT* is frequently mutated. Coloured areas are estimated density distributions. **d**, Significant copy-number losses identified by two-sided hypothesis testing using GISTIC2.0, corrected for multiple-hypothesis testing. Numbers in parentheses indicate the number of genes in significant regions when analysing medulloblastomas without known drivers ($n = 42$). Significant regions with known cancer-associated genes are labelled with the representative cancer-associated gene. **e**, Aneuploidy in chromophobe renal cell carcinomas and pancreatic neuroendocrine tumours without known drivers. Patients are ordered on the *y* axis by tumour type and then by presence of whole-genome duplication (bottom) or not (top).

these affecting known cancer-predisposition genes (such as *BRCA1*, *BRCA2* and *ATM*).

PCAWG tumours with no apparent drivers

Although more than 90% of PCAWG cases had identified drivers, we found none in 181 tumours (Extended Data Fig. 4a). Reasons for missing drivers have not yet been systematically evaluated in a pan-cancer cohort, and could arise from either technical or biological causes.

Technical explanations could include poor-quality samples, inadequate sequencing or failures in the bioinformatic algorithms used. We assessed the quality of the samples and found that 4 of the 181 cases with no known drivers had more than 5% tumour DNA contamination in their matched normal sample (Fig. 3a). Using an algorithm designed to correct for this contamination[41], we identified previously missed mutations in genes relevant to the respective cancer types. Similarly, if the fraction of tumour cells in the cancer sample is low through stromal contamination, the detection of driver mutations can be impaired. Most tumours with no known drivers had an average power to detect mutations close to 100%; however, a few had power in the 70–90% range (Fig. 3b and Extended Data Fig. 4b). Even in adequately sequenced genomes, lack of read depth at specific driver loci can impair mutation detection. For example, only around 50% of PCAWG tumours had sufficient coverage to call a mutation (≥90% power) at the two *TERT* promoter hotspots, probably because the high GC content of this region causes biased coverage (Fig. 3c). In fact, 6 hepatocellular carcinomas and 2 biliary cholangiocarcinomas among the 181 cases with no known drivers actually did contain *TERT* mutations, which were discovered after deep targeted sequencing[42].

Finally, technical reasons for missing driver mutations include failures in the bioinformatic algorithms. This affected 35 myeloproliferative neoplasms in PCAWG, in which the *JAK2*^{V617F} driver mutation should have been called. Our somatic variant-calling algorithms rely on 'panels of normals', typically from blood samples, to remove recurrent sequencing artefacts. As 2–5% of healthy individuals carry occult haematopoietic clones[43], recurrent driver mutations in these clones can enter panels of normals.

With regard to biological causes, tumours may be driven by mutations in cancer-associated genes that are not yet described for that tumour type. Using driver discovery algorithms on tumours with no known drivers, no individual genes reached significance for point mutations. However, we identified a recurrent CNA that spanned *SETD2* in

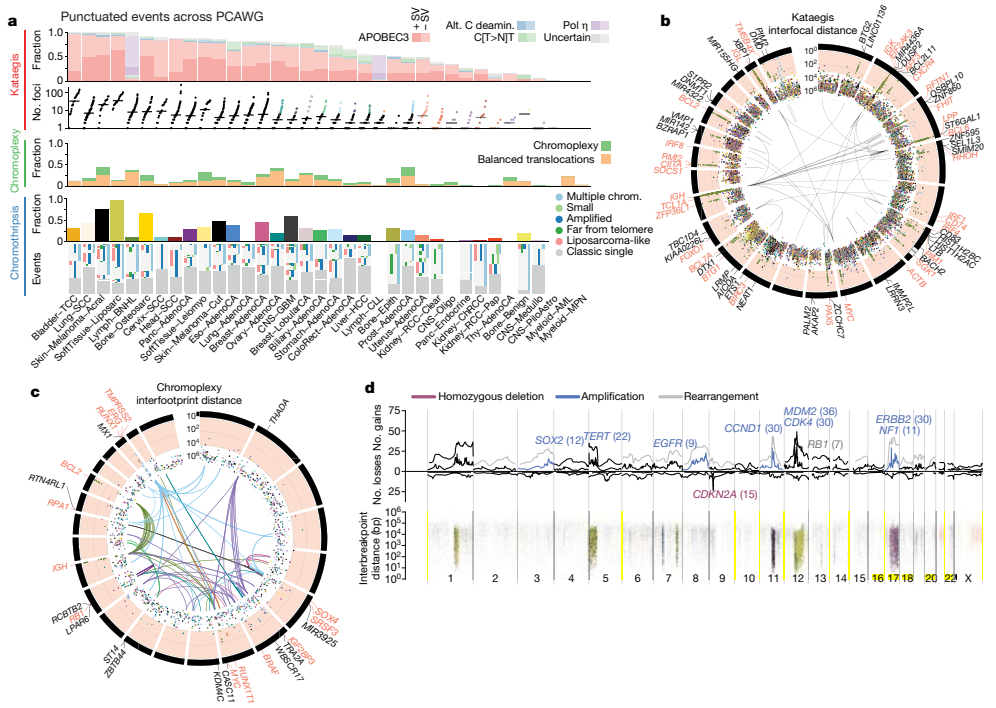


Fig. 4 | Patterns of clustered mutational processes in PCAWG. a, Kataegis. Top, prevalence of different types of kataegis and their association with SVs (≤1 kb from the focus). Bottom, the distribution of the number of foci of kataegis per sample. Chromoplexy. Prevalence of chromoplexy across cancer types, subdivided into balanced translocations and more complex events. Chromothripsis. Top, frequency of chromothripsis across cancer types. Bottom, for each cancer type a column is shown, in which each row is a chromothripsis region represented by five coloured rectangles relating to its categorization. **b**, Circos rainfall plot showing the distances between consecutive kataegis events across PCAWG compared with their genomic position. Lymphoid tumours (khaki, B cell non-Hodgkin's lymphoma; orange, chronic lymphocytic leukaemia) have hypermutation hot spots (≥3 foci with distance ≤1 kb; pale red zone), many of which are near known cancer-associated genes (red annotations) and have associated SVs (≤10 kb from the focus; shown as arcs in the centre). **c**, Circos rainfall plot as in **b** that shows the distance versus the position of consecutive chromoplexy and reciprocal translocation footprints across PCAWG. Lymphoid, prostate and thyroid cancers exhibit recurrent events (≥2 footprints with distance ≤10 kb; pale red zone) that are likely to be driver SVs and are annotated with nearby genes and associated SVs, which are shown as bold and thin arcs for chromoplexy and reciprocal translocations, respectively (colours as in **a**). **d**, Effect of chromothripsis along the genome and involvement of PCAWG driver genes. Top, number of chromothripsis-induced gains or losses (grey) and amplifications (blue) or deletions (red). Within the identified chromothripsis regions, selected recurrently rearranged (light grey), amplified (blue) and homozygously deleted (magenta) driver genes are indicated. Bottom, interbreakpoint distance between all subsequent breakpoints within chromothripsis regions across cancer types, coloured by cancer type. Regions with an average interbreakpoint distance <10 kb are highlighted. C[T>N]T, kataegis with a pattern of thymine mutations in a Cp **T**pT context.

medulloblastomas that lacked known drivers (Fig. 3d), indicating that restricting hypothesis testing to missing-driver cases can improve power if undiscovered genes are enriched in such tumours. Inactivation of *SETD2* in medulloblastoma significantly decreased gene expression ($P = 0.002$) (Extended Data Fig. 4c). Notably, *SETD2* mutations occurred exclusively in medulloblastoma group-4 tumours ($P < 1 \times 10^{-4}$). Group-4 medulloblastomas are known for frequent mutations in other chromatin-modifying genes[44], and our results suggest that *SETD2* loss of function is an additional driver that affects chromatin regulators in this subgroup.

Two tumour types had a surprisingly high fraction of patients without identified driver mutations: chromophobe renal cell carcinoma (44%; 19 out of 43) and pancreatic neuroendocrine cancers (22%; 18 out of 81) (Extended Data Fig. 4a). A notable feature of the missing-driver cases in both tumour types was a remarkably consistent profile of chromosomal aneuploidy—patterns that have previously been reported[45,46] (Fig. 3e). The absence of other identified driver mutations in these patients raises the possibility that certain combinations of whole-chromosome gains and losses may be sufficient to initiate a cancer in the absence of more-targeted driver events such as point mutations or fusion genes of focal CNAs.

Even after accounting for technical issues and novel drivers, 5.3% of PCAWG tumours still had no identifiable driver events. In a research setting, in which we are interested in drawing conclusions about populations of patients, the consequences of technical issues that affect occasional samples will be mitigated by sample size. In a clinical setting, in which we are interested in the driver mutations in a specific patient, these issues become substantially more important. Careful and critical appraisal of the whole pipeline—including sample acquisition, genome sequencing, mapping, variant calling and driver annotation, as done

here—should be required for laboratories that offer clinical sequencing of cancer genomes.

Patterns of clustered mutations and SVs

Some somatic mutational processes generate multiple mutations in a single catastrophic event, typically clustered in genomic space, leading to substantial reconfiguration of the genome. Three such processes have previously been described: (1) chromoplexy, in which repair of co-occurring double-stranded DNA breaks—typically on different chromosomes—results in shuffled chains of rearrangements[47,48] (Extended Data Fig. 5a); (2) kataegis, a focal hypermutation process that leads to locally clustered nucleotide substitutions, biased towards a single DNA strand[49–51] (Extended Data Fig. 5b); and (3) chromothripsis, in which tens to hundreds of DNA breaks occur simultaneously, clustered on one or a few chromosomes, with near-random stitching together of the resulting fragments[52–55] (Extended Data Fig. 5c). We characterized the PCAWG genomes for these three processes (Fig. 4).

Chromoplexy events and reciprocal translocations were identified in 467 (17.8%) samples (Fig. 4a, c). Chromoplexy was prominent in prostate adenocarcinoma and lymphoid malignancies, as previously described[47,48], and—unexpectedly—thyroid adenocarcinoma. Different genomic loci were recurrently rearranged by chromoplexy across the three tumour types, mediated by positive selection for particular fusion genes or enhancer-hijacking events. Of 13 fusion genes or enhancer hijacking events in 48 thyroid adenocarcinomas, at least 4 (31%) were caused by chromoplexy, with a further 4 (31%) part of complexes that contained chromoplexy footprints (Extended Data Fig. 5a). These events generated fusion genes that involved *RET* (two cases) and *NTRK3* (one case)[56], and the juxtaposition of the oncogene *IGF2BP3* with regulatory elements from highly expressed genes (five cases).

Kataegis events were found in 60.5% of all cancers, with particularly high abundance in lung squamous cell carcinoma, bladder cancer, acral melanoma and sarcomas (Fig. 4a, b). Typically, kataegis comprises C > N mutations in a TpC context, which are probably caused by APOBEC activity[49–51], although a T > N conversion in a **T**pT or Cp**T** process (the affected T is highlighted in bold) attributed to error-prone polymerases has recently been described[57]. The APOBEC signature accounted for 81.7% of kataegis events and correlated positively with *APOBEC3B* expression levels, somatic SV burden and age at diagnosis (Supplementary Fig. 5). Furthermore, 5.7% of kataegis events involved the T > N error-prone polymerase signature and 2.3% of events, most notably in sarcomas, showed cytidine deamination in an alternative Gp**C** or Cp**C** context.

Kataegis events were frequently associated with somatic SV breakpoints (Fig. 4a and Supplementary Fig. 6a), as previously described[50,51]. Deletions and complex rearrangements were most-strongly associated with kataegis, whereas tandem duplications and other simple SV classes were only infrequently associated (Supplementary Fig. 6b). Kataegis inducing predominantly T > N mutations in Cp**T**pT context was enriched near deletions, specifically those in the 10–25-kilobase (kb) range (Supplementary Fig. 6c).

Samples with extreme kataegis burden (more than 30 foci) comprise four types of focal hypermutation (Extended Data Fig. 6): (1) off-target somatic hypermutation and foci of T > N at Cp**T**pT, found in B cell non-Hodgkin lymphoma and oesophageal adenocarcinomas, respectively; (2) APOBEC kataegis associated with complex rearrangements, notably found in sarcoma and melanoma; (3) rearrangement-independent APOBEC kataegis on the lagging strand and in early-replicating regions, mainly found in bladder and head and neck cancer; and (4) a mix of the last two types. Kataegis only occasionally led to driver mutations (Supplementary Table 5).

We identified chromothripsis in 587 samples (22.3%), most frequently among sarcoma, glioblastoma, lung squamous cell carcinoma, melanoma and breast adenocarcinoma[18]. Chromothripsis increased with whole-genome duplications in most cancer types (Extended Data Fig. 7a), as previously shown in medulloblastoma[58]. The most recurrently associated driver was *TP53*[52] (pan-cancer odds ratio = 3.22; pan-cancer $P = 8.3 \times 10^{-35}$; $q < 0.05$ in breast lobular (odds ratio = 13), colorectal (odds ratio = 25), prostate (odds ratio = 2.6) and hepatocellular (odds ratio = 3.9) cancers; Fisher–Boschloo tests). In two cancer types (osteosarcoma and B cell lymphoma), women had a higher incidence of chromothripsis than men (Extended Data Fig. 7b). In prostate cancer, we observed a higher incidence of chromothripsis in patients with late-onset than early-onset disease[59] (Extended Data Fig. 7c).

Chromothripsis regions coincided with 3.6% of all identified drivers in PCAWG and around 7% of copy-number drivers (Fig. 4d). These proportions are considerably enriched compared to expectation if selection were not acting on these events (Extended Data Fig. 7d). The majority of coinciding driver events were amplifications (58%), followed by homozygous deletions (34%) and SVs within genes or promoter regions (8%). We frequently observed a ≥2-fold increase or decrease in expression of amplified or deleted drivers, respectively, when these loci were part of a chromothripsis event, compared with samples without chromothripsis (Extended Data Fig. 7e).

Chromothripsis manifested in diverse patterns and frequencies across tumour types, which we categorized on the basis of five characteristics (Fig. 4a). In liposarcoma, for example, chromothripsis events often involved multiple chromosomes, with universal *MDM2* amplification[60] and co-amplification of *TERT* in 4 of 19 cases (Fig. 4d). By contrast, in glioblastoma the events tended to affect a smaller region on a single chromosome that was distant from the telomere, resulting in focal amplification of *EGFR* and *MDM2* and loss of *CDKN2A*. Acral melanomas frequently exhibited *CCND1* amplification, and lung squamous cell carcinomas *SOX2* amplifications. In both cases, these drivers were more-frequently altered by chromothripsis compared with other drivers in the same cancer type and to other cancer types for the same driver (Fig. 4d and Extended Data Fig. 7f). Finally, in chromophobe renal cell carcinoma, chromothripsis nearly always affected chromosome 5 (Supplementary Fig. 7): these samples had breakpoints immediately adjacent to *TERT*, increasing *TERT* expression by 80-fold on average compared with samples without rearrangements ($P = 0.0004$; Mann–Whitney *U*-test).

Timing clustered mutations in evolution

An unanswered question for clustered mutational processes is whether they occur early or late in cancer evolution. To address this, we used molecular clocks to define broad epochs in the life history of each tumour[49,61]. One transition point is between clonal and subclonal mutations: clonal mutations occurred before, and subclonal mutations after, the emergence of the most-recent common ancestor. In regions with copy-number gains, molecular time can be further divided according to whether mutations preceded the copy-number gain (and were themselves duplicated) or occurred after the gain (and therefore present on only one chromosomal copy)[7].

Chromothripsis tended to have greater relative odds of being clonal than subclonal, suggesting that it occurs early in cancer evolution, especially in liposarcomas, prostate adenocarcinoma and squamous cell lung cancer (Fig. 5a). As previously reported, chromothripsis was especially common in melanomas[62]. We identified 89 separate chromothripsis events that affected 66 melanomas (61%); 47 out of 89 events affected genes known to be recurrently altered in melanoma[63] (Supplementary Table 6). Involvement of a region on chromosome 11 that includes the cell-cycle regulator *CCND1* occurred in 21 cases (10 out of 86 cutaneous, and 11 out of 21 acral or mucosal melanomas), typically combining chromothripsis with amplification (19 out of 21 cases) (Extended Data Fig. 8). Co-involvement of other cancer-associated genes in the same chromothripsis event was also frequent, including

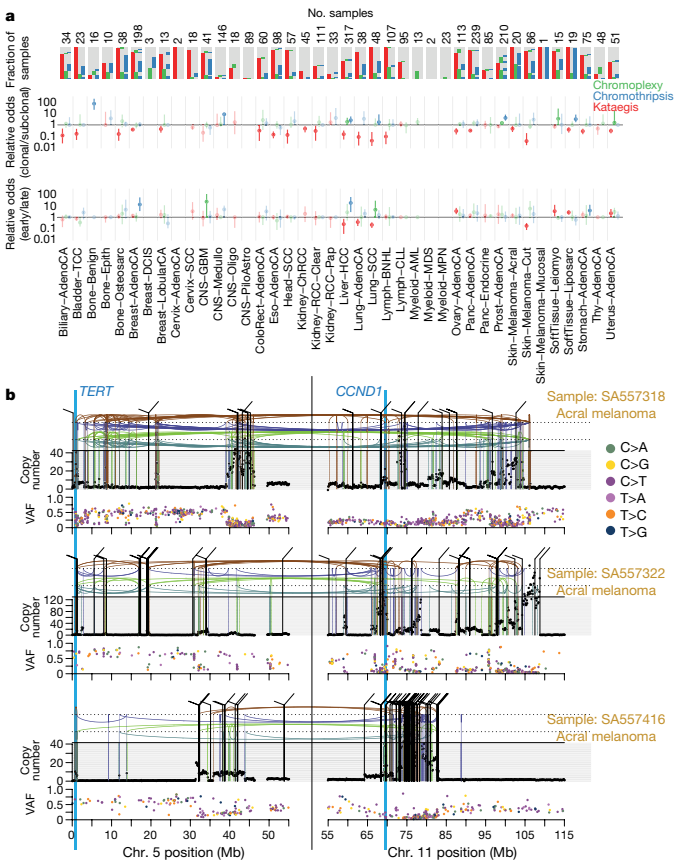


Fig. 5 | Timing of clustered events in PCAWG. a, Extent and timing of chromothripsis, kataegis and chromoplexy across PCAWG. Top, stacked bar charts illustrate co-occurrence of chromothripsis, kataegis and chromoplexy in the samples. Middle, relative odds of clustered events being clonal or subclonal are shown with bootstrapped 95% confidence intervals. Point estimates are highlighted when they do not overlap odds of 1:1. Bottom, relative odds of the events being early or late clonal are shown as above. Sample sizes (number of patients) are shown across the top. **b**, Three representative patients with acral melanoma and chromothripsis-induced amplification that simultaneously affects *TERT* and *CCND1*. The black points (top) represent sequence coverage from individual genomic bins, with SVs shown as coloured arcs (translocation in black, deletion in purple, duplication in brown, tail-to-tail inversion in cyan and head-to-head inversion in green). Bottom, the variant allele fractions of somatic point mutations.

TERT (five cases), *CDKN2A* (three cases), *TP53* (two cases) and *MYC* (two cases) (Fig. 5b). In these co-amplifications, a chromothripsis event involving multiple chromosomes initiated the process, creating a derivative chromosome in which hundreds of fragments were stitched together in a near-random order (Fig. 5b). This derivative then rearranged further, leading to massive co-amplification of the multiple target oncogenes together with regions located nearby on the derivative chromosome.

In these cases of amplified chromothripsis, we can use the inferred number of copies bearing each SNV to time the amplification process. SNVs present on the chromosome before amplification will themselves be amplified and are therefore reported in a high fraction of sequence reads (Fig. 5b and Extended Data Fig. 8). By contrast, late SNVs that occur after the amplification has concluded will be present on only one chromosome copy out of many, and thus have a low variant allele fraction. Regions of *CCND1* amplification had few—sometimes zero—mutations at high variant allele fraction in acral melanomas, in contrast to later *CCND1* amplifications in cutaneous melanomas, in which hundreds to thousands of mutations typically predated amplification (Fig. 5b and Extended Data Fig. 9a, b). Thus, both chromothripsis and the subsequent amplification generally occurred very early during the evolution of acral melanoma. By comparison, in lung squamous cell carcinomas, similar patterns of chromothripsis followed by *SOX2* amplification are characterized by many amplified SNVs, suggesting a later event in the evolution of these cancers (Extended Data Fig. 9c).

Notably, in cancer types in which the mutational load was sufficiently high, we could detect a larger-than-expected number of SNVs on an intermediate number of DNA copies, suggesting that they appeared during the amplification process (Supplementary Fig. 8).

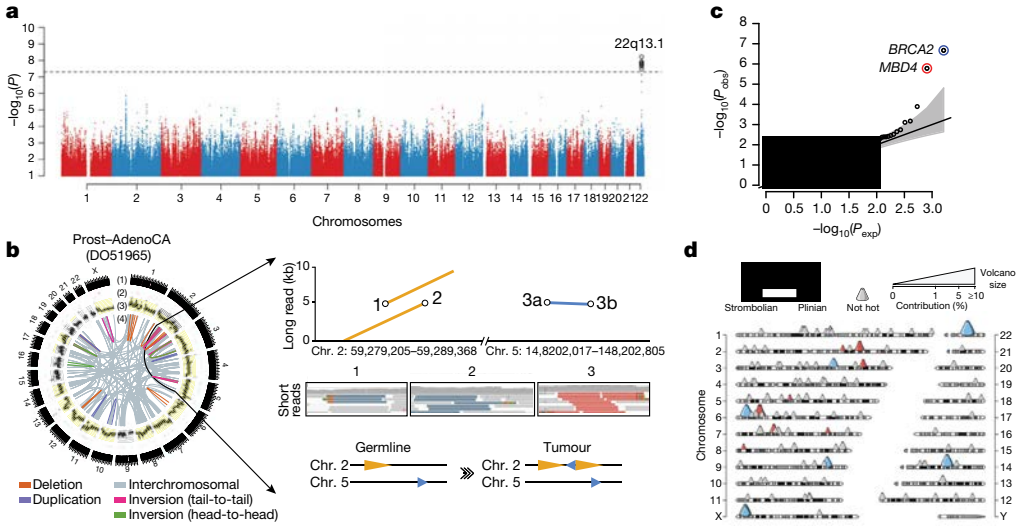


Fig. 6 | Germline determinants of the somatic mutation landscape. **a**, Association between common (MAF > 5%) germline variants and somatic APOBEC3B-like mutagenesis in individuals of European ancestry ($n = 1{,}201$). Two-sided hypothesis testing was performed with PLINK v.1.9. To mitigate multiple-hypothesis testing, the significance threshold was set to genome-wide significance ($P < 5 \times 10^{-8}$). **b**, Templated insertion SVs in a *BRCA1*-associated prostate cancer. Left, chromosome bands (1); SVs ≤ 10 megabases (Mb) (2); 1-kb read depth corrected to copy number 0–6 (3); inter- and intrachromosomal SVs > 10 Mb (4). Right, a complex somatic SV composed of a 2.2-kb tandem duplication on chromosome 2 together with a 232-base-pair (bp) inverted templated insertion SV that is derived from chromosome 5 and inserted inbetween the tandem duplication (bottom). Consensus sequence alignment of locally assembled Oxford Nanopore Technologies long sequencing reads to chromosomes 2 and 5 of the human reference genome (top). Breakpoints are circled and marked as 1 (beginning of tandem duplication), 2 (end of tandem duplication) or 3 (inverted templated insertion). For each breakpoint, the middle panel shows Illumina short reads at SV breakpoints. **c**, Association between rare germline PTVs (MAF < 0.5%) and somatic CpG mutagenesis (approximately with signature 1) in individuals of European ancestry ($n = 1{,}201$). Genes highlighted in blue or red were associated with lower or higher somatic mutation rates. Two-sided hypothesis testing was performed using linear-regression models with sex, age at diagnosis and cancer project as variables. To mitigate multiple-hypothesis testing, the significance threshold was set to exome-wide significance ($P < 2.5 \times 10^{-6}$). The black line represents the identity line that would be followed if the observed *P* values followed the null expectation; the shaded area shows the 95% confidence intervals. **d**, Catalogue of polymorphic germline L1 source elements that are active in cancer. The chromosomal map shows germline source L1 elements as volcano symbols. Each volcano is colour-coded according to the type of source L1 activity. The contribution of each source locus (expressed as a percentage) to the total number of transductions identified in PCAWG tumours is represented as a gradient of volcano size, with top contributing elements exhibiting larger sizes.

Germline effects on somatic mutations

We integrated the set of 88 million germline genetic variant calls with somatic mutations in PCAWG, to study germline determinants of somatic mutation rates and patterns. First, we performed a genome-wide association study of somatic mutational processes with common germline variants (minor allele frequency (MAF) > 5%) in individuals with inferred European ancestry. An independent genome-wide association study was performed in East Asian individuals from Asian cancer genome projects. We focused on two prevalent endogenous mutational processes: spontaneous deamination of 5-methylcytosine at CpG dinucleotides[5] (signature 1) and activity of the APOBEC3 family of cytidine deaminases[64] (signatures 2 and 13). No locus reached genome-wide significance ($P < 5 \times 10^{-8}$) for signature 1 (Extended Data Fig. 10a, b). However, a locus at 22q13.1 predicted an APOBEC3B-like mutagenesis at the pan-cancer level[65] (Fig. 6a). The strongest signal at 22q13.1 was driven by rs12628403, and the minor (non-reference) allele was protective against APOBEC3B-like mutagenesis ($\beta = -0.43$, $P = 5.6 \times 10^{-9}$, MAF = 8.2%, $n = 1{,}201$ donors) (Extended Data Fig. 10c). This variant tags a common, approximately 30-kb germline SV that deletes the *APOBEC3B* coding sequence and fuses the *APOBEC3B* 3′ untranslated region with the coding sequence of *APOBEC3A*. The deletion is known to increase breast cancer risk and APOBEC mutagenesis in breast cancer genomes[66,67]. Here, we found that rs12628403 reduces APOBEC3B-like mutagenesis specifically in cancer types with low levels of APOBEC mutagenesis ($\beta_{low} = -0.50$, $P_{low} = 1 \times 10^{-8}$; $\beta_{high} = 0.17$, $P_{high} = 0.2$), and increases APOBEC3A-like mutagenesis in cancer types with high levels of APOBEC mutagenesis ($\beta_{high} = 0.44$, $P_{high} = 8 \times 10^{-4}$; $\beta_{low} = -0.21$, $P_{low} = 0.02$). Moreover, we identified a second, novel locus at 22q13.1 that was associated with APOBEC3B-like mutagenesis across cancer types (rs2142833, $\beta = 0.23$, $P = 1.3 \times 10^{-8}$). We independently validated the association between both loci and APOBEC3B-like mutagenesis using East Asian individuals from Asian cancer genome projects ($\beta_{rs12628403} = 0.57$, $P_{rs12628403} = 4.2 \times 10^{-12}$; $\beta_{rs2142833} = 0.58$, $P_{rs2142833} = 8 \times 10^{-15}$) (Extended Data Fig. 10d). Notably, in a conditional analysis that accounted for rs12628403, we found that rs2142833 and rs12628403 are inherited independently in Europeans ($r^2 < 0.1$), and rs2142833 remained significantly associated with APOBEC3B-like mutagenesis in Europeans ($\beta_{EUR} = 0.17$, $P_{EUR} = 3 \times 10^{-5}$) and East Asians ($\beta_{ASN} = 0.25$, $P_{ASN} = 2 \times 10^{-3}$) (Extended Data Fig. 10e, f). Analysis of donor-matched expression data further suggests that rs2142833 is a *cis*-expression quantitative trait locus (eQTL) for *APOBEC3B* at the pan-cancer level ($\beta = 0.19$, $P = 2 \times 10^{-6}$) (Extended Data Fig. 10g, h), consistent with *cis*-eQTL studies in normal cells[68,69].

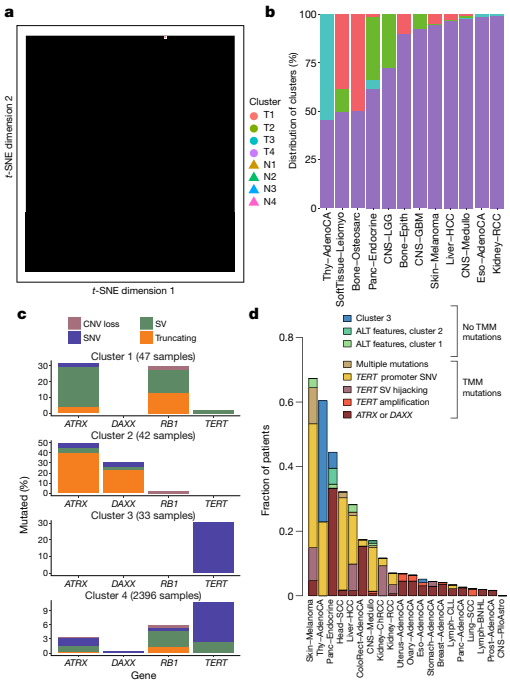


Fig. 7 | Telomere sequence patterns across PCAWG. a, Scatter plot of the clusters of telomere patterns identified across PCAWG using *t*-distributed stochastic neighbour embedding (*t*-SNE), based on $n = 2{,}518$ tumour samples and their matched normal samples. Axes have arbitrary dimensions such that samples with similar telomere profiles are clustered together and samples with dissimilar telomere profiles are far apart with high probability. **b**, Distribution of the four tumour-specific clusters of telomere patterns in selected tumour types from PCAWG. **c**, Distribution of relevant driver mutations associated with alternative lengthening of telomere and normal telomere maintenance across the four clusters. **d**, Distribution of telomere maintenance abnormalities across tumour types with more than 40 patients in PCAWG. Samples were classified as tumour clusters 1–3 if they fell into a relevant cluster without mutations in *TERT*, *ATRX* or *DAXX* and had no ALT phenotype. TMM, telomere maintenance mechanisms.

Second, we performed a rare-variant association study (MAF <0.5%) to investigate the relationship between germline PTVs and somatic DNA rearrangements in individuals with European ancestry (Extended Data Fig. 11a–c). Germline *BRCA2* and *BRCA1* PTVs were associated with an increased burden of small (less than 10 kb) somatic SV deletions ($P = 1 \times 10^{-8}$) and tandem duplications ($P = 6 \times 10^{-13}$), respectively, corroborating recent studies in breast and ovarian cancer[30,70]. In PCAWG data, this pattern also extends to other tumour types, including adenocarcinomas of the prostate and pancreas[6], typically in the setting of biallelic inactivation. In addition, tumours with high levels of small SV tandem duplications frequently exhibited a novel and distinct class of SVs termed 'cycles of templated insertions'[6]. These complex SV events consist of DNA templates that are copied from across the genome, joined into one contiguous sequence and inserted into a single derivative chromosome. We found a significant association between germline *BRCA1* PTVs and templated insertions at the pan-cancer level ($P = 4 \times 10^{-15}$) (Extended Data Fig. 11d, e). Whole-genome long-read sequencing data generated for a *BRCA1*-deficient PCAWG prostate tumour verified the small tandem-duplication and templated-insertion SV phenotypes (Fig. 6b). Almost all (20 out of 21) of *BRCA1*-associated tumours with a templated-insertion SV phenotype displayed combined germline and somatic hits in the gene. Together, these data suggest that biallelic inactivation of *BRCA1* is a driver of the templated-insertion SV phenotype.

Third, rare-variant association analysis revealed that patients with germline *MBD4* PTVs had increased rates of somatic C > T mutation rates at CpG dinucleotides ($P < 2.5 \times 10^{-6}$) (Fig. 6c and Extended Data Fig. 11f, g). Analysis of previously published whole-exome sequencing samples from the TCGA ($n = 8{,}134$) replicated the association between germline *MBD4* PTVs and increased somatic CpG mutagenesis at the pan-cancer level ($P = 7.1 \times 10^{-4}$) (Extended Data Fig. 11h). Moreover, gene-expression profiling revealed a significant but modest correlation between *MBD4* expression and somatic CpG mutation rates between and within PCAWG tumour types (Extended Data Fig. 11i–k). *MBD4* encodes a DNA-repair gene that removes thymidines from T:G mismatches within methylated CpG sites[71], a functionality that would be consistent with a CpG mutational signature in cancer.

Fourth, we assessed long interspersed nuclear elements (LINE-1; L1 hereafter) that mediate somatic retrotransposition events[72–74]. We identified 114 germline source L1 elements capable of active somatic retrotransposition, including 70 that represent insertions with respect to the human reference genome (Fig. 6d and Supplementary Table 7), and 53 that were tagged by single-nucleotide polymorphisms in strong linkage disequilibrium (Supplementary Table 7). Only 16 germline L1 elements accounted for 67% (2,440 out of 3,669) of all L1-mediated transductions[10] detected in the PCAWG dataset (Extended Data Fig. 12a). These 16 hot-L1 elements followed two broad patterns of somatic activity (8 of each), which we term Strombolian and Plinian in analogy to patterns of volcanic activity. Strombolian L1s are frequently active in cancer, but mediate only small-to-modest eruptions of somatic L1 activity in cancer samples (Extended Data Fig. 12b). By contrast, Plinian L1s are more rarely seen, but display aggressive somatic activity. Whereas Strombolian elements are typically relatively common (MAF > 2%) and sometimes even fixed in the human population, all Plinian elements were infrequent (MAF ≤ 2%) in PCAWG donors (Extended Data Fig. 12c; $P = 0.001$, Mann–Whitney *U*-test). This dichotomous pattern of activity and allele frequency may reflect differences in age and selective pressures, with Plinian elements potentially inserted into the human germline more recently. PCAWG donors bear on average between 50 and 60 L1 source elements and between 5 and 7 elements with hot activity (Extended Data Fig. 12d), but only 38% (1,075 out of 2,814) of PCAWG donors carried ≥1 Plinian element. Some L1 germline source loci caused somatic loss of tumour-suppressor genes (Extended Data Fig. 12e). Many are restricted to individual continental population ancestries (Extended Data Fig. 12f–j).

Replicative immortality

One of the hallmarks of cancer is the ability of cancer to evade cellular senescence[21]. Normal somatic cells typically have finite cell division potential; telomere attrition is one mechanism to limit numbers of mitoses[75]. Cancers enlist multiple strategies to achieve replicative immortality. Overexpression of the telomerase gene, *TERT*, which maintains telomere lengths, is especially prevalent. This can be achieved through point mutations in the promoter that lead to de novo transcription factor binding[34,37]; hitching *TERT* to highly active regulatory elements elsewhere in the genome[46,76]; insertions of viral enhancers upstream of the gene[77,78]; and increased dosage through chromosomal amplification, as we have seen in melanoma (Fig. 5b). In addition, there is an 'alternative lengthening of telomeres' (ALT) pathway, in which telomeres are lengthened through homologous recombination, mediated by loss-of-function mutations in the *ATRX* and *DAXX* genes[79].

As reported in a companion paper[13], 16% of tumours in the PCAWG dataset exhibited somatic mutations in at least one of *ATRX*, *DAXX* and *TERT*. *TERT* alterations were detected in 270 samples, whereas 128 tumours had alterations in *ATRX* or *DAXX*, of which 71 were protein-truncating. In the companion paper, which focused on describing patterns of ALT and *TERT*-mediated telomere maintenance[13], 12 features of telomeric sequence were measured in the PCAWG cohort. These included counts of nine variants of the core hexameric sequence, the number of ectopic telomere-like insertions within the genome, the number of genomic breakpoints and telomere length as a ratio between tumour and normal. Here we used the 12 features as an overview of telomere integrity across all tumours in the PCAWG dataset.

On the basis of these 12 features, tumour samples formed 4 distinct subclusters (Fig. 7a and Extended Data Fig. 13a), suggesting that telomere-maintenance mechanisms are more diverse than the well-established *TERT* and ALT dichotomy. Clusters C1 (47 tumours) and C2 (42 tumours) were enriched for traits of the ALT pathway—having longer telomeres, more genomic breakpoints, more ectopic telomere insertions and variant telomere sequence motifs (Supplementary Fig. 9). C1 and C2 were distinguished from one another by the latter having a considerable increase in the number of TTCGGG and TGAGGG variant motifs among the telomeric hexamers. Thyroid adenocarcinomas were markedly enriched among C3 samples (26 out of 33 C3 samples; $P < 10^{-16}$); the C1 cluster (ALT subtype 1) was common among sarcomas; and both pancreatic endocrine neoplasms and low-grade gliomas had a high proportion of samples in the C2 cluster (ALT subtype 2) (Fig. 7b). Notably, some of the thyroid adenocarcinomas and pancreatic neuroendocrine tumours that cluster together (cluster C3) had matched normal samples that also cluster together (normal cluster N3) (Extended Data Fig. 13a) and which share common properties. For example, the GTAGGG repeat was overrepresented among samples in this group (Supplementary Fig. 10).

Somatic driver mutations were also unevenly distributed across the four clusters (Fig. 7c). C1 tumours were enriched for *RB1* mutations or SVs ($P = 3 \times 10^{-5}$), as well as frequent SVs that affected *ATRX* ($P = 6 \times 10^{-14}$), but not *DAXX*. *RB1* and *ATRX* mutations were largely mutually exclusive (Extended Data Fig. 13b). By contrast, C2 tumours were enriched for somatic point mutations in *ATRX* and *DAXX* ($P = 6 \times 10^{-5}$), but not *RB1*. The enrichment of *RB1* mutations in C1 remained significant when only leiomyosarcomas and osteosarcomas were considered, confirming that this enrichment is not merely a consequence of the different distribution of tumour types across clusters. C3 samples had frequent *TERT* promoter mutations (30%; $P = 2 \times 10^{-6}$).

There was a marked predominance of *RB1* mutations in C1. Nearly a third of the samples in C1 contained an *RB1* alteration, which were evenly distributed across truncating SNVs, SVs and shallow deletions (Extended Data Fig. 13c). Previous research has shown that *RB1* mutations are associated with long telomeres in the absence of *TERT* mutations and *ATRX* inactivation[80], and studies using mouse models have shown that knockout of Rb-family proteins causes elongated telomeres[81]. The association with the C1 cluster here suggests that *RB1* mutations can represent another route to activating the ALT pathway, which has subtly different properties of telomeric sequence compared with the inactivation of *DAXX*—these fall almost exclusively in cluster C2.

Tumour types with the highest rates of abnormal telomere maintenance mechanisms often originate in tissues that have low endogenous replicative activity (Fig. 7d). In support of this, we found an inverse correlation between previously estimated rates of stem cell division across tissues[82] and the frequency of telomere maintenance abnormalities ($P = 0.01$, Poisson regression) (Extended Data Fig. 13d). This suggests that restriction of telomere maintenance is an important tumour-suppression mechanism, particularly in tissues with low steady-state cellular proliferation, in which a clone must overcome this constraint to achieve replicative immortality.

Conclusions and future perspectives

The resource reported in this paper and its companion papers has yielded insights into the nature and timing of the many mutational processes that shape large- and small-scale somatic variation in the cancer genome; the patterns of selection that act on these variations; the widespread effect of somatic variants on transcription; the complementary roles of the coding and non-coding genome for both germline and somatic mutations; the ubiquity of intratumoral heterogeneity; and the distinctive evolutionary trajectory of each cancer type. Many of these insights can be obtained only from an integrated analysis of all classes of somatic mutation on a whole-genome scale, and would not be accessible with, for example, targeted exome sequencing.

The promise of precision medicine is to match patients to targeted therapies using genomics. A major barrier to its evidence-based implementation is the daunting heterogeneity of cancer chronicled in these papers, from tumour type to tumour type, from patient to patient, from clone to clone and from cell to cell. Building meaningful clinical predictors from genomic data can be achieved, but will require knowledge banks comprising tens of thousands of patients with comprehensive clinical characterization[83]. As these sample sizes will be too large for any single funding agency, pharmaceutical company or health system, international collaboration and data sharing will be required. The next phase of ICGC, ICGC-ARGO (https:// www.icgc-argo.org/), will bring the cancer genomics community together with healthcare providers, pharmaceutical companies, data science and clinical trials groups to build comprehensive knowledge banks of clinical outcome and treatment data from patients with a wide variety of cancers, matched with detailed molecular profiling.

Extending the story begun by TCGA, ICGC and other cancer genomics projects, the PCAWG has brought us closer to a comprehensive narrative of the causal biological changes that drive cancer phenotypes. We must now translate this knowledge into sustainable, meaningful clinical treatments.

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/s41586-020-1969-6.

1. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
2. Pleasance, E. D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
3. Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
4. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,693 cancer whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1965-x (2020).
5. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* https://doi.org/10.1038/s41586-020-1943-3 (2020).
6. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* https://doi.org/10.1038/s41586-019-1913-9 (2020).
7. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* https://doi.org/10.1038/s41586-019-1907-7 (2020).
8. PCAWG Transcriptome Core Group et al. Genomic basis of RNA alterations in cancer. *Nature* https://doi.org/10.1038/s41586-020-1970-0 (2020).
9. Zhang, Y. et al. High-coverage whole-genome analysis of 1,220 cancers reveals hundreds of genes deregulated by rearrangement-mediated *cis*-regulatory alterations. *Nat. Commun.* https://doi.org/10.1038/s41467-019-13885-w (2020).
10. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0562-0 (2020).
11. Zapatka, M. et al. The landscape of viral associations in human cancers. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0558-9 (2020).
12. Jiao, W. et al. A deep learning system can accurately classify primary and metastatic cancers based on patterns of passenger mutations. *Nat. Commun.* https://doi.org/10.1038/s41467-019-13825-8 (2020).

13. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun*. https://doi.org/10.1038/s41467-019-13824-9 (2020).
14. Yuan, Y. et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet*. https://doi.org/10.1038/s41588-019-0557-x (2020).
15. Akdemir, K. C. et al. Chromatin folding domains disruptions by somatic genomic rearrangements in human cancers. *Nat. Genet*. https://doi.org/10.1038/s41588-019-0564-y (2020).
16. Reyna, M. A. et al. Pathway and network analysis of more than 2,500 whole cancer genomes. *Nat. Commun*. https://doi.org/10.1038/s41467-020-14351-8 (2020).
17. Bailey, M. H. et al. Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. *Nat. Commun*. (2020).
18. Cortes-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet*. https://doi.org/10.1038/s41588-019-0576-7 (2020).
19. Bray, F., Ren, J.-S., Masuyer, E. & Ferlay, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132**, 1133–1145 (2013).
20. Tarver, T. Cancer Facts & Figures 2012. American Cancer Society (ACS). *J. Consum. Health Internet* **16**, 366–367 (2012).
21. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
22. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
23. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
24. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
25. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
26. Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G. & Korbel, J. O. Data analysis: create a cloud commons. *Nature* **523**, 149–151 (2015).
27. Phillips, M. et al. Genomics: data sharing needs international code of conduct. *Nature* https://doi.org/10.1038/d41586-020-00082-9 (2020).
28. Krochmalski, J. *Developing with Docker* (Packt Publishing, 2016).
29. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
30. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
31. Meier, B. et al. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*. **24**, 1624–1636 (2014).
32. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
33. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. **10**, 25 (2018).
34. Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
35. Rheinbay, E. et al. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).
36. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet*. **46**, 1258–1263 (2014).
37. Horn, S. et al. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
38. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet*. **45**, 1127–1133 (2013).
39. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
40. Pearl, L. H., Schierz, A. C., Ward, S. E., Al-Lazikani, B. & Pearl, F. M. G. Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer* **15**, 166–180 (2015).
41. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
42. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet*. **48**, 500–509 (2016).
43. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
44. Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
45. Scarpa, A. et al. Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature* **543**, 65–71 (2017).
46. Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
47. Berger, M. F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
48. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
49. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
50. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
51. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
52. Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148**, 59–71 (2012).
53. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
54. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
55. Zhang, C.-Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
56. The Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
57. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
58. Mardin, B. R. et al. A cell-based model system links chromothripsis with hyperploidy. *Mol. Syst. Biol*. **11**, 828 (2015).
59. Weischenfeldt, J. et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
60. Garsed, D. W. et al. The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653–667 (2014).
61. Durinck, S. et al. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov*. **1**, 137–143 (2011).
62. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
63. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
64. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
65. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet*. **47**, 1067–1072 (2015).
66. Nik-Zainal, S. et al. Association of a germline copy number polymorphism of *APOBEC3A* and *APOBEC3B* with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet*. **46**, 487–491 (2014).
67. Middlebrooks, C. D. et al. Association of germline variants in the *APOBEC3* region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet*. **48**, 1330–1338 (2016).
68. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet*. **45**, 1238–1243 (2013).
69. Stranger, B. E. et al. Population genomics of human gene expression. *Nat. Genet*. **39**, 1217–1224 (2007).
70. Menghi, F. et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl Acad. Sci. USA* **113**, E2373–E2382 (2016).
71. Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301–304 (1999).
72. Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
73. Tubio, J. M. C. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343–1251343 (2014).
74. Helman, E. et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res*. **24**, 1053–1063 (2014).
75. Shay, J. W. & Wright, W. E. Hayflick, his limit, and cellular ageing. *Nat. Rev. Mol. Cell Biol*. **1**, 72–76 (2000).
76. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
77. Totoki, Y. et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet*. **46**, 1267–1273 (2014).
78. Paterlini-Bréchot, P. et al. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. *Oncogene* **22**, 3911–3916 (2003).
79. Heaphy, C. M. et al. Prevalence of the alternative lengthening of telomeres telomere maintenance mechanism in human cancer subtypes. *Am. J. Pathol*. **179**, 1608–1615 (2011).
80. Barthel, F. P. et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet*. **49**, 349–357 (2017).
81. García-Cao, M., Gonzalo, S., Dean, D. & Blasco, M. A. A role for the Rb family of proteins in controlling telomere length. *Nat. Genet*. **32**, 415–419 (2002).
82. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
83. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet*. **49**, 332–340 (2017).
84. O'Connor, B. D. et al. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Res*. **6**, 52 (2017).
85. Zhang, J. et al. The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol*. **37**, 367–369 (2019).
86. Miller, C. A., Qiao, Y., DiSera, T., D'Astous, B. & Marth, G. T. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nat. Methods* **11**, 1189–1189 (2014).
87. Goldman, M. et al. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. Preprint at https://www.biorxiv.org/content/10.1101/326470v6 (2019)
88. Papatheodorou, I. et al. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res*. **46**, D246–D251 (2018).

Article

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Peter J. Campbell[1,2,745*], Gad Getz[3,4,5,6,745*], Jan O. Korbel[7,8,745*], Joshua M. Stuart[9,745*], Jennifer L. Jennings[10,11,745], Lincoln D. Stein[12,13,745*], Marc D. Perry[14,15], Hardeep K. Nahal-Bose[15], B. F. Francis Ouellette[16,17], Constance H. Li[12,18], Esther Rheinbay[3,6,19], G. Petur Nielsen[19], Dennis C. Sgroi[19], Chin-Lee Wu[19], William C. Faquin[19], Vikram Deshpande[19], Paul C. Boutros[12,18,20,21], Alexander J. Lazar[22], Katherine A. Hoadley[23,24], David N. Louis[19], L. Jonathan Dursi[12,25], Christina K. Yung[15], Matthew H. Bailey[26,27], Gordon Saksena[3], Keiran M. Raine[1], Ivo Buchhalter[28,29,30], Kortine Kleinheinz[28,30], Matthias Schlesner[28,31], Junjun Zhang[15], Wenyi Wang[32], David A. Wheeler[33,34], Li Ding[26,27,35], Jared T. Simpson[12,36], Brian D. O'Connor[15,37], Sergei Yakneen[8], Kyle Ellrott[38], Naoki Miyoshi[39], Adam P. Butler[1], Romina Royo[40], Solomon I. Shorser[12], Miguel Vazquez[40,41], Tobias Rausch[8], Grace Tiao[3], Sebastian M. Waszak[8], Bernardo Rodriguez-Martin[42,43,44], Suyash Shringarpure[45], Dai-Ying Wu[46], German M. Demidov[47,48,49], Olivier Delaneau[50,51,52], Shuto Hayashi[39], Seiya Imoto[39], Nina Habermann[8], Ayellet V. Segre[3,53], Erik Garrison[1], Andy Cafferkey[7], Eva G. Alvarez[42,43,44], José María Heredia-Genestar[54], Francesc Muyas[47,48,49], Oliver Drechsel[47,49], Alicia L. Bruzos[42,43,44], Javier Temes[42,43], Jorge Zamora[1,42,43,44], Adrian Baez-Ortega[55], Hyung-Lae Kim[56], R. Jay Mashl[27,57], Kai Ye[58,59], Anthony DiBiase[60], Kuan-lin Huang[27,61], Ivica Letunic[62], Michael D. McLellan[26,27,35], Steven J. Newhouse[7], Tal Shmaya[46], Sushant Kumar[63,64], David C. Wedge[1,65,66], Mark H. Wright[45], Venkata D. Yellapantula[67,68], Mark Gerstein[63,64,69], Ekta Khurana[70,71,72,73], Tomas Marques-Bonet[74,75,76,77], Arcadi Navarro[74,75,76], Carlos D. Bustamante[45,78], Reiner Siebert[79,80], Hidewaki Nakagawa[81], Douglas F. Easton[82,83], Stephan Ossowski[47,48,49], Jose M. C. Tubio[42,43,44], Francisco M. De La Vega[45,46,78], Xavier Estivill[47,84], Denis Yuen[12], George L. Mihaiescu[15], Larsson Omberg[85], Vincent Ferretti[15,86], Radhakrishnan Sabarinathan[87,88,89], Oriol Pich[87,89], Abel Gonzalez-Perez[87,89], Amaro Taylor-Weiner[90], Matthew W. Fittall[91], Jonas Demeulemeester[91,92], Maxime Tarabichi[1,91], Nicola D. Roberts[1], Peter Van Loo[91,92], Isidro Cortés-Ciriano[93,94,95], Lara Urban[7,8], Peter Park[34,95], Bin Zhu[96], Esa Pitkänen[8], Yilong Li[1], Natalie Saini[97], Leszek J. Klimczak[98], Joachim Weischenfeldt[8,99,100], Nikos Sidiropoulos[100], Ludmil B. Alexandrov[1,101], Raquel Rabionet[47,49,102], Georgia Escaramis[47,103,104], Mattia Bosio[40,47,49], Aliaksei Z. Holik[47], Hana Susak[47,49], Aparna Prasad[49], Serap Erkek[8], Claudia Calabrese[7,8], Benjamin Raeder[8], Eoghan Harrington[105], Simon Mayes[106], Daniel Turner[106], Sissel Juul[105], Steven A. Roberts[107], Lei Song[96], Roelof Koster[108], Lisa Mirabello[86], Xing Hua[96], Tomas J. Tanskanen[109], Marta Tojo[44], Jieming Chen[64,110], Lauri A. Aaltonen[111], Gunnar Rätsch[112,113,114,115,116,117], Roland F. Schwarz[7,118,119,120], Atul J. Butte[121], Alvis Brazma[7], Stephen J. Chanock[86], Nilanjan Chatterjee[122,123], Oliver Stegle[7,8,124], Olivier Harismendy[125], G. Steven Bova[126], Dmitry A. Gordenin[97], David Haan[9], Lina Sieverling[127,128], Lars Feuerbach[127], Don Chalmers[129], Yann Joly[130], Bartha Knoppers[130], Fruzsina Molnár-Gábor[131], Mark Phillips[130], Adrian Thorogood[130], David Townend[130], Mary Goldman[132], Nuno A. Fonseca[7,133], Qian Xiang[15], Brian Craft[132], Elena Piñeiro-Yáñez[134], Alfonso Muñoz[7], Robert Petryszak[7], Anja Füllgrabe[7], Fatima Al-Shahrour[134], Maria Keays[7], David Haussler[132,135], John Weinstein[136,137], Wolfgang Huber[8], Alfonso Valencia[40,76], Irene Papatheodorou[7], Jingchun Zhu[132], Yu Fan[32], David Torrents[40,76], Matthias Bieg[138,139], Ken Chen[140], Zechen Chong[141], Kristian Cibulskis[3], Roland Eils[28,30,142,143], Robert S. Fulton[26,27,35], Josep L. Gelpi[40,144], Santiago Gonzalez[7,8], Ivo G. Gut[49,74], Faraz Hach[145,146], Michael Heinold[28,30], Taobo Hu[147], Vincent Huang[12], Barbara Hutter[139,148,149], Natalie Jäger[28], Jongsun Jung[150], Yogesh Kumar[147], Christopher Lalansingh[12], Ignaty Leshchiner[3], Dimitri Livitz[3], Eric Z. Ma[147], Yosef E. Maruvka[3,19,151], Ana Milovanovic[40], Morten Muhlig Nielsen[152], Nagarajan Paramasivam[28,139], Jakob Skou Pedersen[152,153], Montserrat Puiggròs[40], S. Cenk Sahinalp[146,154,155], Iman Sarrafi[146,155], Chip Stewart[3], Miranda D. Stobbe[49,74], Jeremiah A. Wala[3,6,156], Jiayin Wang[27,58,157], Michael Wendl[27,158,159], Johannes Werner[28,160], Zhenggang Wu[147], Hong Xue[147], Takafumi N. Yamaguchi[12], Venkata Yellapantula[67,68], Brandi N. Davis-Dusenbery[161], Robert L. Grossman[162], Youngwook Kim[163,164], Michael C. Heinold[28,30], Jonathan Hinton[1], David R. Jones[1], Andrew Menzies[1], Lucy Stebbings[1], Julian M. Hess[3,151], Mara Rosenberg[3,19], Andrew J. Dunford[3], Manaswi Gupta[3], Marcin Imielinski[165,166], Matthew Meyerson[3,6,156], Rameen Beroukhim[3,6,167], Jüri Reimand[12,18], Priyanka Dhingra[71,73], Francesco Favero[168], Stefan Dentro[1,65,91], Jeff Wintersinger[169,170,171], Vasilisa Rudneva[6], Ji Wan Park[172], Eun Pyo Hong[172], Seong Gu Heo[172], André Kahles[112,113,114,115,116], Kjong-Van Lehmann[112,114,115,173,174], Cameron M. Soulette[37], Yuichi Shiraishi[39], Fenglin Liu[175,176], Yao He[175], Deniz Demircioğlu[177,178], Natalie R. Davidson[112,114,115,117,173], Liliana Greger[7], Siliang Li[179,180], Dongbing Liu[179,180], Stefan G. Stark[115,173,181,182], Fan Zhang[175], Samirkumar B. Amin[183,184,185], Peter Bailey[186], Aurélien Chateigner[15], Milana Frenkel-Morgenstern[187], Yong Hou[179,180], Matthew R. Huska[118], Helena Kilpinen[188], Fabien C. Lamaze[12], Chang Li[179,180], Xiaobo Li[179,180], Xinyue Li[179], Xingmin Liu[179,180], Maximillian G. Marin[37], Julia Markowski[118], Tannistha Nandi[189], Akinyemi I. Ojesina[190,191,192], Qiang Pan-Hammarström[179,193], Peter J. Park[94,95], Chandra Sekhar Pedamallu[3,6,167], Hong Su[179,180], Patrick Tan[189,194,195,196], Bin Tean Teh[194,195,196,197,198], Jian Wang[179], Heng Xiong[179,180], Chen Ye[179,180], Christina Yung[15], Xiuqing Zhang[179], Liangtao Zheng[175], Shida Zhu[179,180], Philip Awadalla[12,13], Chad J. Creighton[199], Kui Wu[179,180], Huanming Yang[179], Jonathan Göke[177,200], Zemin Zhang[175,201], Angela N. Brooks[3,37,156], Matthew W Fittall[91], Iñigo Martincorena[1], Carlota Rubio-Perez[87,89,202], Malene Juul[152], Steven Schumacher[3,203], Ofer Shapira[3,156], David Tamborero[87,89], Loris Mularoni[87,89], Henrik Hornshøj[152], Jordi Deu-Pons[89,204], Ferran Muiños[87,89], Johanna Bertl[152,205], Qianyun Guo[153], Abel Gonzalez-Perez[87,89,206], Qian Xiang[207], Wojciech Bazant[7], Elisabet Barrera[7], Sultan T. Al-Sedairy[208], Axel Aretz[209], Cindy Bell[210], Miguel Betancourt[211], Christiane Buchholz[212], Fabien Calvo[213], Christine Chomienne[214], Michael Dunn[215], Stuart Edmonds[216], Eric Green[217], Shailja Gupta[218], Carolyn M. Hutter[217], Karine Jegalian[219], Nic Jones[220], Youyong Lu[221,222,223], Hitoshi Nakagama[224], Gerd Nettekoven[225], Laura Planko[225], David Scott[220], Tatsuhiro Shibata[226,227], Kiyo Shimizu[228], Michael R. Stratton[1], Takashi Yugawa[228], Giampaolo Tortora[229,230], K. VijayRaghavan[218], Jean C. Zenklusen[231], David Townend[232], Bartha M. Knoppers[130], Brice Aminou[15], Javier Bartolome[40], Keith A. Boroevich[81,233], Rich Boyce[7], Alex Buchanan[38], Niall J. Byrne[15], Zhaohong Chen[234], Sunghoon Cho[235], Wan Choi[236], Peter Clapham[1], Michelle T. Dow[234], Lewis Jonathan Dursi[12,25], Juergen Eils[142,143], Claudiu Farcas[234], Nodirjon Fayzullaev[15], Paul Flicek[7], Allison P. Heath[237], Oliver Hofmann[238], Jongwhi H. Hong[239], Thomas J. Hudson[240,241], Daniel Hübschmann[30,120,142,242,243], Sinisa Ivkovic[244], Seung-Hyup Jeon[236], Wei Jiao[12], Rolf Kabbe[28], Andre Kahles[112,113,114,115,174], Jules N. A. Kerssemakers[28], Hyunghwan Kim[236], Jihoon Kim[245], Michael Koscher[246], Antonios Koures[234], Milena Kovacevic[244], Chris Lawerenz[143], Jia Liu[247], Sanja Mijalkovic[244], Ana Mijalkovic Mijalkovic-Lazic[244], Satoru Miyano[39], Mia Nastic[244], Jonathan Nicholson[1], David Ocana[7], Kazuhiro Ohi[39], Lucila Ohno-Machado[234], Todd D. Pihl[248], Manuel Prinz[28], Petar Radovic[244], Charles Short[7], Heidi J. Sofia[217], Jonathan Spring[162], Adam J. Struck[38], Nebojsa Tijanic[244], David Vicente[40], Zhining Wang[231], Ashley Williams[234], Youngchoon Woo[236], Adam J. Wright[12], Liming Yang[231], Mark P. Hamilton[249], Todd A. Johnson[235], Abdullah Kahraman[250,251,252], Manolis Kellis[3,253], Paz Polak[3,4,6], Richard Sallari[3], Nasa Sinnott-Armstrong[3,45], Christian von Mering[252,254], Sergi Beltran[49,74], Daniela S. Gerhard[255], Marta Gut[49,74], Jean-Rémi Trotta[74], Justin P. Whalley[74], Beifang Niu[256], Shadrielle M. G. Espiritu[12], Shengjie Gao[179], Yi Huang[157,257], Christopher M. Lalansingh[12], Jon W. Teague[1], Michael C. Wendl[27,158,159], Federico Abascal[1], Gary D. Bader[13], Pratiti Bandopadhayay[3,258,259], Jonathan Barenboim[12], Søren Brunak[260,261], Joana Carlevaro-Fita[262,263,264], Dimple Chakravarty[265,266], Calvin Wing Yiu Chan[28,128], Jung Kyoon Choi[267], Klev Diamanti[268], J. Lynn Fink[40,269], Joan Frigola[204], Carlo Gambacorti-Passerini[270], Dale W. Garsed[271], Nicholas J. Haradhvala[3,19], Arif O. Harmanci[64,272], Mohamed Helmy[170], Carl Herrmann[28,30,273], Asger Hobolth[153,205], Ermin Hodzic[155], Chen Hong[127,128], Keren Isaev[12,18], Jose M. G. Izarzugaza[260], Rory Johnson[263,274], Randi Istrup Juul[152], Jaegil Kim[3], Jong K. Kim[275], Jan Komorowski[268,276], Andrés Lanzós[263,264,274], Erik Larsson[112], Donghoon Lee[64], Shantao Li[64], Xiaotong Li[64], Ziao Lin[3,277], Eric Minwei Liu[71,73,278], Lucas Lochovsky[63,64,185], Shaoke Lou[63,64], Tobias Madsen[152], Kathleen Marchal[279,280], Alexander Martinez-Fundichely[71,72,73], Patrick D. McGillivray[63], William Meyerson[64,281], Marta Paczkowska[12], Keunchil Park[282,283], Kiejung Park[284], Tirso Pons[285], Sergio Pulido-Tamayo[279,280], Iker Reyes-Salazar[87], Matthew A. Reyna[286], Mark A. Rubin[274,287,288,289,290], Leonidas Salichos[63,64], Chris Sander[112,156,291,292], Steven E. Schumacher[3,203], Mark Shackleton[271], Ciyue Shen[292,293], Raunak Shrestha[146], Shimin Shuai[12,13], Tatsuhiko Tsunoda[233,294,295,296], Husen M. Umer[268,297], Liis Uusküla-Reimand[298,299], Lieven P. C. Verbeke[280,300], Claes Wadelius[301], Lina Wadi[12], Jonathan Warrell[63,64], Guanming Wu[302], Jun Yu[303], Jing Zhang[64], Xuanping Zhang[157,304], Yan Zhang[64,305,306], Zhongming Zhao[307], Lihua Zou[308], Michael S. Lawrence[3,19,233], Benjamin J. Raphael[286], Peter J. Bailey[186], David Craft[3,309], Mary J. Goldman[132], Hiroyuki Aburatani[310], Hans Binder[311,312], Huy Q. Dinh[313], Simon C. Heath[49,74], Steve Hoffmann[311,312,314,315], Charles David Imbusch[127], Helene Kretzmer[312,315], Peter W. Laird[316], Jose I. Martin-Subero[76,317], Genta Nagae[310,318], Hui Shen[319], Qi Wang[246], Dieter Weichenhan[320], Wanding Zhou[319], Benjamin P. Berman[313,321,322], Benedikt Brors[127,149,323], Christoph Plass[320], Kadir C. Akdemir[140], David D. L. Bowtell[271], Kathleen H. Burns[324,325], John Busanovich[3,326], Kin Chan[327], Ana Dueso-Barroso[40], Paul A. Edwards[328,329], Dariush Etemadmoghadam[271], James E. Haber[330], David T. W. Jones[331,332], Young Seok Ju[1,267], Marat D. Kazanov[333,334,335], Youngil Koh[336,337], Kiran Kumar[3], Eunjung Alice Lee[338], Jake June-Koo Lee[94,95], Andy G. Lynch[328,329,339], Geoff Macintyre[328], Florian Markowetz[328,329], Fabio C. P. Navarro[63], John V. Pearson[340,341], Karsten Rippe[120], Ralph Scully[342], Izar Villasante[40], Nicola Waddell[340,341], Lixing Yang[343], Xiaotong Yao[165,344], Sung-Soo Yoon[337], Cheng-Zhong Zhang[3,6,156], Erik N. Bergstrom[345], Arnoud Boot[195,346], Kyle Covington[34], Akihiro Fujimoto[81], Mi Ni Huang[195,346], S. M. Ashiqul Islam[101], John R. McPherson[195,346], Sandro Morganella[1], Ville Mustonen[347,348,349], Alvin Wei Tian Ng[350], Stephenie D. Prokopec[12], Ignacio Vázquez-García[1,67,351,352], Yang Wu[195,346], Fouad Yousif[12], Willie Yu[353], Steven G. Rozen[195,196,346], Vasilisa A. Rudneva[8], Suyash S. Shringarpure[45], Daniel J. Turner[106], Tian Xia[354], Gurnit Atwal[12,13,171], David K. Chang[186,355], Susanna L. Cooke[186], Bishoy M. Faltas[117], Syed Haider[12], Vera B. Kaiser[356], Rosa Karlić[357], Mamoru Kato[358], Kirsten Kübler[3,6,19], Adam Margolin[38], Sancha Martin[1,359], Serena Nik-Zainal[1,360,361,362], Christine P'ng[12], Colin A. Semple[356], Jaclyn Smith[38], Ren X. Sun[12], Kevin Thai[15], Derek W. Wright[363,364], Ke Yuan[328,359,365], Andrew V. Biankin[186,355,366,367], Levi Garraway[156], Sean M. Grimmond[368], David J. Adams[1], Pavana Anur[369], Shaolong Cao[32], Elizabeth L. Christie[271], Marek Cmero[370,371,372], Yupeng Cun[373], Kevin J. Dawson[1], Stefan C. Dentro[1,65,91], Amit G. Deshwar[374], Nilgun Donmez[146,155], Ruben M. Drews[328], Moritz Gerstung[7,8], Gavin Ha[3], Kerstin Haase[91], Lara Jerman[8,375], Yuan Ji[376,377], Clemency Jolly[91], Juhee Lee[378], Henry Lee-Six[1], Salem Malikic[146,155], Thomas J. Mitchell[1,329,379], Quaid D. Morris[171,380], Layla Oesper[381], Martin Peifer[373], Myron Peto[382], Daniel Rosebrock[3], Yulia Rubanova[36,171], Adriana Salcedo[12], Subhajit Sengupta[383], Ruian Shi[380], Seung Jun Shin[182], Oliver Spiro[3], Shankar Vembu[380,384], Jeffrey A. Wintersinger[169,170,171], Tsun-Po Yang[373], Kaixian Yu[385], Hongtu Zhu[386,387], Paul T. Spellman[388], John N. Weinstein[136,137], Yiwen Chen[32], Masashi Fujita[81], Leng Han[304], Takanori Hasegawa[39], Mitsuhiro Komura[39], Jun Li[32], Shinichi Mizuno[389], Eigo Shimizu[39], Yumeng Wang[32,390], Yanxun Xu[391], Rui Yamaguchi[39], Fan Yang[380], Yang Yang[304], Christopher J. Yoon[267], Yuan Yuan[32], Han Liang[32], Malik Alawi[392,393], Ivan Borozan[12], Daniel S. Brewer[394,395], Colin S. Cooper[395,396,397], Nikita Desai[15], Adam Grundhoff[392,398], Murat Iskar[399], Xiaoping Su[400], Marc Zapatka[399], Peter Lichter[148,399], Kathryn Alsop[271], Timothy J. C. Bruxner[269], Angelika N. Christ[269], Stephen M. Cordner[401], Prue A. Cowin[402], Ronny Drapkin[403], Sian Fereday[402], Joshy George[185], Anne Hamilton[402], Oliver Holmes[340,341], Jillian A. Hung[404,405], Karin S. Kassahn[269,406], Stephen H. Kazakoff[340,341], Catherine J. Kennedy[407,408], Conrad R. Leonard[340,341], Linda Mileshkin[271], David K. Miller[269,355,409], Gisela Mir Arnau[402], Chris Mitchell[402], Felicity Newell[340,341], Katia Nones[340,341], Ann-Marie Patch[340,341], Michael C. Quinn[340,341], Darrin F. Taylor[269], Heather Thorne[402], Nadia Traficante[402],

Ravikiran Vedururu[402], Nick M. Waddell[341], Paul M. Waring[410], Scott Wood[340,341], Qinying Xu[340,341], Anna deFazio[411,412,413], Matthew J. Anderson[269], Davide Antonello[414], Andrew P. Barbour[415,416], Claudio Bassi[414], Samantha Bersani[417], Ivana Cataldo[417,418], Lorraine A. Chantrill[355,419], Yoke-Eng Chiew[411], Angela Chou[355,420], Sara Cingarlini[229], Nicole Cloonan[421], Vincenzo Corbo[418,422], Maria Vittoria Davi[423], Fraser R. Duthie[186,424], Anthony J. Gill[355,420], Janet S. Graham[186,425], Ivon Harliwong[269], Nigel B. Jamieson[186,367,426], Amber L. Johns[355,409], James G. Kench[355,420,427], Luca Landoni[414], Rita T. Lawlor[418], Andrea Mafficini[418], Neil D. Merrett[414,428], Marco Miotto[414], Elizabeth A. Musgrove[186], Adnan M. Nagrial[355], Karin A. Oien[410,429], Marina Pajic[355], Mark Pinese[430], Alan J. Robertson[269], Ilse Rooman[355], Borislav C. Rusev[418], Jaswinder S. Samra[414,420], Maria Scardoni[417], Christopher J. Scarlett[355,431], Aldo Scarpa[418], Elisabetta Sereni[414], Katarzyna O. Sikora[418], Michele Simbolo[422], Morgan L. Taschuk[15], Christopher W. Toon[355], Caterina Vicentini[418], Jianmin Wu[355], Nikolajs Zeps[432,433], Andreas Behren[434], Hazel Burke[435], Jonathan Cebon[434], Rebecca A. Dagg[436], Ricardo De Paoli-Iseppi[437], Ken Dutton-Regester[340], Matthew A. Field[438], Anna Fitzgerald[439], Peter Hersey[435], Valerie Jakrot[435], Peter A. Johansson[340], Hojabr Kakavand[437], Richard F. Kefford[440], Loretta M. S. Lau[441], Georgina V. Long[442], Hilda A. Pickett[441], Antonia L. Pritchard[340], Gulietta M. Pupo[443], Robyn P. M. Saw[442], Sarah-Jane Schramm[444], Catherine A. Shang[439], Ping Shang[442], Andrew J. Spillane[442], Jonathan R. Stretch[442], Varsha Tembe[411,444], John F. Thompson[442], Ricardo E. Vilain[445], James S. Wilmott[442], Jean Y. Yang[446], Nicholas K. Hayward[340,435], Graham J. Mann[411,447], Richard A. Scolyer[412,442,445,448], John Bartlett[449,450], Prashant Bavi[451], Dianne E. Chadwick[452], Michelle Chan-Seng-Yue[451], Sean Cleary[451,453], Ashton A. Connor[453,454], Karolina Czajka[241], Robert E. Denroche[451], Neesha C. Dhani[455], Jenna Eagles[241], Steven Gallinger[451,453,454], Robert C. Grant[451,454], David Hedley[455], Michael A. Hollingsworth[456], Gun Ho Jang[451], Jeremy Johns[241], Sangeetha Kalimuthu[451], Sheng-Ben Liang[457], Ilinca Lungu[451,458], Xuemei Luo[12], Faridah Mbabaali[241], Treasa A. McPherson[454], Jessica K. Miller[241], Malcolm J. Moore[455], Faiyaz Notta[451,459], Danielle Pasternack[241], Gloria M. Petersen[460], Michael H. A. Roehrl[18,451,461,462,463], Michelle Sam[241], Iris Selander[454], Stefano Serra[410], Sagedeh Shahabi[457], Sarah P. Thayer[456], Lee E. Timms[241], Gavin W. Wilson[12,451], Julie M. Wilson[451], Bradly G. Wouters[464], John D. McPherson[241,451,465], Timothy A. Beck[15,466], Vinayak Bhandari[12], Colin C. Collins[146], Neil E. Fleshner[467], Natalie S. Fox[12], Michael Fraser[12], Lawrence E. Heisler[468], Emilie Lalonde[12], Julie Livingstone[12], Alice Meng[469], Veronica Y. Sabelnykova[12], Yu-Jia Shiah[12], Theodorus Van der Kwast[470], Robert G. Bristow[18,471,472,473,474], Shuai Ding[475], Daiming Fan[476], Lin Li[179], Yongzhan Nie[476,477], Xiao Xiao[157], Rui Xing[222,478], Shanlin Yang[475], Yingyan Yu[479], Yong Zhou[179], Rosamonde E. Banks[480], Guillaume Bourque[481,482], Paul Brennan[483], Louis Letourneau[484], Yasser Riazalhosseini[482], Ghislaine Scelo[483], Naveen Vasudev[480,485], Juris Viksna[486], Mark Lathrop[482], Jörg Tost[487], Sung-Min Ahn[488], Samuel Aparicio[489], Laurent Arnould[490], M. R. Aure[491], Shriram G. Bhosle[1], Ewan Birney[7], Ake Borg[492], Sandrine Boyault[493], Arie B. Brinkman[494], Jane E. Brock[495], Annegien Broeks[496], Anne-Lise Børresen-Dale[491], Carlos Caldas[497,498], Suet-Feung Chin[497,498], Helen Davies[1,360,361], Christine Desmedt[499,500], Luc Dirix[501], Serge Dronov[1], Anna Ehinger[502], Jorunn E. Eyfjord[503], Aquila Fatima[203], John A. Foekens[504], P. Andrew Futreal[505], Øystein Garred[506,507], Dilip D. Giri[508], Dominik Glodzik[1], Dorthe Grabau[509], Holmfridur Hilmarsdottir[503], Gerrit K. Hooijer[510], Jocelyne Jacquemier[511], Se Jin Jang[512], Jon G. Jonasson[503], Jos Jonkers[513], Hyung-Yong Kim[511], Tari A. King[514,515], Stian Knappskog[1,516], Gu Kong[511], Savitri Krishnamurthy[517], Sunil R. Lakhani[518], Anita Langerød[491], Denis Larsimont[519], Hee Jin Lee[512], Jeong-Yeon Lee[520], Ming Ta Michael Lee[505], Ole Christian Lingjærde[521], Gaetan MacGrogan[522], John W. M. Martens[504], Sarah O'Meara[1], Iris Pauporté[214], Sarah Pinder[523], Xavier Pivot[524], Elena Provenzano[525], Colin A. Purdie[526], Manasa Ramakrishna[1], Kamna Ramakrishnan[1], Jorge Reis-Filho[508], Andrea L. Richardson[203], Markus Ringnér[492], Javier Bartolomé Rodriguez[40], F. Germán Rodríguez-González[261], Gilles Romieu[527], Roberto Salgado[410], Torill Sauer[521], Rebecca Shepherd[1], Anieta M. Sieuwerts[504], Peter T. Simpson[518], Marcel Smid[504], Christos Sotiriou[234], Paul N. Span[528], Ólafur Andri Stefánsson[529], Alasdair Stenhouse[530], Henk G. Stunnenberg[180,531], Fred Sweep[532], Benita Kiat Tee Tan[533], Gilles Thomas[534], Alastair M. Thompson[530], Stefania Tommasi[535], Isabelle Treilleux[536,537], Andrew Tutt[203], Naoto T. Ueno[387], Steven Van Laere[501], Gert G. Van den Eynden[501], Peter Vermeulen[501], Alain Viari[418], Anne Vincent-Salomon[531], Bernice H. Wong[538], Lucy Yates[1], Xueqing Zou[1], Carolien H. M. van Deurzen[539], Marc J. van de Vijver[410], Laura van't Veer[540], Ole Ammerpohl[541,542,543], Sietse Aukema[542,543,544], Anke K. Bergmann[545], Stephan H. Bernhart[311,312,315], Arndt Borkhardt[546], Christoph Borst[547], Birgit Burkhardt[548], Alexander Claviez[549], Maria Elisabeth Goebler[550], Andrea Haake[541], Siegfried Haas[547], Martin Hansmann[551], Jessica I. Hoell[546], Michael Hummel[552], Dennis Karsch[553], Wolfram Klapper[544], Michael Kneba[553], Markus Kreuz[554], Dieter Kube[555], Ralf Küppers[556], Dido Lenze[552], Markus Loeffler[554], Cristina López[80,541], Luisa Mantovani-Löffler[557], Peter Möller[558], German Ott[559], Bernhard Radlwimmer[399], Julia Richter[541,544], Marius Rohde[560], Philip C. Rosenstiel[561], Andreas Rosenwald[562], Markus B. Schilhabel[561], Stefan Schreiber[563], Peter F. Stadler[311,312,315], Peter Staib[564], Stephan Stilgenbauer[565], Stephanie Sungalee[8], Monika Szczepanowski[544], Umut H. Toprak[30,566], Lorenz H. P. Trümper[555], Rabea Wagener[80,541], Thorsten Zenz[149], Volker Hovestadt[399], Christof von Kalle[120], Marcel Kool[246,331], Andrey Korshunov[246], Pablo Landgraf[567,568], Hans Lehrach[569], Paul A. Northcott[570], Stefan M. Pfister[246,331,571], Guido Reifenberger[568], Hans-Jörg Warnatz[569], Stephan Wolf[572], Marie-Laure Yaspo[569], Yassen Assenov[573], Clarissa Gerhauser[320], Sarah Minner[574], Thorsten Schlomm[99,575], Ronald Simon[576], Guido Sauter[576], Holger Sültmann[149,577], Nidhan K. Biswas[578], Arindam Maitra[578], Partha P. Majumder[578], Rajiv Sarin[579], Stefano Barbi[422], Giada Bonizzato[418], Cinzia Cantù[418], Angelo P. Dei Tos[580], Matteo Fassan[581], Sonia Grimaldi[418], Claudio Luchini[417], Giuseppe Malleo[414], Giovanni Marchegiani[414], Michele Milella[229], Salvatore Paiella[414], Antonio Pea[414], Paolo Pederzoli[414], Andrea Ruzzenente[414], Roberto Salvia[414], Nicola Sperandio[418], Yasuhito Arai[226], Natsuko Hama[226], Nobuyoshi Hiraoka[582], Fumie Hosoda[226], Hiromi Nakamura[226], Hidenori Ojima[583], Takuji Okusaka[584], Yasushi Totoki[226], Tomoko Urushidate[227], Masashi Fukayama[585], Shumpei Ishikawa[586], Hitoshi Katai[587], Hiroto Katoh[586], Daisuke Komura[586], Hirofumi Rokutan[358], Mihoko Saito-Adachi[358], Akihiro Suzuki[310,588], Hirokazu Taniguchi[589], Kenji Tatsuno[310], Tetsuo Ushiku[585], Shinichi Yachida[226,590], Shogo Yamamoto[310], Hiroshi Aikata[591], Koji Arihiro[591], Shun-ichi Ariizumi[592], Kazuaki Chayama[591], Mayuko Furuta[81], Kunihito Gotoh[593], Shinya Hayami[594], Satoshi Hirano[595], Yoshiiku Kawakami[591], Kazuhiro Maejima[81], Toru Nakamura[595], Kaoru Nakano[81], Hideki Ohdan[591], Aya Sasaki-Oku[81], Hiroko Tanaka[39], Masaki Ueno[594], Masakazu Yamamoto[592], Hiroki Yamaue[594], Su Pin Choo[596], Ioana Cutcutache[195,346], Narong Khuntikeo[414,597], Choon Kiat Ong[598], Chawalit Pairojkul[410], Irinel Popescu[599], Keun Soo Ahn[600], Marta Aymerich[601], Armando Lopez-Guillermo[602], Carlos López-Otín[603], Xose S. Puente[603], Elias Campo[604,605], Fernanda Amary[606], Daniel Baumhoer[607], Sam Behjati[1], Bodil Bjerkehagen[607,608], P. A. Futreal[505], Ola Myklebost[516], Nischalan Pillay[609], Patrick Tarpey[610], Roberto Tirabosco[611], Olga Zaikova[612], Adrienne M. Flanagan[613], Jacqueline Boultwood[614], David T. Bowen[1], Mario Cazzola[615], Anthony R. Green[329], Eva Hellstrom-Lindberg[616], Luca Malcovati[615], Jyoti Nangalia[617], Elli Papaemmanuil[1], Paresh Vyas[340,618], Yeng Ang[619], Hugh Barr[620], Duncan Beardsmore[621], Matthew Eldridge[328], James Gossage[622], Nicola Grehan[361], George B. Hanna[623], Stephen J. Hayes[624,625], Ted R. Hupp[626], David Khoo[627], Jesper Lagergren[616,628], Laurence B. Lovat[188], Shona MacRae[136], Maria O'Donovan[361], J. Robert O'Neill[629], Simon L. Parsons[630], Shaun R. Preston[631], Sonia Puig[632], Tom Roques[633], Grant Sanders[24], Sharmila Sothi[634], Simon Tavaré[328], Olga Tucker[635], Richard Turkington[636], Timothy J. Underwood[637], Ian Welch[638], Rebecca C. Fitzgerald[361], Daniel M. Berney[639], Johann S. De Bono[396], Declan Cahill[640], Niedzica Camacho[396], Nening M. Dennis[640], Tim Dudderidge[640,641], Sandra E. Edwards[396], Cyril Fisher[640], Christopher S. Foster[642,643], Mohammed Ghori[1], Pelvender Gill[618], Vincent J. Gnanapragasam[379,644], Gunes Gundem[278], Freddie C. Hamdy[645], Steve Hawkins[328], Steven Hazell[640], William Howat[379], William B. Isaacs[646], Katalin Karaszi[618], Jonathan D. Kay[188], Vincent Khoo[640], Zsofia Kote-Jarai[396], Barbara Kremeyer[1], Pardeep Kumar[640], Adam Lambert[618], Daniel A. Leongamornlert[1,396], Naomi Livni[640], Yong-Jie Lu[639,647], Hayley J. Luxton[188], Luke Marsden[618], Charlie E. Massie[328], Lucy Matthews[396], Erik Mayer[640,648], Ultan McDermott[1], Sue Merson[396], David E. Neal[328,379], Anthony Ng[649], David Nicol[640], Christopher Ogden[640], Edward W. Rowe[640], Nimish C. Shah[379], Sarah Thomas[640], Alan Thompson[640], Clare Verrill[618,650], Tapio Visakorpi[126], Anne Y. Warren[379,651], Hayley C. Whitaker[188], Hongwei Zhang[647], Nicholas van As[640], Rosalind A. Eeles[396,640], Adam Abeshouse[278], Nishant Agrawal[162], Rehan Akbani[361,652], Hikmat Al-Ahmadie[278], Monique Albert[450], Kenneth Aldape[400,653], Adrian Ally[654], Elizabeth L. Appelbaum[27,188], Joshua Armenia[655], Sylvia Asa[630,656], J. Todd Auman[657], Miruna Balasundaram[654], Saianand Balu[24], Jill Barnholtz-Sloan[658,659], Oliver F. Bathe[660,661], Stephen B. Baylin[123,641], Christopher Benz[662], Andrew Berchuck[663], Mario Berrios[664], Darell Bigner[665], Michael Birrer[19], Tom Bodenheimer[24], Lori Boice[632], Moiz S. Bootwalla[664], Marcus Bosenberg[666], Reanne Bowlby[654], Jeffrey Boyd[667], Russell R. Broaddus[400], Malcolm Brock[668], Denise Brooks[654], Susan Bullman[3,167], Samantha J. Caesar-Johnson[231], Thomas E. Carey[669], Rebecca Carlsen[654], Robert Cerfolio[670], Vishal S. Chandan[671], Hsiao-Wei Chen[619,655], Andrew D. Cherniack[3,156,167], Jeremy Chien[672], Juok Cho[3], Eric Chuah[654], Carrie Cibulskis[3], Leslie Cope[673], Matthew G. Cordes[27,633], Erin Curley[674], Bogdan Czerniak[400,627], Ludmila Danilova[673], Ian J. Davis[675], Timothy Defreitas[3], John A. Demchok[231], Noreen Dhalla[654], Rajiv Dhir[676], HarshaVardhan Doddapaneni[34], Adel El-Naggar[400,627], Ina Felau[231], Martin L. Ferguson[677], Gaetano Finocchiaro[678], Kwun M. Fong[679], Scott Frazer[3], William Friedman[680], Catrina C. Fronick[27,633], Lucinda A. Fulton[27], Stacey B. Gabriel[3], Jianjiong Gao[655], Nils Gehlenborg[3,681], Jeffrey E. Gershenwald[682,683], Ronald Ghossein[508], Nasra H. Giama[684], Richard A. Gibbs[34], Carmen Gomez[685], Ramaswamy Govindan[26], D. Neil Hayes[24,686,687], Apurva M. Hegde[136,137], David I. Heiman[3], Zachary Heins[278], Austin J. Hepperla[24], Andrea Holbrook[664], Robert A. Holt[654], Alan P. Hoyle[24], Ralph H. Hruban[673], Jianhong Hu[34], Mei Huang[632], David Huntsman[688], Jason Huse[278], Christine A. Iacobuzio-Donahue[508], Michael Ittmann[689,690], Joy C. Jayaseelan[34], Stuart R. Jefferys[24], Corbin D. Jones[691], Steven J. M. Jones[692], Hartmut Juhl[693], Koo Jeong Kang[694], Beth Karlan[695], Katayoon Kasaian[692], Electron Kebebew[696,697], Hark Kyun Kim[698], Viktoriya Korchina[34], Ritika Kundra[619,655], Phillip H. Lai[664], Eric Lander[3], Xuan Le[699], Darlene Lee[654], Douglas A. Levine[278,700], Lora Lewis[34], Tim Ley[701], Haiyan Irene Li[654], Pei Lin[3], W. M. Linehan[702], Fei Fei Liu[380], Yiling Lu[137], Lisa Lype[703], Yussanne Ma[654], Dennis T. Maglinte[664,704], Elaine R. Mardis[27,667,705], Jeffrey Marks[414,706], Marco A. Marra[654], Thomas J. Matthew[57], Michael Mayo[654], Karen McCune[707], Samuel R. Meier[3], Shaowu Meng[24], Piotr A. Mieczkowski[23], Tom Mikkelsen[708], Christopher A. Miller[27], Gordon B. Mills[709], Richard A. Moore[654], Carl Morrison[410,710], Lisle E. Mose[24], Catherine D. Moser[684], Andrew J. Mungall[654], Karen Mungall[654], David Mutch[711], Donna M. Muzny[712], Jerome Myers[713], Yulia Newton[37], Michael S. Noble[3], Peter O'Donnell[714], Brian Patrick O'Neill[715], Angelica Ochoa[278], Joong-Won Park[716], Joel S. Parker[717], Harvey Pass[718], Alessandro Pastore[112], Nathan A. Pennell[719], Charles M. Perou[720], Nicholas Petrelli[721], Olga Potapova[722], Janet S. Rader[723], Suresh Ramalingam[724], W. Kimryn Rathmell[725], Victor Reuter[508], Sheila M. Reynolds[703], Matthew Ringel[726], Jeffrey Roach[727], Lewis R. Roberts[684], A. Gordon Robertson[654], Sara Sadeghi[654], Charles Saller[728], Francisco Sanchez-Vega[619,655], Dirk Schadendorf[148,729], Jacqueline E. Schein[654], Heather K. Schmidt[27], Nikolaus Schultz[655], Raja Seethala[730], Yasin Senbabaoglu[112], Troy Shelton[674], Yan Shi[24], Juliann Shih[3,167], Ilya Shmulevich[703], Craig Shriver[731], Sabina Signoretti[167,263,732], Janae V. Simons[24], Samuel Singer[414,733], Payal Sipahimalani[654], Tara J. Skelly[23], Karen Smith-McCune[707], Nicholas D. Socci[112], Matthew G. Soloway[717], Anil K. Sood[734], Angela Tam[654], Donghui Tan[23], Roy Tarnuzzer[231], Nina Thiessen[654], R. Houston Thompson[735], Leigh B. Thorne[632], Ming Tsao[630,656], Christopher Umbricht[324,621,736], David J. Van Den Berg[664], Erwin G. Van Meir[737], Umadevi Veluvolu[23], Douglas Voet[3], Linghua Wang[34], Paul Weinberger[738],

Article

Daniel J. Weisenberger[664], Dennis Wigle[739], Matthew D. Wilkerson[23], Richard K. Wilson[27,740], Boris Winterhoff[741], Maciej Wiznerowicz[742,743], Tina Wong[27,654], Winghing Wong[744], Liu Xi[34], Christina Yau[662], Hailei Zhang[3], Hongxin Zhang[655] & Jiashan Zhang[231]

[1]Wellcome Sanger Institute, Hinxton, UK. [2]Department of Haematology, University of Cambridge, Cambridge, UK. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. [5]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [6]Harvard Medical School, Boston, MA, USA. [7]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [8]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. [9]Biomolecular Engineering Department, University of California Santa Cruz, Santa Cruz, CA, USA. [10]Adaptive Oncology Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [11]International Cancer Genome Consortium (ICGC)/ICGC Accelerating Research in Genomic Oncology (ICGC-ARGO) Secretariat, Toronto, Ontario, Canada. [12]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [13]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [14]Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, USA. [15]Genome Informatics Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [16]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada. [17]Genome Informatics, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [18]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [19]Massachusetts General Hospital, Boston, MA, USA. [20]Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. [21]University of California Los Angeles, Los Angeles, CA, USA. [22]Department of Pathology, Department of Genomic Medicine and Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [23]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [24]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [25]The Hospital for Sick Children, Toronto, Ontario, Canada. [26]Alvin J. Siteman Cancer Center, Washington University School of Medicine, St Louis, MO, USA. [27]The McDonnell Genome Institute, Washington University, St Louis, MO, USA. [28]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [29]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center, Heidelberg, Germany. [30]Institute of Pharmacy and Molecular Biotechnology, and BioQuant, Heidelberg University, Heidelberg, Germany. [31]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [32]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [33]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [34]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [35]Department of Genetics and Department of Medicine, Washington University in St Louis, St Louis, MO, USA. [36]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [37]University of California Santa Cruz, Santa Cruz, CA, USA. [38]Computational Biology Program, Oregon Health & Science University, Portland, OR, USA. [39]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [40]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [41]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway. [42]Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [43]Department of Zoology, Genetics and Physical Anthropology, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [44]The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. [45]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [46]Annai Systems, Carlsbad, CA, USA. [47]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [48]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [49]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [50]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [51]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [52]Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland. [53]Department of Ophthalmology, Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA. [54]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra (UPF), Barcelona, Spain. [55]Department of Veterinary Medicine, Transmissible Cancer Group, University of Cambridge, Cambridge, UK. [56]Department of Biochemistry, College of Medicine, Ewha Womans University, Seoul, South Korea. [57]Division of Oncology, Washington University School of Medicine, St Louis, MO, USA. [58]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [59]The First Affiliated Hospital, Xi'an Jiaotong University, Xi'an, China. [60]Independent Consultant, Wellesley, MA, USA. [61]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [62]Biobyte Solutions, Heidelberg, Germany. [63]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [64]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [65]Big Data Institute, Li Ka Shing Centre, University of Oxford, Oxford, UK. [66]Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. [67]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [68]The McDonnell Genome Institute at Washington University School of Medicine, and Department of Genetics and Department of Medicine, Siteman Cancer Center, Washington University in St Louis, St Louis, MO, USA. [69]Department of Computer Science, Yale University, New Haven, CT, USA. [70]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [71]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [72]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [73]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [74]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [75]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra (UPF), Barcelona, Spain. [76]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [77]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. [78]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [79]Human Genetics, University of Kiel, Kiel, Germany. [80]Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany. [81]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [82]Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK. [83]Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK. [84]Quantitative Genomics Laboratories (qGenomics), Barcelona, Spain. [85]Sage Bionetworks, Seattle, WA, USA. [86]Department of Biochemistry and Molecular Medicine, University of Montreal, Montreal, Quebec, Canada. [87]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [88]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India. [89]Research Program on Biomedical Informatics, Universitat Pompeu Fabra (UPF), Barcelona, Spain. [90]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [91]The Francis Crick Institute, London, UK. [92]University of Leuven, Leuven, Belgium. [93]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. [94]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [95]Ludwig Center at Harvard Medical School, Boston, MA, USA. [96]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [97]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [98]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [99]Department of Urology, Charité Universitätsmedizin Berlin, Berlin, Germany. [100]Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. [101]Department of Bioengineering and Department of Cellular and Molecular Medicine, Moores Cancer Center, University of California San Diego, La Jolla, CA, USA. [102]Department of Genetics, Microbiology and Statistics, University of Barcelona, IRSJD, IBUB, Barcelona, Spain. [103]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [104]Research Group on Statistics, Econometrics and Health (GRECS), UdG, Barcelona, Spain. [105]Oxford Nanopore Technologies, New York, NY, USA. [106]Applications Department, Oxford Nanopore Technologies, Oxford, UK. [107]School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA, USA. [108]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [109]Department of Medical and Clinical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland. [110]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT, USA. [111]Applied Tumor Genomics Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland. [112]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [113]Department of Biology, ETH Zurich, Zurich, Switzerland. [114]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [115]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. [116]University Hospital Zurich, Zurich, Switzerland. [117]Weill Cornell Medical College, New York, NY, USA. [118]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. [119]German Cancer Consortium (DKTK), Partner site Berlin, Berlin, Germany. [120]German Cancer Research Center (DKFZ), Heidelberg, Germany. [121]Bakar Computational Health Sciences Institute and Department of Pediatrics, University of California, San Francisco, CA, USA. [122]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [123]Department of Oncology, The Johns Hopkins School of Medicine, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, MD, USA. [124]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [125]Department of Medicine and Moores Cancer Center, Division of Biomedical Informatics, UC San Diego School of Medicine, San Diego, CA, USA. [126]Faculty of Medicine and Health Technology, Tampere University and Tays Cancer Center, Tampere University Hospital, Tampere, Finland. [127]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [128]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [129]Centre for Law and Genetics, University of Tasmania, Hobart, Tasmania, Australia. [130]Centre of Genomics and Policy, McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada. [131]Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany. [132]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [133]CIBIO/InBIO, Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Portugal. [134]Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [135]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [136]Cancer Unit, MRC University of Cambridge, Cambridge, UK. [137]Department of Bioinformatics and Computational Biology and Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [138]Center for Digital Health, Berlin Institute of Health (BIH) and Charitè–Universitätsmedizin Berlin, Berlin, Germany. [139]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer

Research Center (DKFZ), Heidelberg, Germany. [140]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [141]Department of Genetics and Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, USA. [142]Heidelberg University, Heidelberg, Germany. [143]New BIH Digital Health Center, Berlin Institute of Health (BIH) and Charité–Universitätsmedizin Berlin, Berlin, Germany. [144]Department of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain. [145]Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada. [146]Vancouver Prostate Centre, Vancouver, British Columbia, Canada. [147]Division of Life Science and Applied Genomics Center, Hong Kong University of Science and Technology, Hong Kong, China. [148]German Cancer Consortium (DKTK), Heidelberg, Germany. [149]National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. [150]Genome Integration Data Center, Syntekabio, Daejon, South Korea. [151]Massachusetts General Hospital Center for Cancer Research, Charlestown, MA, USA. [152]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark. [153]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. [154]Indiana University, Bloomington, IN, USA. [155]Simon Fraser University, Burnaby, British Columbia, Canada. [156]Dana-Farber Cancer Institute, Boston, MA, USA. [157]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. [158]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [159]Department of Mathematics, Washington University in St Louis, St Louis, MO, USA. [160]Department of Biological Oceanography, Leibniz Institute of Baltic Sea Research, Rostock, Germany. [161]Seven Bridges Genomics, Charlestown, MA, USA. [162]University of Chicago, Chicago, IL, USA. [163]Department of Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul, South Korea. [164]Samsung Genome Institute, Seoul, South Korea. [165]New York Genome Center, New York, NY, USA. [166]Weill Cornell Medicine, New York, NY, USA. [167]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [168]Rigshospitalet, Copenhagen, Denmark. [169]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [170]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [171]Vector Institute, Toronto, Ontario, Canada. [172]Department of Medical Genetics, College of Medicine, Hallym University, Chuncheon, South Korea. [173]Department of Biology, ETH Zurich, Zurich, Switzerland. [174]University Hospital Zurich, Zurich, Switzerland. [175]Peking University, Beijing, China. [176]School of Life Sciences, Peking University, Beijing, China. [177]Computational and Systems Biology, Genome Institute of Singapore, Singapore, Singapore. [178]School of Computing, National University of Singapore, Singapore, Singapore. [179]BGI-Shenzhen, Shenzhen, China. [180]China National GeneBank-Shenzhen, Shenzhen, China. [181]Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [182]Korea University, Seoul, South Korea. [183]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [184]Quantitative & Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX, USA. [185]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [186]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden, UK. [187]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. [188]University College London, London, UK. [189]Genome Institute of Singapore, Singapore, Singapore. [190]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [191]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [192]O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA. [193]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. [194]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. [195]Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore. [196]SingHealth, Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore, Singapore. [197]Institute of Molecular and Cell Biology, Singapore, Singapore. [198]Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore, Singapore. [199]Department of Medicine, Baylor College of Medicine, Houston, TX, USA. [200]National Cancer Centre Singapore, Singapore, Singapore. [201]BIOPIC, ICG and College of Life Sciences, Peking University, Beijing, China. [202]Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. [203]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [204]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [205]Department of Mathematics, Aarhus University, Aarhus, Denmark. [206]Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Spain. [207]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [208]King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia. [209]DLR Project Management Agency, Bonn, Germany. [210]Genome Canada, Ottawa, Ontario, Canada. [211]Instituto Carlos Slim de la Salud, Mexico City, Mexico. [212]Federal Ministry of Education and Research, Berlin, Germany. [213]Institut Gustave Roussy, Villejuif, France. [214]Institut National du Cancer (INCA), Boulogne-Billancourt, France. [215]The Wellcome Trust, London, UK. [216]Prostate Cancer Canada, Toronto, Ontario, Canada. [217]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [218]Department of Biotechnology, Ministry of Science & Technology, Government of India, New Delhi, Delhi, India. [219]Science Writer, Garrett Park, MD, USA. [220]Cancer Research UK, London, UK. [221]Chinese Cancer Genome Consortium, Shenzhen, China. [222]Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. [223]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. [224]National Cancer Center, Tokyo, Japan. [225]German Cancer Aid, Bonn, Germany. [226]Division of Cancer Genomics, National Cancer Center Research Institute, National Cancer Center, Tokyo, Japan. [227]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan. [228]Japan Agency for Medical Research and Development, Chiyoda-ku, Tokyo, Japan. [229]Medical Oncology, University and Hospital Trust of Verona, Verona, Italy. [230]University of Verona, Verona, Italy. [231]National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [232]CAPHRI Research School, Maastricht University, Maastricht, The Netherlands. [233]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [234]University of California San Diego, San Diego, CA, USA. [235]PDXen Biosystems, Seoul, South Korea. [236]Electronics and Telecommunications Research Institute, Daejeon, South Korea. [237]Children's Hospital of Philadelphia, Philadelphia, PA, USA. [238]University of Melbourne Centre for Cancer Research, Melbourne, Victoria, Australia. [239]Syntekabio, Daejon, South Korea. [240]AbbVie, North Chicago, IL, USA. [241]Genomics Research Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [242]Department of Pediatric Immunology, Hematology and Oncology, University Hospital, Heidelberg, Germany. [243]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg, Germany. [244]Seven Bridges, Charlestown, MA, USA. [245]Health Sciences Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA. [246]Functional and Structural Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [247]Leidos Biomedical Research, McLean, VA, USA. [248]CSRA Incorporated, Fairfax, VA, USA. [249]Department of Internal Medicine, Stanford University, Stanford, CA, USA. [250]Clinical Bioinformatics, Swiss Institute of Bioinformatics, Geneva, Switzerland. [251]Institute for Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland. [252]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [253]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [254]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. [255]Office of Cancer Genomics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [256]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. [257]Geneplus-Shenzhen, Shenzhen, China. [258]Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA, USA. [259]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [260]Technical University of Denmark, Lyngby, Denmark. [261]University of Copenhagen, Copenhagen, Denmark. [262]Department for BioMedical Research, University of Bern, Bern, Switzerland. [263]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern, Switzerland. [264]Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland. [265]Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [266]Department of Urology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [267]Korea Advanced Institute of Science and Technology, Daejeon, South Korea. [268]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. [269]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. [270]University of Milano Bicocca, Monza, Italy. [271]Sir Peter MacCallum Department of Oncology, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. [272]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX, USA. [273]Health Data Science Unit, University Clinics, Heidelberg, Germany. [274]Department for Biomedical Research, University of Bern, Bern, Switzerland. [275]Research Core Center, National Cancer Centre Korea, Goyang-si, South Korea. [276]Institute of Computer Science, Polish Academy of Sciences, Warsawa, Poland. [277]Harvard University, Cambridge, MA, USA. [278]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [279]Department of Information Technology, Ghent University, Ghent, Belgium. [280]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [281]Yale School of Medicine, Yale University, New Haven, CT, USA. [282]Division of Hematology-Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea. [283]Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul, South Korea. [284]Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan, South Korea. [285]Spanish National Cancer Research Centre, Madrid, Spain. [286]Department of Computer Science, Princeton University, Princeton, NJ, USA. [287]Bern Center for Precision Medicine, University Hospital of Bern, University of Bern, Bern, Switzerland. [288]Englander Institute for Precision Medicine, Weill Cornell Medicine and New York Presbyterian Hospital, New York, NY, USA. [289]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [290]Pathology and Laboratory, Weill Cornell Medical College, New York, NY, USA. [291]cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [292]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [293]cBio Center, Dana-Farber Cancer Institute, Boston, MA, USA. [294]CREST, Japan Science and Technology Agency, Tokyo, Japan. [295]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo, Japan. [296]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [297]Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. [298]Department of Gene Technology, Tallinn University of Technology, Tallinn, Estonia. [299]Genetics & Genome Biology Program, SickKids Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada. [300]Department of Information Technology, Ghent University, Interuniversitair Micro-Electronica Centrum (IMEC), Ghent, Belgium. [301]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. [302]Oregon Health & Sciences University, Portland, OR, USA. [303]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, China. [304]The University of Texas Health Science Center at Houston, Houston, TX, USA. [305]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA.

[306]The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH, USA. [307]The University of Texas School of Biomedical Informatics (SBMI) at Houston, Houston, TX, USA. [308]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [309]Physics Division, Optimization and Systems Biology Lab, Massachusetts General Hospital, Boston, MA, USA. [310]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan. [311]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany. [312]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany. [313]Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [314]Computational Biology, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Jena, Germany. [315]Transcriptome Bioinformatics, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany. [316]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI, USA. [317]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [318]Research Center for Advanced Science and Technology, The University of Tokyo, Minato-ku, Tokyo, Japan. [319]Van Andel Research Institute, Grand Rapids, MI, USA. [320]Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [321]Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [322]The Hebrew University Faculty of Medicine, Jerusalem, Israel. [323]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [324]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [325]McKusick-Nathans Institute of Genetic Medicine, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [326]Foundation Medicine, Cambridge, MA, USA. [327]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. [328]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [329]University of Cambridge, Cambridge, UK. [330]Brandeis University, Waltham, MA, USA. [331]Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany. [332]Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. [333]A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. [334]Oncology and Immunology, Dmitry Rogachev National Research Center of Pediatric Hematology, Moscow, Russia. [335]Skolkovo Institute of Science and Technology, Moscow, Russia. [336]Center for Medical Innovation, Seoul National University Hospital, Seoul, South Korea. [337]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [338]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [339]School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. [340]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [341]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [342]Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA. [343]Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL, USA. [344]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY, USA. [345]Department of Bioengineering, and Department of Cellular and Molecular Medicine, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. [346]Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. [347]Department of Computer Science, University of Helsinki, Helsinki, Finland. [348]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [349]Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland. [350]Programme in Cancer & Stem Cell Biology, Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. [351]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK. [352]Department of Statistics, Columbia University, New York, NY, USA. [353]Duke-NUS Medical School, Singapore, Singapore. [354]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. [355]The Kinghorn Cancer Centre, Cancer Division, Garvan Institute of Medical Research, University of New South Wales, Sydney, New South Wales, Australia. [356]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, UK. [357]Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia. [358]Department of Bioinformatics, Division of Cancer Genomics, National Cancer Center Research Institute, National Cancer Center, Tokyo, Japan. [359]University of Glasgow, Glasgow, UK. [360]Academic Department of Medical Genetics, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [361]MRC Cancer Unit, University of Cambridge, Cambridge, UK. [362]The University of Cambridge School of Clinical Medicine, Cambridge, UK. [363]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [364]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden, UK. [365]School of Computing Science, University of Glasgow, Glasgow, UK. [366]South Western Sydney Clinical School, Faculty of Medicine, University of New South Wales, Liverpool, New South Wales, Australia. [367]West of Scotland Pancreatic Unit, Glasgow Royal Infirmary, Glasgow, UK. [368]University of Melbourne Centre for Cancer Research, Melbourne, Victoria, Australia. [369]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [370]Department of Surgery, University of Melbourne, Parkville, Victoria, Australia. [371]The Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia. [372]Walter + Eliza Hall Institute, Parkville, Victoria, Australia. [373]University of Cologne, Cologne, Germany. [374]The Edward S. Rogers Sr Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. [375]University of Ljubljana, Ljubljana, Slovenia. [376]Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA. [377]Research Institute, NorthShore University HealthSystem, Evanston, IL, USA. [378]Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, USA. [379]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [380]University of Toronto, Toronto, Ontario, Canada. [381]Department of Computer Science, Carleton College, Northfield, MN, USA. [382]Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [383]Center for Psychiatric Genetics, NorthShore University HealthSystem, Evanston, IL, USA. [384]Argmix Consulting, North Vancouver, British Columbia, Canada. [385]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [386]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [387]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [388]Molecular and Medical Genetics, Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. [389]Department of Health Sciences, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan. [390]Baylor College of Medicine, Houston, TX, USA. [391]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA. [392]Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. [393]University Medical Center Hamburg-Eppendorf, Bioinformatics Core, Hamburg, Germany. [394]Earlham Institute, Norwich, UK. [395]Norwich Medical School, University of East Anglia, Norwich, UK. [396]The Institute of Cancer Research, London, UK. [397]University of East Anglia, Norwich, UK. [398]German Center for Infection Research (DZIF), Partner Site Hamburg-Borstel-Lübeck-Riems, Hamburg, Germany. [399]Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [400]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [401]Victorian Institute of Forensic Medicine, Southbank, Victoria, Australia. [402]Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. [403]University of Pennsylvania, Philadelphia, PA, USA. [404]Centre for Cancer Research, The Westmead Institute for Medical Research, Sydney, New South Wales, Australia. [405]Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. [406]Genetics and Molecular Pathology, SA Pathology, Adelaide, South Australia, Australia. [407]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. [408]Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. [409]Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. [410]Department of Clinical Pathology, University of Melbourne, Melbourne, Victoria, Australia. [411]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. [412]Department of Gynaecological Oncology, Westmead Hospital, Sydney, New South Wales, Australia. [413]Westmead Clinical School, The Westmead Institute for Medical Research, Sydney, New South Wales, Australia. [414]Department of Surgery, Pancreas Institute, University and Hospital Trust of Verona, Verona, Italy. [415]Department of Surgery, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [416]Surgical Oncology Group, Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia. [417]Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona, Italy. [418]ARC-Net Centre for Applied Research on Cancer, University and Hospital Trust of Verona, Verona, Italy. [419]Illawarra Shoalhaven Local Health District L3 Illawarra Cancer Care Centre, Wollongong Hospital, Wollongong, New South Wales, Australia. [420]Department of Pathology, University of Sydney, Sydney, New South Wales, Australia. [421]School of Biological Sciences, The University of Auckland, Auckland, New Zealand. [422]Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy. [423]Department of Medicine, Section of Endocrinology, University and Hospital Trust of Verona, Verona, Italy. [424]Department of Pathology, Queen Elizabeth University Hospital, Glasgow, UK. [425]Department of Medical Oncology, Beatson West of Scotland Cancer Centre, Glasgow, UK. [426]Academic Unit of Surgery, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow Royal Infirmary, Glasgow, UK. [427]Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia. [428]Discipline of Surgery, Western Sydney University, Penrith, New South Wales, Australia. [429]Institute of Cancer Sciences, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [430]The Kinghorn Cancer Centre, Cancer Division, Garvan Institute of Medical Research, University of New South Wales, Sydney, New South Wales, Australia. [431]School of Environmental and Life Sciences, Faculty of Science, The University of Newcastle, Ourimbah, New South Wales, Australia. [432]Eastern Clinical School, Monash University, Melbourne, Victoria, Australia. [433]Epworth HealthCare, Richmond, Victoria, Australia. [434]Olivia Newton-John Cancer Research Institute, La Trobe University, Heidelberg, Victoria, Australia. [435]Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. [436]Children's Hospital at Westmead, The University of Sydney, Sydney, New South Wales, Australia. [437]Melanoma Institute Australia, The University of Sydney, Sydney, New South Wales, Australia. [438]Australian Institute of Tropical Health and Medicine, James Cook University, Douglas, Queensland, Australia. [439]Bioplatforms Australia, North Ryde, New South Wales, Australia. [440]Melanoma Institute Australia, Macquarie University, Wollstonecraft, New South Wales, Australia. [441]Children's Medical Research Institute, Sydney, New South Wales, Australia. [442]Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. [443]Centre for Cancer Research, The Westmead Millennium Institute for Medical Research, University of Sydney, Westmead Hospital, Sydney, New South Wales, Australia. [444]Translational Cancer Research Centre, The University of Sydney at the Westmead Institute, Sydney, New South Wales, Australia. [445]Discipline of Pathology, Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. [446]School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia. [447]Melanoma Institute Australia, The University of Sydney, Wollstonecraft, New South Wales, Australia. [448]Royal Prince Alfred Hospital, Sydney, New South Wales, Australia. [449]Diagnostic Development, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [450]Ontario

Tumour Bank, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [451]PanCuRx Translational Research Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [452]BioSpecimen Sciences Program, University Health Network, Toronto, Ontario, Canada. [453]Hepatobiliary/Pancreatic Surgical Oncology Program, University Health Network, Toronto, Ontario, Canada. [454]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [455]Division of Medical Oncology, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [456]University of Nebraska Medical Center, Omaha, NE, USA. [457]BioSpecimen Sciences Program, University Health Network, Toronto, Ontario, Canada. [458]Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [459]University Health Network, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [460]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [461]BioSpecimen Sciences, Laboratory Medicine (Toronto), Medical Biophysics, PanCuRX, Toronto, Ontario, Canada. [462]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [463]Department of Pathology, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [464]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [465]Department of Biochemistry and Molecular Medicine, University California at Davis, Sacramento, CA, USA. [466]Human Longevity, San Diego, CA, USA. [467]Department of Surgical Oncology, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [468]Genome Informatics Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [469]STTARR Innovation Facility, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [470]Department of Pathology, Toronto General Hospital, Toronto, Ontario, Canada. [471]CRUK Manchester Institute and Centre, Manchester, UK. [472]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [473]Manchester Cancer Research Centre, Cancer Division, FBMH, University of Manchester, Manchester, UK. [474]Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [475]Hefei University of Technology, Anhui, China. [476]State Key Laboratory of Cancer Biology and Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Shaanxi, China. [477]Fourth Military Medical University, Shaanxi, China. [478]Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing, China. [479]Department of Surgery, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China. [480]Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK. [481]Canadian Center for Computational Genomics, McGill University, Montreal, Quebec, Canada. [482]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [483]International Agency for Research on Cancer, Lyon, France. [484]McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada. [485]St James Institute of Oncology, University of Leeds, St James's University Hospital, Leeds, UK. [486]Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia. [487]Centre National de Génotypage, CEA - Institute de Génomique, Evry, France. [488]Department of Oncology, Gil Medical Center, Gachon University, Incheon, South Korea. [489]Department of Molecular Oncology, BC Cancer Agency, Vancouver, British Columbia, Canada. [490]Los Alamos National Laboratory, Los Alamos, NM, USA. [491]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway. [492]Lund University, Lund, Sweden. [493]Translational Research Lab, Centre Léon Bérard, Lyon, France. [494]Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. [495]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [496]Department of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [497]Li Ka Shing Centre, Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [498]Department of Oncology, University of Cambridge, Cambridge, UK. [499]Breast Cancer Translational Research Laboratory J. C. Heuson, Institut Jules Bordet, Brussels, Belgium. [500]Laboratory for Translational Breast Cancer Research, Department of Oncology, KU Leuven, Leuven, Belgium. [501]Translational Cancer Research Unit, GZA Hospitals St-Augustinus, Center for Oncological Research, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. [502]Department of Gynecology & Obstetrics and Department of Clinical Sciences, Skåne University Hospital, Lund University, Lund, Sweden. [503]Icelandic Cancer Registry, Icelandic Cancer Society, Reykjavik, Iceland. [504]Department of Medical Oncology, Josephine Nefkens Institute and Cancer Genomics Centre, Erasmus Medical Center, Rotterdam, The Netherlands. [505]National Genotyping Center, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [506]Department of Pathology, Oslo University Hospital Ulleval, Oslo, Norway. [507]Faculty of Medicine and Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [508]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [509]Department of Pathology, Skåne University Hospital, Lund University, Lund, Sweden. [510]Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands. [511]Department of Pathology, College of Medicine, Hanyang University, Seoul, South Korea. [512]Department of Pathology, Asan Medical Center, College of Medicine, Ulsan University, Songpa-gu, Seoul, South Korea. [513]The Netherlands Cancer Institute, Amsterdam, The Netherlands. [514]Department of Surgery, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Boston, MA, USA. [515]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [516]Department of Clinical Science, University of Bergen, Bergen, Norway. [517]Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [518]The University of Queensland Centre for Clinical Research, The Royal Brisbane & Women's Hospital, Herston, Queensland, Australia. [519]Department of Pathology, Institut Jules Bordet, Brussels, Belgium. [520]Institute for Bioengineering and Biopharmaceutical Research (IBBR), Hanyang University, Seoul, South Korea. [521]University of Oslo, Oslo, Norway. [522]Institut Bergonié, Bordeaux, France. [523]Department of Research Oncology, Guy's Hospital, King's Health Partners AHSC, King's College London School of Medicine, London, UK. [524]University Hospital of Minjoz, INSERM UMR 1098, Besançon, France. [525]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, UK. [526]East of Scotland Breast Service, Ninewells Hospital, Aberdeen, UK. [527]Oncologie Sénologie, ICM Institut Régional du Cancer, Montpellier, France. [528]Department of Radiation Oncology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [529]University of Iceland, Reykjavik, Iceland. [530]Dundee Cancer Centre, Ninewells Hospital, Dundee, UK. [531]Institut Curie, INSERM Unit 830, Paris, France. [532]Department of Laboratory Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [533]Department of General Surgery, Singapore General Hospital, Singapore, Singapore. [534]INCa-Synergie, Centre Léon Bérard, Universite Lyon, Lyon, France. [535]Giovanni Paolo II/I.R.C.C.S. Cancer Institute, Bari, Italy. [536]Department of Biopathology, Centre Léon Bérard, Lyon, France. [537]Université Claude Bernard Lyon 1, Villeurbanne, France. [538]NCCS-VARI Translational Research Laboratory, National Cancer Centre Singapore, Singapore, Singapore. [539]Department of Pathology, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands. [540]Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [541]Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany. [542]Institute of Human Genetics, University of Ulm, Ulm, Germany. [543]University Hospital of Ulm, Ulm, Germany. [544]Hematopathology Section, Institute of Pathology, Christian-Albrechts-University, Kiel, Germany. [545]Department of Human Genetics, Hannover Medical School, Hannover, Germany. [546]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Düsseldorf, Germany. [547]Department of Internal Medicine/Hematology, Friedrich-Ebert-Hospital, Neumünster, Germany. [548]Pediatric Hematology and Oncology, University Hospital Muenster, Muenster, Germany. [549]Department of Pediatrics, University Hospital Schleswig-Holstein, Kiel, Germany. [550]Department of Medicine II, University of Würzburg, Würzburg, Germany. [551]Senckenberg Institute of Pathology, University of Frankfurt Medical School, Frankfurt, Germany. [552]Institute of Pathology, Charité–University Medicine Berlin, Berlin, Germany. [553]Department for Internal Medicine II, University Hospital Schleswig-Holstein, Kiel, Germany. [554]Institute for Medical Informatics Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [555]Department of Hematology and Oncology, Georg-Augusts-University of Göttingen, Göttingen, Germany. [556]Institute of Cell Biology (Cancer Research), University of Duisburg-Essen, Essen, Germany. [557]MVZ Department of Oncology, PraxisClinic am Johannisplatz, Leipzig, Germany. [558]Institute of Pathology, Ulm University and University Hospital of Ulm, Ulm, Germany. [559]Department of Pathology, Robert-Bosch-Hospital, Stuttgart, Germany. [560]Pediatric Hematology and Oncology, University Hospital Giessen, Giessen, Germany. [561]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [562]Institute of Pathology, University of Wuerzburg, Wuerzburg, Germany. [563]Department of General Internal Medicine, University Kiel, Kiel, Germany. [564]Clinic for Hematology and Oncology, St-Antonius-Hospital, Eschweiler, Germany. [565]Department for Internal Medicine III, University of Ulm and University Hospital of Ulm, Ulm, Germany. [566]Neuroblastoma Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [567]Department of Pediatric Oncology and Hematology, University of Cologne, Cologne, Germany. [568]University of Düsseldorf, Düsseldorf, Germany. [569]Department of Vertebrate Genomics/Otto Warburg Laboratory Gene Regulation and Systems Biology of Cancer, Max Planck Institute for Molecular Genetics, Berlin, Germany. [570]St Jude Children's Research Hospital, Memphis, TN, USA. [571]Heidelberg University Hospital, Heidelberg, Germany. [572]Genomics and Proteomics Core Facility High Throughput Sequencing Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany. [573]Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany. [574]University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [575]Martini-Clinic, Prostate Cancer Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [576]Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [577]Division of Cancer Genome Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [578]National Institute of Biomedical Genomics, Kalyani, India. [579]Advanced Centre for Treatment Research & Education in Cancer, Tata Memorial Centre, Navi Mumbai, India. [580]Department of Pathology, General Hospital of Treviso, Department of Medicine, University of Padua, Treviso, Italy. [581]Department of Medicine (DIMED), Surgical Pathology Unit, University of Padua, Padua, Italy. [582]Department of Hepatobiliary and Pancreatic Oncology, Hepatobiliary and Pancreatic Surgery Division, Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. [583]Department of Pathology, Keio University School of Medicine, Tokyo, Japan. [584]Department of Hepatobiliary and Pancreatic Oncology, National Cancer Center Hospital, Tokyo, Japan. [585]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [586]Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [587]Gastric Surgery Division, Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Tokyo, Japan. [588]Department of Gastroenterology and Hepatology, Yokohama City University Graduate School of Medicine, Kanagawa, Japan. [589]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, University of Tokyo, Tokyo, Japan. [590]Department of Cancer Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan. [591]Hiroshima University, Hiroshima, Japan. [592]Tokyo Women's Medical University, Tokyo, Japan. [593]Osaka International Cancer Center, Osaka, Japan. [594]Wakayama Medical University, Wakayama, Japan. [595]Hokkaido University, Sapporo, Japan. [596]Division of Medical Oncology, National Cancer Centre, Singapore, Singapore. [597]Cholangiocarcinoma Screening and Care Program and Liver Fluke and Cholangiocarcinoma Research Centre, Faculty of Medicine, Khon Kaen University,

Khon Kaen, Thailand. [598]Lymphoma Genomic Translational Research Laboratory, National Cancer Centre, Singapore, Singapore. [599]Center of Digestive Diseases and Liver Transplantation, Fundeni Clinical Institute, Bucharest, Romania. [600]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, School of Medicine, Keimyung University Dongsan Medical Center, Daegu, South Korea. [601]Pathology, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. [602]Hematology, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. [603]Department of Biochemistry and Molecular Biology, Faculty of Medicine, University Institute of Oncology-IUOPA, Oviedo, Spain. [604]Anatomia Patológica, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. [605]Spanish Ministry of Science and Innovation, Madrid, Spain. [606]Royal National Orthopaedic Hospital (Bolsover), London, UK. [607]Department of Pathology, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway. [608]Institute of Clinical Medicine and Institute of Oral Biology, University of Oslo, Oslo, Norway. [609]Research Department of Pathology, University College London Cancer Institute, London, UK. [610]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [611]Royal National Orthopaedic Hospital (Stanmore), London, UK. [612]Division of Orthopaedic Surgery, Oslo University Hospital, Oslo, Norway. [613]Department of Pathology (Research), University College London Cancer Institute, London, UK. [614]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [615]University of Pavia, Pavia, Italy. [616]Karolinska Institute, Stockholm, Sweden. [617]Wellcome Sanger Institute, Hinxton, UK. [618]University of Oxford, Oxford, UK. [619]Salford Royal NHS Foundation Trust, Salford, UK. [620]Gloucester Royal Hospital, Gloucester, UK. [621]Royal Stoke University Hospital, Stoke-on-Trent, UK. [622]St Thomas's Hospital, London, UK. [623]Imperial College NHS Trust, Imperial College London, London, UK. [624]Department of Histopathology, Salford Royal NHS Foundation Trust, Salford, UK. [625]Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. [626]Edinburgh Royal Infirmary, Edinburgh, UK. [627]Barking Havering and Redbridge University Hospitals NHS Trust, Romford, UK. [628]King's College London and Guy's and St Thomas' NHS Foundation Trust, London, UK. [629]Cambridge Oesophagogastric Centre, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [630]Nottingham University Hospitals NHS Trust, Nottingham, UK. [631]St Luke's Cancer Centre, Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. [632]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [633]Norfolk and Norwich University Hospital NHS Trust, Norwich, UK. [634]University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK. [635]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [636]Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK. [637]School of Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. [638]Wythenshawe Hospital, Manchester, UK. [639]Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [640]Royal Marsden NHS Foundation Trust, London and Sutton, London, UK. [641]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [642]HCA Laboratories, London, UK. [643]University of Liverpool, Liverpool, UK. [644]Academic Urology Group, Department of Surgery, University of Cambridge, Cambridge, UK. [645]University of Oxford, Oxford, Oxford, UK. [646]Department of Urology, James Buchanan Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [647]Second Military Medical University, Shanghai, China. [648]Department of Surgery and Cancer, Imperial College London, London, UK. [649]The Chinese University of Hong Kong, Shatin, Hong Kong, China. [650]Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, Headington, Oxford, UK. [651]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [652]Department of Bioinformatics and Computational Biology and Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [653]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. [654]Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, British Columbia, Canada. [655]Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [656]University Health Network, Toronto, Ontario, Canada. [657]Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [658]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA. [659]Research Health Analytics and Informatics, University Hospitals Cleveland Medical Center, Cleveland, OH, USA. [660]Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada. [661]Department of Surgery and Department of Oncology, University of Calgary, Calgary, Alberta, Canada. [662]Buck Institute for Research on Aging, Novato, CA, USA. [663]Duke University Medical Center, Durham, NC, USA. [664]USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA. [665]The Preston Robert Tisch Brain Tumor Center, Duke University Medical Center, Durham, NC, USA. [666]Department of Dermatology and Department of Pathology, Yale University, New Haven, CT, USA. [667]Fox Chase Cancer Center, Philadelphia, PA, USA. [668]Department of Surgery, Division of Thoracic Surgery, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [669]University of Michigan Comprehensive Cancer Center, Ann Arbor, MI, USA. [670]University of Alabama at Birmingham, Birmingham, AL, USA. [671]Division of Anatomic Pathology, Mayo Clinic, Rochester, MN, USA. [672]Division of Experimental Pathology, Mayo Clinic, Rochester, MN, USA. [673]Department of Oncology, The Johns Hopkins School of Medicine, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, MD, USA. [674]International Genomics Consortium, Phoenix, AZ, USA. [675]Department of Pediatrics and Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [676]Department of Pathology, UPMC Shadyside, Pittsburgh, PA, USA. [677]Center for Cancer Genomics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [678]Department of Neuro-Oncology, Istituto Neurologico Besta, Milan, Italy. [679]University of Queensland Thoracic Research Centre, The Prince Charles Hospital, Brisbane, Queensland, Australia. [680]Department of Neurosurgery, University of Florida, Gainesville, FL, USA. [681]Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [682]Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [683]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [684]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA. [685]Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA. [686]Department of Internal Medicine, Division of Medical Oncology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [687]University of Tennessee Health Science Center for Cancer Research, Memphis, TN, USA. [688]Centre for Translational and Applied Genomics, British Columbia Cancer Agency, Vancouver, British Columbia, Canada. [689]Department of Pathology & Immunology, Baylor College of Medicine, Houston, TX, USA. [690]Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA. [691]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [692]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [693]Indivumed, Hamburg, Germany. [694]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, School of Medicine, Keimyung University Dong-san Medical Center, Daegu, South Korea. [695]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [696]Department of Surgery, School of Medicine and Health Science, The George Washington University, Washington, DC, USA. [697]Endocrine Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [698]National Cancer Center, Gyeonggi, South Korea. [699]ILSbio, LLC Biobank, Chestertown, MD, USA. [700]Gynecologic Oncology, NYU Laura and Isaac Perlmutter Cancer Center, New York University, New York, NY, USA. [701]Division of Oncology, Stem Cell Biology Section, Washington University School of Medicine, St Louis, MO, USA. [702]Urologic Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [703]Institute for Systems Biology, Seattle, WA, USA. [704]Center for Personalized Medicine, Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, USA. [705]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [706]Department of Surgery, Duke University, Durham, NC, USA. [707]Department of Obstetrics, Gynecology and Reproductive Services, University of California San Francisco, San Francisco, CA, USA. [708]Department of Neurology and Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA. [709]Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. [710]Department of Pathology, Roswell Park Cancer Institute, Buffalo, NY, USA. [711]Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, Washington University School of Medicine, St Louis, MO, USA. [712]Department of Palliative, Rehabilitation and Integrative Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [713]Penrose St Francis Health Services, Colorado Springs, CO, USA. [714]The University of Chicago, Chicago, IL, USA. [715]Department of Neurology, Mayo Clinic, Rochester, MN, USA. [716]Center for Liver Cancer, Research Institute and Hospital, National Cancer Center, Gyeonggi, South Korea. [717]Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [718]NYU Langone Medical Center, New York, NY, USA. [719]Department of Hematology and Medical Oncology, Cleveland Clinic, Cleveland, OH, USA. [720]Department of Genetics, Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [721]Helen F. Graham Cancer Center at Christiana Care Health Systems, Newark, DE, USA. [722]Cureline, South San Francisco, CA, USA. [723]Department of Obstetrics and Gynecology, Medical College of Wisconsin, Milwaukee, WI, USA. [724]Hematology and Medical Oncology, Winship Cancer Institute of Emory University, Atlanta, GA, USA. [725]Vanderbilt Ingram Cancer Center, Vanderbilt University, Nashville, TN, USA. [726]Ohio State University College of Medicine and Arthur G. James Comprehensive Cancer Center, Columbus, OH, USA. [727]Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [728]Analytical Biological Services, Wilmington, DE, USA. [729]Department of Dermatology, University Hospital Essen, Westdeutsches Tumorzentrum and German Cancer Consortium, Essen, Germany. [730]University of Pittsburgh, Pittsburgh, PA, USA. [731]Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, MD, USA. [732]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [733]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [734]Department of Gynecologic Oncology and Reproductive Medicine, and Center for RNA Interference and Non-Coding RNA, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [735]Department of Urology, Mayo Clinic, Rochester, MN, USA. [736]Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [737]Department of Neurosurgery, Department of Hematology and Department of Medical Oncology, Winship Cancer Institute and School of Medicine, Emory University, Atlanta, GA, USA. [738]Georgia Regents University Cancer Center, Augusta, GA, USA. [739]Thoracic Oncology Laboratory, Mayo Clinic, Rochester, MN, USA. [740]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [741]Department of Obstetrics & Gynecology, Division of Gynecologic Oncology, Mayo Clinic, Rochester, MN, USA. [742]International Institute for Molecular Oncology, Poznań, Poland. [743]Poznan University of Medical Sciences, Poznań, Poland. [744]Edison Family Center for Genome Sciences and Systems Biology, Washington University, St Louis, MO, USA. [745]These authors jointly supervised this work: Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Lincoln D. Stein. *e-mail: pc8@sanger.ac.uk; gadgetz@broadinstitute.org; korbel@embl.de; jstuart@ucsc.edu; lincoln.stein@gmail.com

Methods

Samples

We compiled an inventory of matched tumour–normal whole-cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naive, primary cancers, although a small number of donors had multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (1) matched tumour and normal specimen pair; (2) a minimal set of clinical fields; and (3) characterization of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads.

We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014 (Extended Data Table 1). After quality assurance (Supplementary Methods 2.5), data from 176 donors were excluded as unusable, 75 had minor issues that could affect some analyses (grey-listed donors) and 2,583 had data of optimal quality (white-listed donors) (Supplementary Table 1). Across the 2,658 white- and grey-listed donors, whole-genome sequences were available from 2,605 primary tumours and 173 metastases or local recurrences. Matching normal samples were obtained from blood (2,064 donors), tissue adjacent to the primary tumour (87 donors) or from distant sites (507 donors). Whole-genome sequencing data were available for tumour and normal DNA for the entire cohort. The mean read coverage was 39× for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38× and 60× (Supplementary Fig. 1). The majority of specimens (65.3%) were sequenced using 101-bp paired-end reads. An additional 28% were sequenced with 100-bp paired-end reads. Of the remaining specimens, 4.7% were sequenced with read lengths longer than 101 bp, and 1.9% with read lengths shorter than 100 bp. The distribution of read lengths by tumour cohort is shown in Supplementary Fig. 11. Median read length for whole-genome sequencing paired-end reads was 101 bp (mean = 106.2, s.d. = 16.7; minimum–maximum = 50–151). RNA-sequencing data were collected and re-analysed centrally for 1,222 donors, including 1,178 primary tumours, 67 metastases or local recurrences and 153 matched normal tissue samples adjacent to the primary tumour.

Demographically, the cohort included 1,469 men (55%) and 1,189 women (45%), with a mean age of 56 years (range, 1–90 years) (Supplementary Table 1). Using population ancestry-differentiated single nucleotide polymorphisms, the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects (Supplementary Table 1).

We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary[89]. Overall, the PCAWG dataset comprises 38 distinct tumour types (Extended Data Table 1 and Supplementary Table 1). Although the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely owing to differences among contributing ICGC/TCGA groups in the numbers of sequenced samples.

Uniform processing and somatic variant calling

To generate a consistent set of somatic mutation calls that could be used for cross-tumour analyses, we analysed all 6,835 samples using a uniform set of algorithms for alignment, variant calling and quality control (Extended Data Fig. 1, Supplementary Fig. 2, Supplementary Table 3 and Supplementary Methods 2). We used the BWA-MEM algorithm[90] to align each tumour and normal sample to human reference build hs37d5 (as used in the 1000 Genomes Project[91]). Somatic mutations were identified in the aligned data using three established pipelines, which were run independently on each tumour–normal pair. Each of the three pipelines—labelled 'Sanger'[92–95], 'EMBL/DKFZ'[96,97] and 'Broad'[98–101] after the computational biology groups that created or assembled them—consisted of multiple software packages for calling somatic SNVs, small indels, CNAs and somatic SVs (with intrachromosomal SVs defined as those >100 bp). Two additional variant algorithms[102,103] were included to further improve accuracy across a broad range of clonal and subclonal mutations. We tested different merging strategies using validation data, and choses the optimal method for each variant type to generate a final consensus set of mutation calls (Supplementary Methods S2.4).

Somatic retrotransposition events, including Alu and LINE-1 insertions[72], L1-mediated transductions[73] and pseudogene formation[104], were called using a dedicated pipeline[73]. We removed these retrotransposition events from the somatic SV call-set. Mitochondrial DNA mutations were called using a published algorithm[105]. RNA-sequencing data were uniformly processed to quantify normalized gene-level expression, splicing variation and allele-specific expression, and to identify fusion transcripts, alternative promoter usage and sites of RNA editing[8].

Integration, phasing and validation of germline variant call-sets

Calls of common (≥1% frequency in PCAWG) and rare (<1%) germline variants including single-nucleotide polymorphisms, indels, SVs and mobile-element insertions (MEIs) were generated using a population-scale genetic polymorphism-detection approach[91,106]. The uniform germline data-processing workflow comprised variant identification using six different variant-calling algorithms[96,107,108] and was orchestrated using the Butler workflow system[109].

We performed call-set benchmarking, merging, variant genotyping and statistical haplotype-block phasing[91] (Supplementary Methods 3.4). Using this strategy, we identified 80.1 million germline single-nucleotide polymorphisms, 5.9 million germline indels, 1.8 million multi-allelic short (<50 bp) germline variants, as well as germline SVs ≥ 50 bp in size including 29,492 biallelic deletions and 27,254 MEIs (Supplementary Table 2). We statistically phased this germline variant set using haplotypes from the 1000 Genomes Project[91] as a reference panel, yielding an N50-phased block length of 265 kb based on haploid chromosomes from donor-matched tumour genomes. Precision estimates for germline SNVs and indels were >99% for the phased merged call-set, and sensitivity estimates ranged from 92% to 98%.

Core alignment and variant calling by cloud computing

The requirement to uniformly realign and call variants on nearly 5,800 whole genomes (tumour plus normal) presented considerable computational challenges, and raised ethical issues owing to the use of data from different jurisdictions (Extended Data Table 2). To process the data, we adopted a cloud-computing architecture[26] in which the alignment and variant calling was spread across 13 data centres on 3 continents, representing a mixture of commercial, infrastructure-as-a-service, academic cloud compute and traditional academic high-performance computer clusters (Supplementary Table 3). Together, the effort used 10 million CPU-core hours.

To generate reproducible variant calling across the 13 data centres, we built the core pipelines into Docker containers[28], in which the workflow description, required code and all associated dependencies were packaged together in stand-alone packages. These heavily tested, extensively validated workflows are available for download (Box 1).

Validation, benchmarking and merging of somatic variant calls

To evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep-sequencing validation experiment (Supplementary Notes 1). We selected a pilot set of 63 representative tumour–normal pairs, on which we ran the 3 core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the PCAWG SNV Calling Methods Working Group. Sufficient DNA remained for 50 of the 63 cases for validation, which was performed by hybridization of tumour and matched normal DNA to a custom RNA bait set, followed

by deep sequencing, as previously described[29]. Although performed using the same sequencing chemistry as the original whole-genome sequencing analyses, the considerably greater depth achieved in the validation experiment enabled accurate assessment of sensitivity and precision of variant calls. Variant calls in repeat-masked regions were not tested, owing to the challenge of designing reliable validation probes in these areas.

The 3 core pipelines had individual estimates of sensitivity of 80–90% to detect a true somatic SNV called by any of the 13 pipelines; with >95% of SNV calls made by each of the core pipelines being genuine somatic variants (Fig. 1a). For indels—a more-challenging class of variants to identify in short-read sequencing data—the 3 core algorithms had individual sensitivity estimates in the range of 40–50%, with precision 70–95% (Fig. 1b). Validation of SV calls is inherently more difficult, as methods based on PCR or hybridization to RNA baits often fail to isolate DNA that spans the breakpoint. To assess the accuracy of SV calls, we therefore used the property that an SV must either generate a copy-number change or be balanced, whereas artefactual calls will not respect this property. For individual SV-calling algorithms, we estimated precision to be in the range of 80–95% for samples in the 63-sample pilot dataset.

Next, we examined multiple methods for merging calls made by several algorithms into a single definitive call-set to be used for downstream analysis. The final consensus calls for SNVs were based on a simple approach that required two or more methods to agree on a call. For indels, because methods were less concordant, we used stacked logistic regression[110,111] to integrate the calls. The merged SV set includes all calls made by two or more of the four primary SV-calling algorithms[96,100,112,113]. Consensus CNA calls were obtained by joining the outputs of six individual CNA-calling algorithms with SV consensus breakpoints to obtain base-pair resolution CNAs (Supplementary Methods 2.4.3). Consensus purity and ploidy were derived, and a multitier system was developed for consensus copy-number calls (Supplementary Methods 2.4.3, and described in detail elsewhere[7]).

Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (90% confidence interval, 88–98%) and 95% (90% confidence interval, 71–99%), respectively, for SNVs (Extended Data Fig. 2). For somatic indels, sensitivity and precision were 60% (90% confidence interval, 34–72%) and 91% (90% confidence interval, 73–96%), respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one calling pipeline; precision was estimated to be 97.5%. That is, 97.5% of SVs in the merged SV call-set had an associated copy-number change or balanced partner rearrangement. The improvement in calling accuracy from combining different pipelines was most noticeable in variants that had low variant allele fractions, which are likely to originate from subclonal populations of the tumour (Fig. 1c, d). There remains much work to be done to improve indel calling software; we still lack sensitivity for calling even fully clonal complex indels from short-read sequencing data.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The PCAWG-generated alignments, somatic variant calls, annotations and derived datasets are available for general research use for browsing and download at http://dcc.icgc.org/pcawg/ (Box 1 and Supplementary Table 4). In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identifying information, such as germline alleles and underlying read data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Beyond the core sequence data and variant call-sets, the analyses in this paper used a number of datasets that were derived from the variant calls (Supplementary Table 4). The individual datasets are available at Synapse (https://www.synapse.org/), and are denoted with synXXXXX accession numbers; all these datasets are also mirrored at https://dcc.icgc.org, with full links, filenames, accession numbers and descriptions detailed in Supplementary Table 4. The datasets encompass: clinical data from each patient including demographics, tumour stage and vital status (syn10389158); harmonized tumour histopathology annotations using a standardised hierarchical ontology (syn1038916); inferred purity and ploidy values for each tumour sample (syn8272483); driver mutations for each patient from their cancer genome spanning all classes of variant, and coding versus non-coding drivers (syn11639581); mutational signatures inferred from PCAWG donors (syn11804065), including APOBEC mutagenesis (syn7437313); and transcriptional data from RNA sequencing, including gene expression levels (syn5553985, syn5553991, syn8105922) and gene fusions (syn10003873, syn7221157).

Code availability

Computational pipelines for calling somatic mutations are available to the public at https://dockstore.org/organizations/PCAWG/collections/PCAWG. A range of data-visualization and -exploration tools are also available for the PCAWG data (Box 1).

89. NCI SEER. *ICD-O-3 Coding Materials* (2018).
90. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
91. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
92. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1–15.9.17 (2016).
93. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
94. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
95. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
96. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
97. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
98. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
99. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
100. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
101. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
102. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).
103. 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).
104. Cooke, S. L. et al. Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5**, 3644 (2014).
105. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, e02935 (2014).
106. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
107. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
108. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

109. Yakneen, S., Waszak, S. M., Gertz, M. & Korbel, J. O. & PCAWG Consortium. Butler enables rapid cloud-based analysis of thousands of human genomes. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-019-0360-3 (2020).
110. Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics* **15**, 154 (2014).
111. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
112. Campbell, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
113. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

Acknowledgements We thank research participants who donated samples and data, the physicians and clinical staff who contributed to sample annotation and collection, and the numerous funding agencies that contributed to the collection and analysis of this dataset.

Author contributions Writing committee leads: Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein. Head of project management: Jennifer L. Jennings. Sample collection: major contributions from Marc D. Perry, Hardeep K. Nahal-Bose; led by B. F. Francis Ouellette. Histopathology harmonization: major contribution from Constance H. Li; further contributions from Esther Rheinbay, G. Petur Nielsen, Dennis C. Sgroi, Chin-Lee Wu, William C. Faquin, Vikram Deshpande, Paul C. Boutros, Alexander J. Lazar, Katherine A. Hoadley; led by Lincoln D. Stein, David N. Louis. Uniform processing, somatic, germline variant calling: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Junjun Zhang, Wenyi Wang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Core alignment, variant calling by cloud computing: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Adam P. Butler, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez. Integration, phasing, validation of germline variant callsets: major contributions from Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu; further contributions from Sergei Yakneen, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, Alicia L. Bruzos, Javier Temes, Jorge Zamora, L. Jonathan Dursi, Adrian Baez-Ortega, Hyung-Lae Kim, Matthew H. Bailey, R. Jay Mashl, Kai Ye, Ivo Buchhalter, Anthony DiBiase, Kuan-lin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Matthias Schlesner, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Jared T. Simpson, Li Ding, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton; led by Stephan Ossowski, Jose M. C. Tubio, Gad Getz, Francisco M. De La Vega, Xavier Estivill, Jan O. Korbel. Validation, benchmarking, merging of somatic variant calls: major contribution from L. Jonathan Dursi; further contributions from David A. Wheeler, Christina K. Yung; led by Li Ding, Jared T. Simpson. Data and code availability: major contribution from Junjun Zhang; further contributions from Christina K. Yung, Sergei Yakneen, Denis Yuen, George L. Mihaiescu, Larsson Omberg; led by Vincent Ferretti. Pan-cancer burden of somatic mutations: major contribution from Junjun Zhang; led by Peter J. Campbell. Panorama of driver mutations in human cancer: led by Radhakrishnan Sabarinathan, Oriol Pich, Abel Gonzalez-Perez. PCAWG tumours with no apparent driver mutations: major contribution from Esther Rheinbay; further contributions from Amaro Taylor-Weiner, Radhakrishnan Sabarinathan; led by Peter J. Campbell, Gad Getz. Patterns, oncogenicity of kataegis, chromoplexy: major contributions from Matthew W. Fittall, Jonas Demeulemeester, Maxime Tarabichi; further contributions from Nicola D. Roberts, Peter J. Campbell, Jan O. Korbel; led by Peter Van Loo. Patterns, oncogenicity of chromothripsis: major contributions from Maxime Tarabichi, Jonas Demeulemeester, Matthew W. Fittall; further contributions from Isidro Cortes-Ciriano, Lara Urban, Peter J. Park, Peter J. Campbell, Jan O. Korbel; led by Peter Van Loo. Timing-clustered mutational processes during tumour evolution: major contributions from Jonas Demeulemeester, Maxime Tarabichi, Matthew W. Fittall; further contributions from Jan O. Korbel, Peter J. Campbell; led by Peter Van Loo. Germline effects on somatic mutation: major contributions from Sebastian M. Waszak, Bin Zhu, Bernardo Rodriguez-Martin, Esa Pitkanen, Tobias Rausch; further contributions from Yilong Li, Natalie Saini, Leszek J. Klimczak, Joachim Weischenfeldt, Nikos Sidiropoulos, Ludmil B. Alexandrov, Francesc Muyas, Raquel Rabionet, Georgia Escaramis, Adrian Baez-Ortega, Mattia Bosio, Aliaksei Z. Holik, Hana Susak, Eva G. Alvarez, Alicia L. Bruzos, Javier Temes, Aparna Prasad, Nina Habermann, Serap Erkek, Lara Urban, Claudia Calabrese, Benjamin Raeder, Eoghan Harrington, Simon Mayes, Daniel Turner, Sissel Juul, Steven A. Roberts, Lei Song, Roelof Koster, Lisa Mirabello, Xing Hua, Tomas J. Tanskanen, Marta Tojo, David C. Wedge, Jorge Zamora, Jieming Chen, Lauri A. Aaltonen, Gunnar Ratsch, Roland F. Schwarz, Atul J. Butte, Alvis Brazma, Peter J. Campbell, Stephen J. Chanock, Nilanjan Chatterjee, Oliver Stegle, Olivier Harismendy; led by G. Steven Bova, Dmitry A. Gordenin, Jose M. C. Tubio, Douglas F. Easton, Xavier Estivill, Jan O. Korbel. Replicative immortality: major contribution from David Haan; further contributions from Lina Sieverling, Lars Feuerbach; led by Lincoln D. Stein, Joshua M. Stuart. Ethical considerations of genomic cloud computing: led by Don Chalmers, Yann Joly, Bartha Knoppers, Fruzsina Molnar-Gabor, Jan O. Korbel, Mark Phillips, Adrian Thorogood, David Townend. Online resources for data access, visualization, exploration and analysis: major contributions from Mary Goldman, Junjun Zhang, Nuno A. Fonseca; further contributions from Qian Xiang, Brian Craft, Elena Pineiro-Yanez, Alfonso Munoz, Robert Petryszak, Anja Fullgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu; led by Brian D. O'Connor, Lincoln D. Stein, Alvis Brazma, Vincent Ferretti, Miguel Vazquez. The 63-sample pilot-analysis validation process: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenk Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Processing of validation data: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez, Joachim Weischenfeldt, Denis Yuen, Adam P. Butler, Brandi N. Davis-Dusenbery, Roland Eils, Vincent Ferretti, Robert L. Grossman, Olivier Harismendy, Youngwook Kim, Hidewaki Nakagawa, Steven J. Newhouse, David Torrents; led by Lincoln D. Stein. Whole-genome sequencing somatic variant calling: major contribution from Junjun Zhang; further contributions from Christina K. Yung, Solomon I. Shorser. Whole-genome alignment: Keiran M. Raine, Junjun Zhang, Brian D. O'Connor. DKFZ pipeline: Kortine Kleinheinz, Tobias Rausch, Jan O. Korbel, Ivo Buchhalter, Michael C. Heinold, Barbara Hutter, Natalie Jager, Nagarajan Paramasivam, Matthias Schlesner. EMBL pipeline: Joachim Weischenfeldt. Sanger pipeline: Keiran M. Raine, Jonathan Hinton, David R. Jones, Andrew Menzies, Lucy Stebbings, Adam P. Butler. Broad pipeline: Gordon Saksena, Dimitri Livitz, Esther Rheinbay, Julian M. Hess, Ignaty Leshchiner, Chip Stewart, Grace Tiao, Jeremiah A. Wala, Amaro Taylor-Weiner, Mara Rosenberg, Andrew J. Dunford, Manaswi Gupta, Marcin Imielinski, Matthew Meyerson, Rameen Beroukhim, Gad Getz. MuSE Pipeline: Yu Fan, Wenyi Wang. Consensus somatic SNV/indel annotation: Andrew Menzies, Matthias Schlesner, Juri Reimand, Priyanka Dhingra, Ekta Khurana. Somatic SNV, indel merging: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep L. Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenk Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; major contributions from Li Ding, Jared T. Simpson. Somatic SV merging: Joachim Weischenfeldt, Francesco Favero, Yilong Li. Somatic CNA merging: Stefan Dentro, Jeff Wintersinger, Ignaty Leshchiner. Oxidative artefact filtration: Dimitri Livitz, Ignaty Leshchiner, Chip Stewart, Esther Rheinbay, Gordon Saksena, Gad Getz. Strand bias filtration: Matthias Bieg, Ivo Buchhalter, Johannes Werner, Matthias Schlesner. miniBAM generation: Jeremiah A. Wala, Gordon Saksena, Rameen Beroukhim, Gad Getz. Germline variant identification from whole-genome sequencing: major contributions from Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu; further contributions from Sergei Yakneen, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, Alicia L. Bruzos, Jorge Zamora, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, L. Jonathan Dursi, Adrian Baez-Ortega, Hyung-Lae Kim, Matthew H. Bailey, R. Jay Mashl, Kai Ye, Ivo Buchhalter, Vasilisa Rudneva, Ji Wan Park, Eun Pyo Hong, Seong Gu Heo, Anthony DiBiase, Kuan-lin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Matthias Schlesner, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Jared T. Simpson, Li Ding, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton; led by Stephan Ossowski, Jose M. C. Tubio, Gad Getz, Francisco M. De La Vega, Xavier Estivill, Jan O. Korbel. RNA-sequencing analysis: major contributions from Nuno A. Fonseca, Andre Kahles, Kjong-Van Lehmann, Lara Urban, Cameron M. Soulette, Yuichi Shiraishi, Fenglin Liu, Yao He, Deniz Demircioglu, Natalie R. Davidson, Claudia Calabrese, Junjun Zhang, Marc D. Perry, Qian Xiang; further contributions from Liliana Greger, Siliang Li, Dongbing Liu, Stefan G. Stark, Fan Zhang, Samirkumar B. Amin, Peter Bailey, Aurelien Chateigner, Isidro Cortes-Ciriano, Brian Craft, Serap Erkek, Milana Frenkel-Morgenstern, Mary Goldman, Katherine A. Hoadley, Yong Hou, Matthew R. Huska, Ekta Khurana, Helena Kilpinen, Jan O. Korbel, Fabien C. Lamaze, Chang Li, Xiaobo Li, Xinyue Li, Xingmin Liu, Maximillian G. Marin, Julia Markowski, Tannistha Nandi, Morten Muhlig Nielsen, Akinyemi I. Ojesina, Qiang Pan-Hammarstrom, Peter J. Park, Chandra Sekhar Pedamallu, Jakob Skou Pedersen, Reiner Siebert, Hong Su, Patrick Tan, Bin Tean Teh, Jian Wang, Sebastian M. Waszak, Heng Xiong, Sergei Yakneen, Chen Ye, Christina Yung, Xiuqing Zhang, Liangtao Zheng, Jingchun Zhu, Shida Zhu, Philip Awadalla, Chad J. Creighton, Matthew Meyerson, B. F. Francis Ouellette, Kui Wu, Huanming Yang; led by Jonathan Goke, Roland F. Schwarz, Oliver Stegle, Zemin Zhang, Alvis Brazma, Gunnar Ratsch, Angela N. Brooks. Clustering of tumour genomes based on telomere maintenance-related features: major contribution from David Haan; led by Lincoln D. Stein, Joshua M. Stuart. Clustered mutational processes in PCAWG: major contributions from Jonas Demeulemeester, Maxime Tarabichi, Matthew W. Fittall; led by Peter J. Campbell, Jan O. Korbel, Peter Van Loo. Tumours without detected driver mutations: Esther Rheinbay, Amaro Taylor-Weiner, Radhakrishnan Sabarinathan, Peter J. Campbell, Gad Getz. Panorama of driver mutations in human cancer: major contributions from Radhakrishnan Sabarinathan, Oriol Pich; further contributions from Inigo Martincorena, Carlota Rubio-Perez, Malene Juul, Jeremiah A. Wala, Steven Schumacher, Ofer Shapira, Nikos Sidiropoulos, Sebastian M. Waszak, David Tamborero, Loris Mularoni, Esther Rheinbay, Henrik Hornshoj, Jordi Deu-Pons, Ferran Muinos, Johanna Bertl, Qianyun Guo, Chad J. Creighton, Joachim Weischenfeldt, Jan O. Korbel, Gad Getz, Peter J. Campbell, Jakob Skou Pedersen, Rameen Beroukhim; led by Abel Gonzalez-Perez. Pilot benchmarking, variant consensus development and validation: major contribution from L. Jonathan Dursi; further contributions from Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Yu Fan, David Torrents, Matthias Bieg, Paul C. Boutros, Ken Chen, Zechen Chong, Kristian Cibulskis, Oliver Drechsel, Roland Eils, Robert S. Fulton, Josep

Gelpi, Mark Gerstein, Santiago Gonzalez, Gad Getz, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Hyung-Lae Kim, Natalie Jager, Jongsun Jung, Sushant Kumar, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Ivica Letunic, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, R. Jay Mashl, Michael D. McLellan, Ana Milovanovic, Morten Muhlig Nielsen, Brian D. O'Connor, Stephan Ossowski, Nagarajan Paramasivam, Jakob Skou Pedersen, Marc D. Perry, Montserrat Puiggros, Romina Royo, Esther Rheinbay, S. Cenk Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Grace Tiao, Jeremiah A. Wala, Jiayin Wang, Wenyi Wang, Sebastian M. Waszak, Joachim Weischenfeldt, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Sergei Yakneen, Takafumi N. Yamaguchi, Kai Ye, Venkata Yellapantula, Junjun Zhang, David A. Wheeler; led by Li Ding, Jared T. Simpson. Production somatic variant calling on the PCAWG compute cloud: major contributions from Christina K. Yung, Brian D. O'Connor, Sergei Yakneen, Junjun Zhang; further contributions from Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, Keiran M. Raine, Romina Royo, Gordon Saksena, Matthias Schlesner, Solomon I. Shorser, Miguel Vazquez, Joachim Weischenfeldt, Denis Yuen, Adam P. Butler, Brandi N. Davis-Dusenbery, Roland Eils, Vincent Ferretti, Robert L. Grossman, Olivier Harismendy, Youngwook Kim, Hidewaki Nakagawa, Steven J Newhouse, David Torrents; led by Lincoln D. Stein. PCAWG data portals: major contributions from Mary Goldman, Junjun Zhang, Nuno A. Fonseca, Isidro Cortes-Ciriano; further contributions from Qian Xiang, Brian Craft, Elena Pineiro-Yanez, Brian D O'Connor, Wojciech Bazant, Elisabet Barrera, Alfonso Munoz, Robert Petryszak, Anja Fullgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu; led by Vincent Ferretti, Miguel Vazquez.

Competing interests Gad Getz receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig and POLYSOLVER. Hikmat Al-Ahmadie is consultant for AstraZeneca and Bristol-Myers Squibb. Samuel Aparicio is a founder and shareholder of Contextual Genomics. Pratiti Bandopadhayay receives grant funding from Novartis for an unrelated project. Rameen Beroukhim owns equity in Ampressa Therapeutics. Andrew Biankin receives grant funding from Celgene, AstraZeneca and is a consultant for or on advisory boards of AstraZeneca, Celgene, Elstar Therapeutics, Clovis Oncology and Roche. Ewan Birney is a consultant for Oxford Nanopore, Dovetail and GSK. Marcus Bosenberg is a consultant for Eli Lilly. Atul Butte is a cofounder of and consultant for Personalis, NuMedii, a consultant for Samsung, Geisinger Health, Mango Tree Corporation, Regenstrief Institute and in the recent past a consultant for 10x Genomics and Helix, a shareholder in Personalis, a minor shareholder in Apple, Twitter, Facebook, Google, Microsoft, Sarepta, 10x Genomics, Amazon, Biogen, CVS, Illumina, Snap and Sutro and has received honoraria and travel reimbursement for invited talks from Genentech, Roche, Pfizer, Optum, AbbVie and many academic institutions and health systems. Carlos Caldas has served on the Scientific Advisory Board of Illumina. Lorraine Chantrill acted on an advisory board for AMGEN Australia in the past 2 years. Andrew D. Cherniack receives research funding from Bayer. Helen Davies is an inventor on a number of patent applications that encompass the use of mutational signatures. Francisco De La Vega was employed at Annai Systems during part of the project. Ronny Drapkin serves on the scientific advisory board of Repare Therapeutics and Siamab Therapeutics. Rosalind Eeles has received an honorarium for the GU-ASCO meeting in San Francisco in January 2016 as a speaker, a honorarium and support from Janssen for the RMH FR meeting in November 2017 as a speaker (title: genetics and prostate cancer), a honorarium for an University of Chicago invited talk in May 2018 as speaker and an educational honorarium paid by Bayer & Ipsen to attend GU Connect 'Treatment sequencing for mCRPC patients within the changing landscape of mHSPC' at a venue at ESMO, Barcelona, on 28 September 2019. Paul Flicek is a member of the scientific advisory boards of Fabric Genomics and Eagle Genomics. Ronald Ghossein is a consultant for Veracyte. Dominik Glodzik is an inventor on a number of patent applications that encompass the use of mutational signatures. Eoghan Harrington is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Yann Joly is responsible for the Data Access Compliance Office (DACO) of ICGC 2009-2018. Sissel Juul is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Vincent Khoo has received personal fees and non-financial support from Accuray, Astellas, Bayer, Boston Scientific and Janssen. Stian Knappskog is a coprincipal investigator on a clinical trial that receives research funding from AstraZeneca and Pfizer. Ignaty Leshchiner is a consultant for PACT Pharma. Carlos López-Otín has ownership interest (including stock and patents) in DREAMgenics. Matthew Meyerson is a scientific advisory board chair of, and consultant for, OrigiMed, has obtained research funding from Bayer and Ono Pharma and receives patent royalties from LabCorp. Serena Nik-Zainal is an inventor on a number of patent applications that encompass the use of mutational signatures. Nathan Pennell has done consulting work with Merck, Astrazeneca, Eli Lilly and Bristol-Myers Squibb. Xose S. Puente has ownership interest (including stock and patents in DREAMgenics. Benjamin J. Raphael is a consultant for and has ownership interest (including stock and patents) in Medley Genomics. Jorge Reis-Filho is a consultant for Goldman Sachs and REPARE Therapeutics, member of the scientific advisory board of Volition RX and Paige.AI and an ad hoc member of the scientific advisory board of Ventana Medical Systems, Roche Tissue Diagnostics, InVicro, Roche, Genentech and Novartis. Lewis R. Roberts has received grant support from ARIAD Pharmaceuticals, Bayer, BTG International, Exact Sciences, Gilead Sciences, Glycotest, RedHill Biopharma, Target PharmaSolutions and Wako Diagnostics and has provided advisory services to Bayer, Exact Sciences, Gilead Sciences, GRAIL, QED Therapeutics and TAVEC Pharmaceuticals. Richard A. Scolyer has received fees for professional services from Merck Sharp & Dohme, GlaxoSmithKline Australia, Bristol-Myers Squibb, Dermpedia, Novartis Pharmaceuticals Australia, Myriad, NeraCare GmbH and Amgen. Tal Shmaya is employed at Annai Systems. Reiner Siebert has received speaker honoraria from Roche and AstraZeneca. Sabina Signoretti is a consultant for Bristol-Myers Squibb, AstraZeneca, Merck, AACR and NCI and has received funding from Bristol-Myers Squibb, AstraZeneca, Exelixis and royalties from Biogenex. Jared Simpson has received research funding and travel support from Oxford Nanopore Technologies. Anil K. Sood is a consultant for Merck and Kiyatec, has received research funding from M-Trap and is a shareholder in BioPath. Simon Tavaré is on the scientific advisory board of Ipsen and a consultant for Kallyope. John F. Thompson has received honoraria and travel support for attending advisory board meetings of GlaxoSmithKline and Provectus and has received honoraria for participation in advisory boards for MSD Australia and BMS Australia. Daniel Turner is a full-time employee of Oxford Nanopore Technologies and is a stock holder. Naveen Vasudev has received speaker honoraria and/or consultancy fees from Bristol-Myers Squibb, Pfizer, EUSA pharma, MSD and Novartis. Jeremiah A. Wala is a consultant for Nference. Daniel J. Weisenberger is a consultant for Zymo Research. Dai-Ying Wu is employed at Annai Systems. Cheng-Zhong Zhang is a cofounder and equity holder of Pillar Biosciences, a for-profit company that specializes in the development of targeted sequencing assays. The other authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1969-6.
Correspondence and requests for materials should be addressed to P.J.C., G.G., J.O.K., J.M.S. or L.D.S.
Peer review information *Nature* thanks Arul Chinnaiyan, Ben Lehner, Nicolas Robine and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01927-8

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[1,2,21], Beatriz Garcia-Torre[1,21], Martí Duran-Ferrer[1,2], Kevin J. Dawson[5], Marta Kulis[1], Ander Diaz-Navarro[2,6], Neus Villamor[1,2,7], Juan L. Melero[8], Vicente Chapaprieta[1], Ana Dueso-Barroso[3], Julio Delgado[1,2,7,9], Riccardo Moia[10], Sara Ruiz-Gil[4], Domenica Marchese[4], Ariadna Giró[1,2], Núria Verdaguer-Dot[1], Mónica Romo[1], Guillem Clot[1,2], Maria Rozman[1,7], Gerard Frigola[7], Alfredo Rivas-Delgado[1,7], Tycho Baumann[2,7,20], Miguel Alcoceba[2,11], Marcos González[2,11], Fina Climent[12], Pau Abrisqueta[13], Josep Castellví[13], Francesc Bosch[13], Marta Aymerich[1,2,7], Anna Enjuanes[1], Sílvia Ruiz-Gaspà[1], Armando López-Guillermo[1,2,7,9], Pedro Jares[1,2,7,9], Sílvia Beà[1,2,7,9], Salvador Capella-Gutierrez[3], Josep Ll. Gelpí[3,9], Núria López-Bigas[14,15,16], David Torrents[3,16], Peter J. Campbell[5], Ivo Gut[4,15], Davide Rossi[17], Gianluca Gaidano[10], Xose S. Puente[2,6], Pablo M. Garcia-Roves[9,18], Dolors Colomer[1,2,7,9], Holger Heyn[4,15], Francesco Maura[19], 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)^{high}–B cell receptor (BCR)^{low}-signaling transcriptional axis in RT and showed that OXPHOS 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 evolution[1] 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 studies[4]. A better knowledge of this process might translate into anticipation-based treatment strategies[5]. RT in CLL represents a paradigmatic model of cancer evolution occurring rarely in treatment-naive patients with CLL but found in 4–20% of patients after chemoimmunotherapy (CIT) and targeted therapies[6]. RT sometimes occurs within the first months after treatment initiation[7–9], suggesting selection of pre-existing subclones[10]. 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; $n = 17$),

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. [3]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [5]Wellcome Sanger Institute, Hinxton, UK. [6]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. [7]Hospital Clínic of Barcelona, Barcelona, Spain. [8]Omniscope, Barcelona, Spain. [9]Universitat de Barcelona, Barcelona, Spain. [10]Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy. [11]Biología Molecular e Histocompatibilidad, IBSAL-Hospital Universitario, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain. [12]Hospital Universitari de Bellvitge-Institut d'Investigació Biomédica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [13]Department of Hematology, Vall d'Hebron Institute of Oncology, Vall d'Hebron University Hospital, Barcelona, Spain. [14]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [15]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [16]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [17]Oncology Institute of Southern Switzerland, Bellinzona, Switzerland. [18]Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [19]Myeloma Service, Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA. [20]Present address: Hospital Universitario 12 de Octubre, Madrid, Spain. [21]These authors contributed equally: Ferran Nadeu, Romina Royo, Ramon Massoni-Badosa, Heribert Playa-Albinyana, Beatriz Garcia-Torre. ✉e-mail: nadeu@recerca.clinic.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[12–18,21–24] (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)-κB (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. [16]) 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 CLL[26]. *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^{R110} mutation, which triggers cell-autonomous BCR signaling[27] (Fig. 1c).

In addition to the high frequency of CNAs previously identified in RT[11,14], 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 eta) 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 melphalan[34] or ganciclovir[35], 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 allogenic 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 *CIITA*) (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

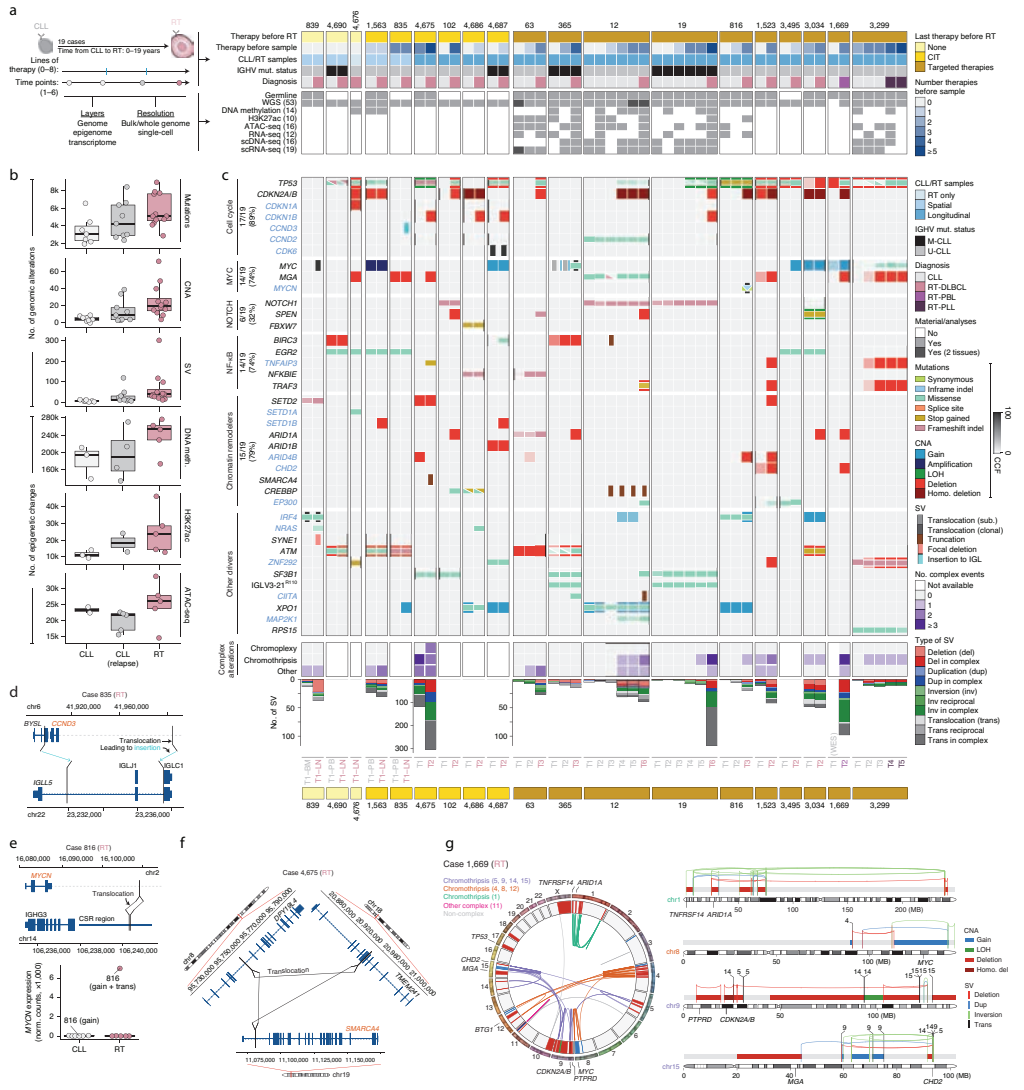


Fig. 1 | The genomic landscape of RT. a, Summary of the study. mut., mutation. **b**, Increase in genomic alterations and epigenetic changes compared to healthy naive and memory B cells over the disease course. Center line indicates median; box limits indicate upper and lower quartiles; whiskers indicate 1.5 × 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.

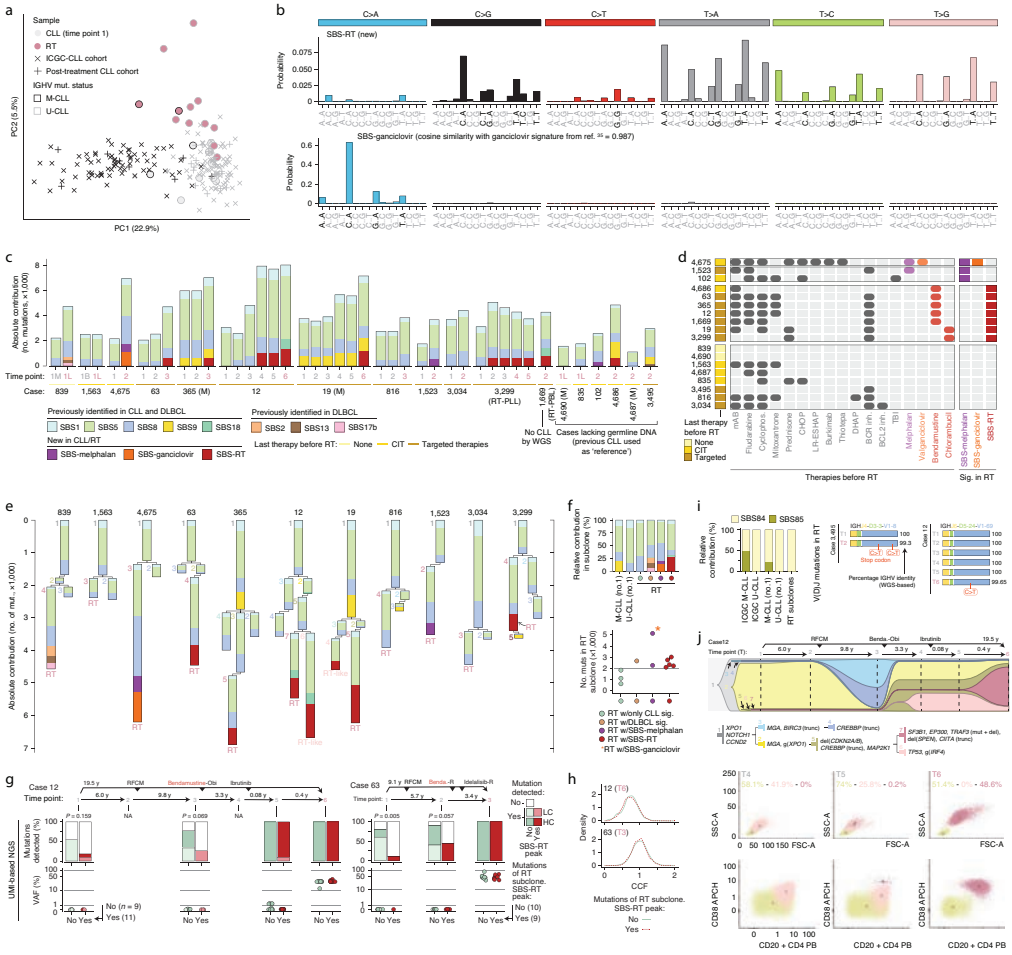


Fig. 2 | Mutational processes in RT. a, Principal component analysis (PCA) of the 96-mutational profile of CLL and RT. **b**, Signatures identified de novo in CLL/RT not reported in COSMIC. The main peaks of each signature are labeled in black. **c**, Contribution of mutational processes in CLL/RT. RT 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 (muts) in RT subclones (bottom). w/, with. **g**, Detection (top) 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. **h**, 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. **i**, 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). **j**, 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[34,36].

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. 2e). 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[37–39] 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 intraclonal 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

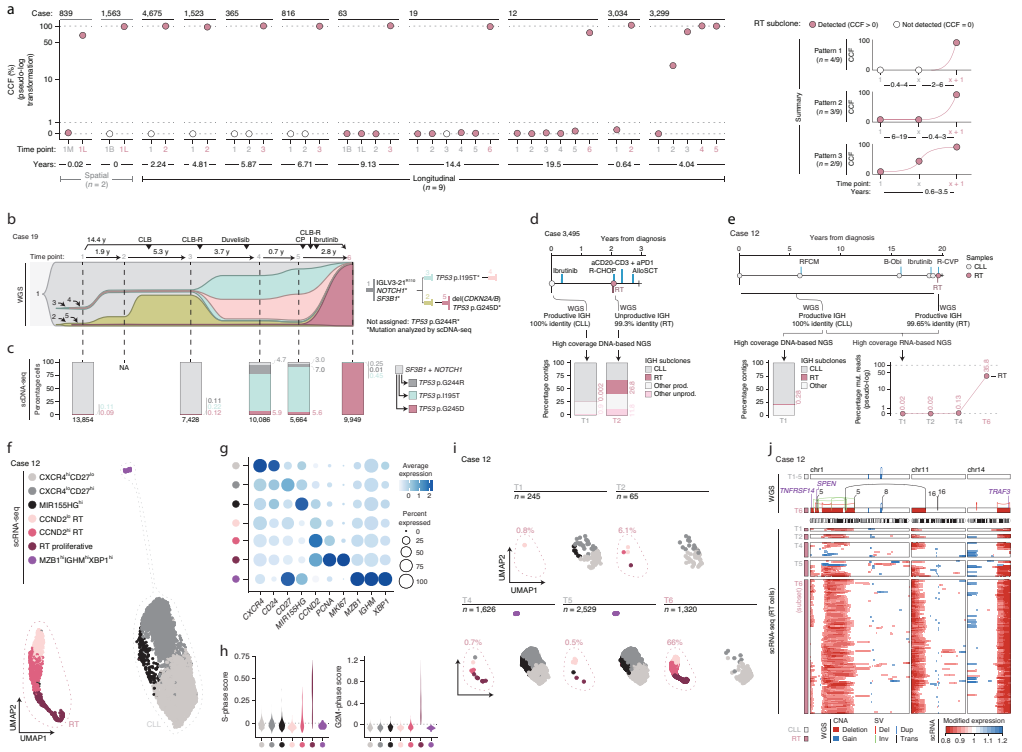


Fig. 3 | Early seeding of RT. a, Evolution of the RT subclone along the disease course based on WGS. Time lapse between the first and last sample analyzed (bottom). RT time points are marked in a rose color. Summary of the three patterns observed (right). **b**, Fish plot showing the clonal evolution 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. **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. 3j 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.[43] 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. 7j). Overall, our integrative analyses uncovered a widespread early seeding of RT cells up to 19 years before their expansion and clinical manifestation.

OXPHOS^{high}–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 CLL[44]. Among them, 24 were enriched and upregulated in RT (Supplementary Table 7). The top TF was TEAD4, which

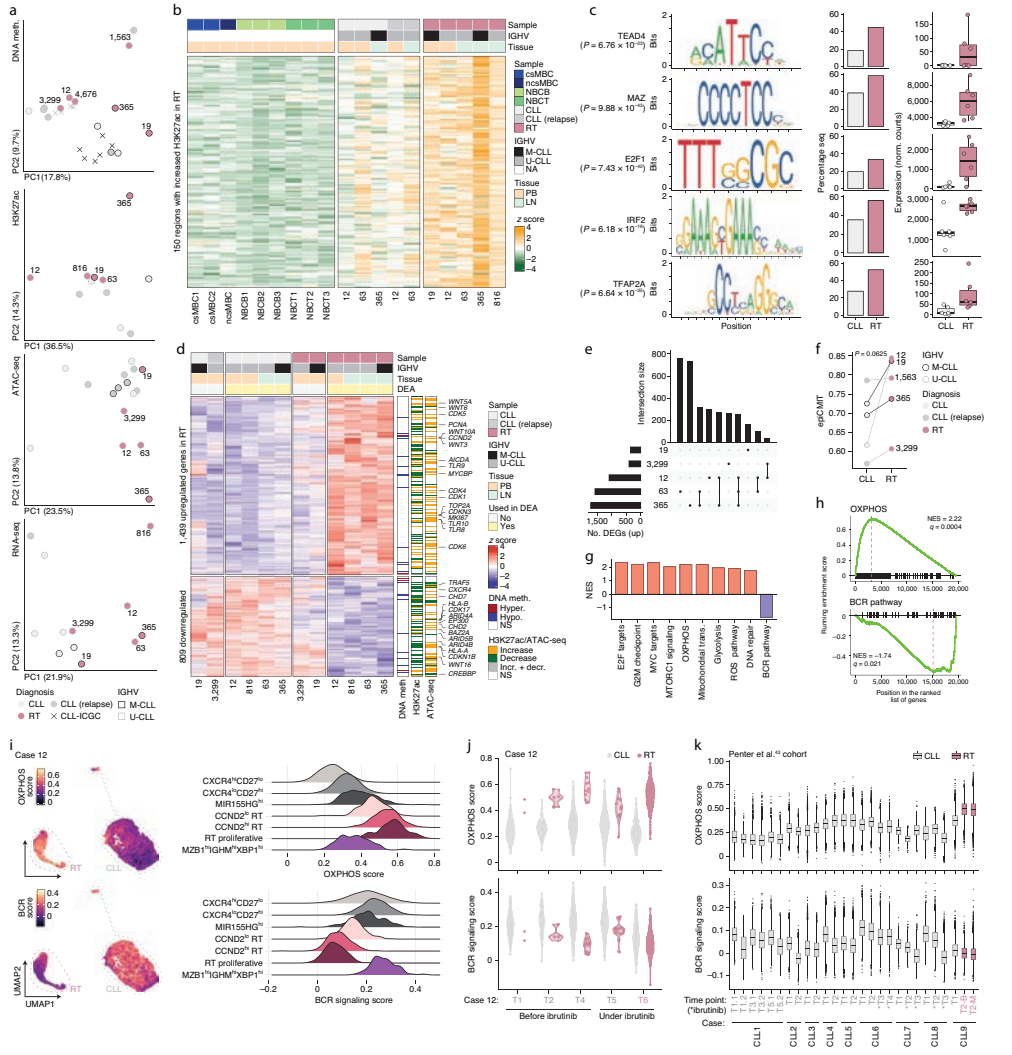


Fig. 4 | Proliferation, OXPHOS and BCR pathways dominate the epigenome and transcriptome of RT. a, PCA of the bulk epigenetic and transcriptomic layers analyzed. **b**, Heat map showing 150 regions with increased H3K27ac levels in RT. **c**, TF enriched within the ATAC peaks identified in the regions 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 1.5× 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). **i**, OXPHOS and BCR signaling scores depicted at single-cell level for case 12 (all time points together). RT and CLL 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 1.5× interquartile range; points indicate outliers. B, peripheral blood; M, bone marrow. *Sample collected under treatment with ibrutinib.

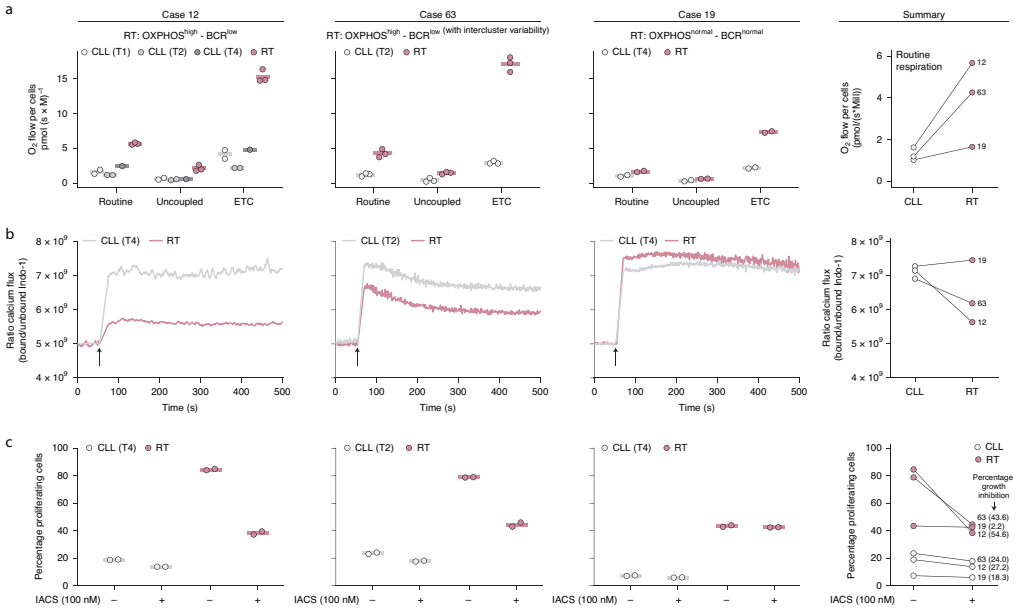


Fig. 5 | Cellular respiration, BCR signaling and OXPHOS inhibition in RT cells. a, Oxygen consumption of intact CLL and RT cells of three patients at routine respiration (routine), oligomycin-inhibited leak respiration (uncoupled) and uncoupler-stimulated ETC. Each dot represents a technical replicate. 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 5×10^9 Indo-1 ratio for 60 s before cell stimulation with F(ab′)2 anti-human IgM + H_2O_2 at 37 °C. Then, Ca^{2+} 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 pathway[45] and co-operates with MYCN[46]. 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 CLL[47], whereas they are necessary to induce MYC target genes, OXPHOS and glycolysis in activated healthy B cells[48].

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 regulation[49]. 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 epiCMIT score[49] (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)[50], Toll-like receptors (*TLR9* among others)[51] 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 $IGLV3\text{–}21^{R110}$ 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.[43]. Case CLL9, which

developed RT in the absence of any therapy, showed a remarkably higher OXPHOS 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 OXPHOS^{high}–BCR^{low} axis reminiscent of that observed in the de novo DLBCL subtype characterized by high OXPHOS (DLBCL-OXPHOS) and insensitive to BCR inhibition[52–54]. This axis might explain the selection and rapid expansion of small RT subclones under therapy with BCR inhibitors.

OXPHOS and BCR activity in RT. We next validated experimentally the OXPHOS and BCR activity of RT in samples of patients 12, 19 and 63. Respirometry assays confirmed that OXPHOS^{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, OXPHOS^{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 Ca^{2+} mobilization upon BCR stimulation with IgM showed that BCR^{low} RT cells (patients 12 and 63) had a lower Ca^{2+} 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} mutation[27] (Fig. 5b, Supplementary Fig. 4a,b and Supplementary Table 25).

To determine the biological effect of OXPHOS^{high} in RT, we performed in vitro proliferation assays using IACS-010759 (100 nM), an OXPHOS inhibitor that targets mitochondrial complex I (Supplementary Figs. 3e and 4c and Supplementary Table 25). OXPHOS^{high} RT (patients 12 and 63) had a higher proliferation at 72 h compared to OXPHOS^{normal} RT (patients 19) and all of them were higher than their respective CLL. OXPHOS inhibition resulted in a marked decrease in proliferation in OXPHOS^{high} RT (mean 49.1%), which contrasted with that observed in OXPHOS^{normal} RT (2.2% decrease) and CLL (23.2% decrease) (Fig. 5c and Supplementary Fig. 4d). Overall, these results confirm the role of OXPHOS^{high} phenotype in high proliferation of RT and suggest its potential therapeutic value in RT as proposed for other neoplasms[53–57].

Discussion

The genome of RT is characterized by a compendium of driver alterations in cell cycle, MYC, NOTCH and NF-κB 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 OXPHOS pathway and downregulation of BCR signaling, the latter potentially compensated by activating Toll-like, MYC and MAPK pathways[17,51,58,59]. 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 signaling[60]. Finally, we also uncovered that OXPHOS inhibition reduced the proliferation of RT cells in vitro, a finding worth exploring in future therapeutic strategies[55,57].

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. Dentro, 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. Beà, 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. Chigrinova, 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 leukaemia. *Nat. Commun.* **8**, 15746 (2017).
28. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
29. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
30. Maura, 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. Gaiti, F. et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
38. Gemenetzi, 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: fractions 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. Penter, 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–MYCN 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. Ntoufa, S., Vilia, M. G., Stamatopoulos, K., Ghia, P. & Muzio, M. Toll-like receptors signaling: a complex network for NF-κB 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. Molina, 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**, eaau1167 (2019).
58. Varano, G. et al. The B-cell receptor controls fitness of MYC-driven lymphoma cells via GSK3β 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 leukaemia treated with BCR signalling inhibitors. *Br. J. Haematol.* **177**, 324–328 (2017).

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/cryopreserved 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/cryopreserved 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; leukocyte 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 $TCR\gamma\delta^+$ 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 (PECy7) versus CD20 (PacB), CD5 (PerCPCy5.5) versus CD20 (PacB) and CD20 (PacB) versus CD38 (APC-H7) plots to evaluate the expression of these B cell markers and the assignation 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. 07962363001) 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)[28] (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 (v.0.7.15)[61]. BAM files were generated and optical/PCR duplicates flagged 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. Mean coverage was 33× and 119× for WGS and WES, respectively (Supplementary Table 1).

Immunoglobulin gene characterization. Immunoglobulin gene rearrangements were characterized using IgCaller (v.1.2)[62]. The rearranged sequences obtained were reviewed on the Integrative Genomics Viewer (IGV; v.2.9.2)[63] and annotated using IMGT/V-QUEST (https://www.imgt.org/IMGT_vquest) and ARResT/AssignSubsets (http://bat.infspire.org/arrest/assignsubsets).

Tumor versus normal SNVs and indel calling. SNVs were called using Sidrón[28], CaVEMan (cgpCaVEManWrapper, v.1.12.0)[64], Mutect2 (Genome Analysis Toolkit (GATK) v.4.0.2.0)[65] and MuSE (v.1.0 rc)[66] and normalized using bcftools (v.1.8)[67]. 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 (v.0.9.4)[68], Pindel (cgpPindel, v.2.2.3)[69], SvABA (v.7.0.2)[70], Mutect2 (GATK v.4.0.2.0)[65] and Platypus (v.0.8.1)[71]. 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 bcftools[67]. 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 (v.4.3t)[72] 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 study[73].

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). Mini-BAM files were obtained using Picard tools and variant calling was performed using Mutect2 (GATK v.4.0.4.0)[65], VarScan2 (v.2.4.3)[74], VarDictJava (v.1.4)[75], LoFreq (v.2.1.3.1)[76], outLyzer (v.1.0)[77] and freebayes (v.1.1.0, https://github.com/freebayes/freebayes). Variants were normalized using bcftools (v.1.9)[67] and annotated using snpEff/snpSift (v.4.3t)[72]. 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/WGS[28] were considered as polymorphisms.

Tumor versus normal CNA calling. CNAs were called using Battenberg (cgpBattenberg, v.3.2.2)[78] and ASCAT (ascatNgs, v.4.1.0)[79]. 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 (v.0.9.3)[80]. CNAs <500 kb, with an absolute $\log_2$ copy ratio ($\log_2$CR) < 0.3 or located within any of the immunoglobulin loci were removed. CNAs were classified as gains if $\log_2\text{CR} > 0.3$, deletions if $\log_2\text{CR} < -0.3$, high-copy gains if $\log_2\text{CR} > 1.1$ and homozygous deletions if $\log_2\text{CR} < -1.1$. The $\log_2$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 $\log_2$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 (v.0.9.4)[68], BRASS (v.6.0.5)[81], SvABA (v.7.0.2)[70] and DELLY2 (v.0.8.1)[82]. 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 (v.1.2)[62] was used to call SVs within any of the immunoglobulin loci. All SVs were visually inspected using IGV[83]. SVs were categorized into simple or complex events. Chromothripsis[83] 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 criteria[83,84]. Chromoplexy was determined by the presence of ≥3 chained chromosomal rearrangements, where chains were identified using a window of 50 kb[85,86]. Cycles of templated insertions were defined as copy number gains in ≥3 chromosomes linked by SVs[87]. 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 allogenic 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 v.4.0.0, 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)[78,88]. 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'[88]. Clusters not assigned to the reconstructed phylogenetic tree were excluded. Fish plots were generated using the TimeScape R package (v.1.6.0). The CCF of indels was calculated integrating read counts, CNAs and tumor purity[89]. Driver indels subjected to validation by scDNA-seq and/or relevant to the tumor 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 ≤1Kb)[29,32]. We integrated the mutations identified in this CLL/RT cohort together with those of 147 CLL treatment-naive samples (ICGC-CLL)[28] and 27 new CLL collected at relapse post-treatment (mean coverage 31.5×; 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 steps[30]:

1. Extraction: de novo signature extraction was performed using a hierarchical Dirichlet process (HDP, v.0.1.5; https://github.com/nicolaroberts/hdp), SignatureAnalyzer (v.0.0.7)[90], SigProfiler (SigProfilerExtractor, v.1.0.8)[32] and sigfit (v.2.0.0; https://github.com/kgori/sigfit). HDP was run with four independent posterior sampling chains, followed by 20,000 burn-in iterations and the collection of 200 posterior samples off 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 (v.3.2)[32] if their cosine similarity was >0.85. Otherwise, the extracted signature was decomposed into '*n*' COSMIC signatures using an expectation maximization (EM) algorithm[91]. 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 Kucab et al.[33] (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 treatment[35] (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, v.3.0.1) 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 therapy[34] was also included in this step. To avoid the so-called inter-sample bleeding effect[30], we iteratively removed the less-contributing signature if its removal decreased the cosine similarity between the original and reconstructed 96-profile <0.01 (ref. [32]). SBS1 and SBS5 were added if addition improved the cosine similarity[32]. Similarly, SBS9 was added in CLL/RT samples classified as M-CLL if addition improved the cosine similarity. We also ran mSigAct (v.2.1.1; https://github.com/steverozen/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 profile[92]. The reference epigenomes of CLL[44] 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 (Posied_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 ≥1 kb with absolute replication timing >0.5 (ref. [93]). 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 0.005 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 package[34]. The replication strand was annotated based on the left/right replication direction of the timing transition regions[94]. The transcriptional strand was annotated using the TxDb.Hsapiens.UCSC.hg19.knownGene R package (v.3.2.2). Finally, kataegis was defined as a genomic region having six or more mutations with an average inter-mutation distance ≤1 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; v.1.15 with parameters: -g CCTACACGACGCTCTTCCGATCT -a AGATCGGAAGAGCACACGTCTGAA -A AGATCGGAAGAGCGTCGTGTAGG -G TTCAGACGTGTGCTCTTCCGATCT -e 0.1 -O 9 -m 20 -n 2). Trimmed FASTQ reads were converted to unmapped BAM using Picard's FastqToSam tool (v.2.10.2). UMI information was extracted and stored as a tag using fgbio ExtractUmisFromBam (http://fulcrumgenomics.github.io/fgbio/; v.1.3.0 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 fgbio 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 (v.0.7.15)[61]. Mean coverage was determined using Picard's CollectTargetedPcrMetrics (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 (v.1.8, parameters: -B -Q 13 -q 10 -d 100,000 -a FORMAT/DP,FORMAT/AD,FORMAT/ADF,FORMAT/ADR -O v)[67]. 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 FR1-CDR1-FR2-CDR2-FR3 sequence of the rearranged IGHV gene were defined as subclones. Raw FASTQ files were trimmed using Trimmomatic (v.0.36)[95] 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 (v.2.3.1, Invivoscribe Technologies, cat. no. 75000009), 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 IMGT/HighV-QUEST (v.1.8.3; 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 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 (v.2.7.9a; https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md) to the hg38 human genome. IGV[63] 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). R and core Bioconductor packages, including minfi (v.1.34.0)[98], were used to integrate and normalize DNA methylation data[49]. 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.ilm10b4.hg19 package (v.0.6). Tumor cell content of each sample was inferred from DNA methylation[49] 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 epitypes and epiCMIT. We compared the DNA methylation status of each CpG to the mean of such CpGs 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 (epitypes) and epiCMIT score for each CLL and RT sample were calculated[49].

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 (v.0.7.7, parameter: -q 5)[61], duplicated reads were marked using Picard tools (v.2.8.1) and low-quality and duplicated reads were removed using SAMtools (v.1.3.1, parameters: -b -F 4 -q 5 -b -F 1,024)[67]. PhantomPeakQualTools (v.1.1.0) were used to generate wiggle plots and for extracting the predominant insert-size. Peaks were called using MACS2 (v.2.1.1.20160309, parameters for H3K27ac: -g hs -q 0.05 -keep-dup all -nomodel -extsize insert-size; parameters for ATAC-seq: -g hs -q 0.05–keep-dup all -f BAM –nomodel –shift −96 –extsize 200; no input control)[97]. Peaks with q values $<1\times10^{-3}$ 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 cells[44] 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 genomecov function (bedtools, v.2.25.0). 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 4 bp downstream) before using the genomecov function. Variance stabilizing transformation (VST) values were calculated for all consensus peaks using DESeq2 (v.1.28.1)[98], 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, v.3.36.0). 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 $\log_2$FC 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 $\log_2$FC 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 $\log_2$FC > 1). The ATAC-seq RT-specific signature encompassed differential regions common in 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 Jaspar 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 ≥ 10% and overexpressed in RT (bulk RNA-seq, $\log_2$FC > 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 (2×76 bp, Illumina) or NextSeq 2000 (2×100 bp, Illumina). All samples had a tumor purity ≥92% as assessed by flow cytometry (Supplementary Table 11).

Data analysis. Ribosomal RNA reads were filter out using SortMeRNA (v.4.3.2)[99]. Non-ribosomal reads were trimmed using Trimmomatic (v.0.38)[95]. Gene-level counts (GRCh38.p13, Ensembl release 100) were calculated using kallisto (v.0.46.1)[100] and tximport (v.1.14.2). A paired DEA was conducted using DESeq2 (v.1.26.0)[98]. Adjusted P value <0.01 and absolute $\log_2$(fold change) > 1 were used to identify DEGs. Gene set enrichment analysis (GSEA) was conducted using a pre-ranked gene list ordered by $-\log_{10}(P)\times$(sign of fold change) using the 'GSEA' function (clusterProfiler R package, v.3.14.3). We focused on C2 (curated) and Hallmark gene sets from the Molecular Signatures Database (v.7.4) 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 Tapestri Platform (Mission Bio, cat. no. 191335) and a commercial 32-gene panel (Tapestri 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 1× 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 Tapestri Pipeline (v.1, 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 (v.3.7). Loom files generated were analyzed using the Tapestri Insights (v.2.2, 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

developed RT in the absence of any therapy, showed a remarkably higher OXPHOS 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 OXPHOS^{high}–BCR^{low} axis reminiscent of that observed in the de novo DLBCL subtype characterized by high OXPHOS (DLBCL-OXPHOS) and insensitive to BCR inhibition[52–54]. This axis might explain the selection and rapid expansion of small RT subclones under therapy with BCR inhibitors.

OXPHOS and BCR activity in RT. We next validated experimentally the OXPHOS and BCR activity of RT in samples of patients 12, 19 and 63. Respirometry assays confirmed that OXPHOS^{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, OXPHOS^{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 Ca^{2+} mobilization upon BCR stimulation with IgM showed that BCR^{low} RT cells (patients 12 and 63) had a lower Ca^{2+} 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} mutation[27] (Fig. 5b, Supplementary Fig. 4a,b and Supplementary Table 25).

To determine the biological effect of OXPHOS^{high} in RT, we performed in vitro proliferation assays using IACS-010759 (100 nM), an OXPHOS inhibitor that targets mitochondrial complex I (Supplementary Figs. 3e and 4c and Supplementary Table 25). OXPHOS^{high} RT (patients 12 and 63) had a higher proliferation at 72 h compared to OXPHOS^{normal} RT (patients 19) and all of them were higher than their respective CLL. OXPHOS inhibition resulted in a marked decrease in proliferation in OXPHOS^{high} RT (mean 49.1%), which contrasted with that observed in OXPHOS^{normal} RT (2.2% decrease) and CLL (23.2% decrease) (Fig. 5c and Supplementary Fig. 4d). Overall, these results confirm the role of OXPHOS^{high} phenotype in high proliferation of RT and suggest its potential therapeutic value in RT as proposed for other neoplasms[53–57].

Discussion

The genome of RT is characterized by a compendium of driver alterations in cell cycle, MYC, NOTCH and NF-κB 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 OXPHOS pathway and downregulation of BCR signaling, the latter potentially compensated by activating Toll-like, MYC and MAPK pathways[17,51,58,59]. 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 signaling[60]. Finally, we also uncovered that OXPHOS inhibition reduced the proliferation of RT cells in vitro, a finding worth exploring in future therapeutic strategies[55,57].

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. Dentro, 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. Beà, 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. Chigrinova, 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).

on the BCLLATLAS_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 BCLLATLAS_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 $\log_2$FC > 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, logfc.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 $\log_2$FC) 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/CTAT/cnv/gencode_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.[43] 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 publication[43]. 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 O_2k-respirometers (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 r.p.m. 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 µl 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 µM, Sigma-Aldrich, cat. no. R8875, CAS, 83-79-4) and antimycin A (2.5 µM, 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 chamber[109]. The same experimental workflow was used to study cellular respiration in CLL and RT cells after 1 h of treatment with IACS-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 10^6 cells ml^{-1}. After 6 h of incubation at 37 °C and 5% CO_2, cells were centrifuged and resuspended on RPMI-1640 with 4 µM 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% CO_2. 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 µg ml^{-1} anti-human F(ab')2 IgM (Southern Biotech, cat. no. 2022-01) and 3.3 mM H_2O_2 (Sigma-Aldrich, cat. no. H1009). Finally, 2 µM 4-hydroxytamoxifen (4-OHT) (Sigma-Aldrich, cat. no. H6278) was added to all conditions before continue recording for up to 8 min. Intracellular Ca^{2+} release was measured on LSRFortessa (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, singlet identification in FSC-A versus FCS-H plot, tumoral cells (CD19$^+$CD5$^+$) in CD19 (Super Bright 600) versus CD5 (PE-Cy5) plot and Ca^{2+} 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 10^7 cells ml^{-1} and labeled with 0.5 µM 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× insulin-transferrin-selenium (Merk, cat. no. I3146), 10 mM HEPES (Fisher Scientific, cat. no. BP299), 50 µM 2-mercaptoethanol (Gibco, cat. no. 21985-023), 1× Non-Essential Amino Acids (Gibco, cat. no. 11140-050), 1 mM sodium pyruvate (Gibco, cat. no. 11360-070) and 50 µg ml^{-1} gentamicin (Gibco, cat. no. 15710-064) at a concentration of 10^6 cells ml^{-1} supplemented with 0.2 µM CpG DNA TLR9 ligand (ODN2006-TL9; InvivoGen, cat. no. TLRL-2006) and 15 ng ml^{-1} recombinant human IL-15 (R&D Systems, cat. no. 247-ILB-025)[110]. 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 LSRFortessa (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, singlet 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/ferrannadeu/RichterTransformation. 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 epitypes and epiCMIT from DNA methylation data can be found at https://github.com/Duran-FerrerM/Pan-B-cell-methylome.

References

61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Nadeu, F. et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* **11**, 3390 (2020).
63. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
64. 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. Muller, 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**, i333–i339 (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 chromothripsis 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. Dentro, 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. & Beerenwinkel, 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. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* https://doi.org/10.1093/gigascience/giy059 (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., Steinlechner-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 Clínic, 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 (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221, to E.C.), the European Research Council under the European Union's Horizon 2020 Research and Innovation Program (810287, BCLLatlas, to E.C., J.I.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-Amgen Fellowship in Clinical/Translational Cancer Research, 21-40-11-NADE to F.N.), the European Hematology Association (EHA Junior Research Grant 2021, RG-202012-00245 to F.N.), the Lady Tata Memorial Trust (International Award for Research in Leukaemia 2021-2022, LADY_TATA_21_3223 to F.N.), the Generalitat de Catalunya Suport Grups de Recerca AGAUR (2017-SGR-1142 to E.C., 2017-SGR-736 to J.I.M.-S. and 2017-SGR-1009 to D.C.), the Accelerator award CRUK/AIRC/AECC joint funder partnership (AECC_AA17_SUBERO to J.I.M.-S.), the Fundació La Marató de TV3 (201924-30 to J.I.M.-S.), the Centro de Investigación Biomédica en Red Cáncer (CIBERONC; CB16/12/00225, CB16/12/00334, CB16/12/00236), the Ministerio de Ciencia e Innovación (PID2020-117185RB-I00 to X.S.P.), the Fundación Asociación Española Contra el Cáncer (FUNCAR-PRYGN211258SUÁR to X.S.P.), the Associazione Italiana per la Ricerca sul Cancro Foundation (AIRC 5×1,000 no. 21198 to G.G.) 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 Institució 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

F.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., E.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.I.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

F.N. 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.

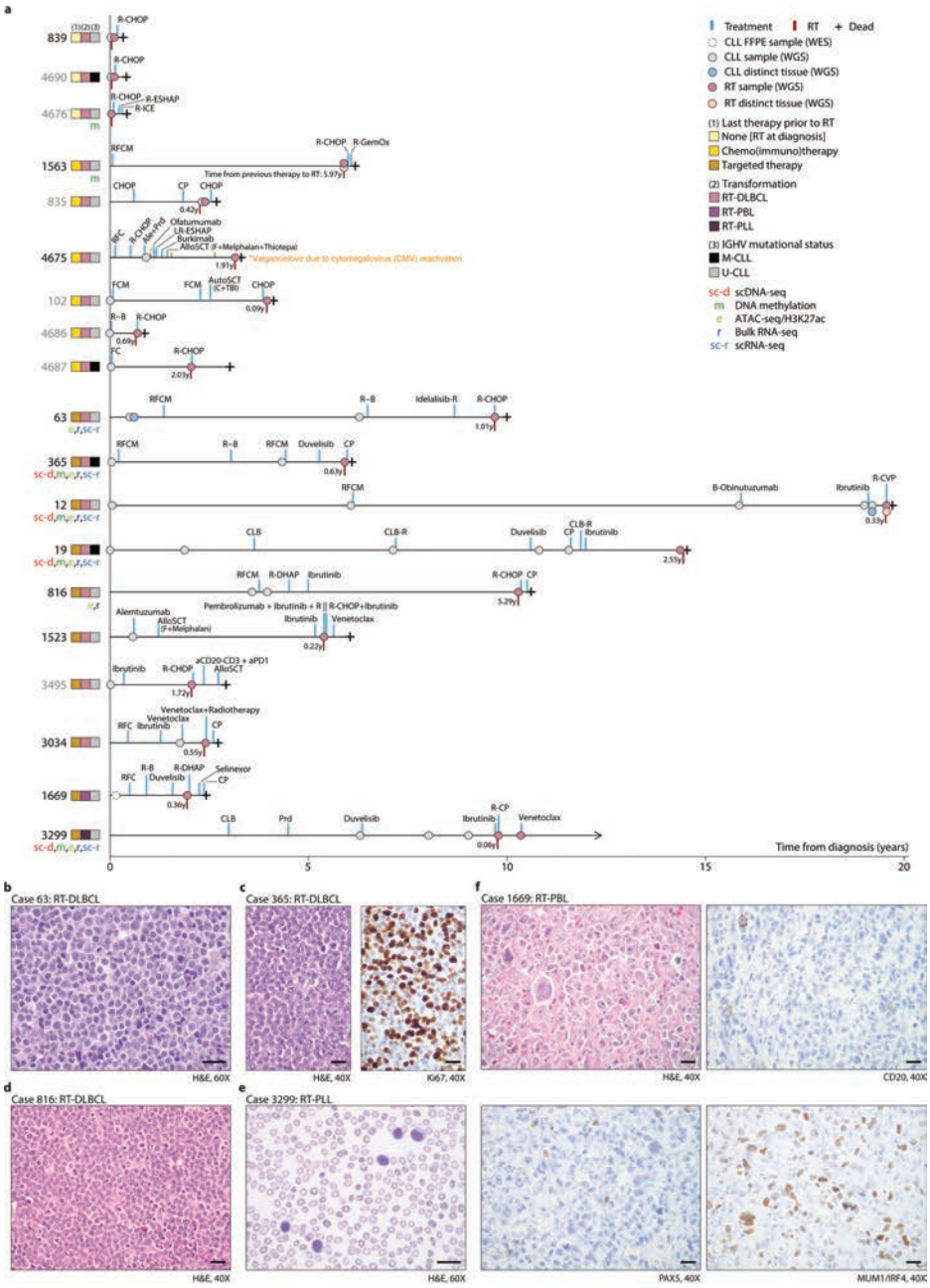
Additional information

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.

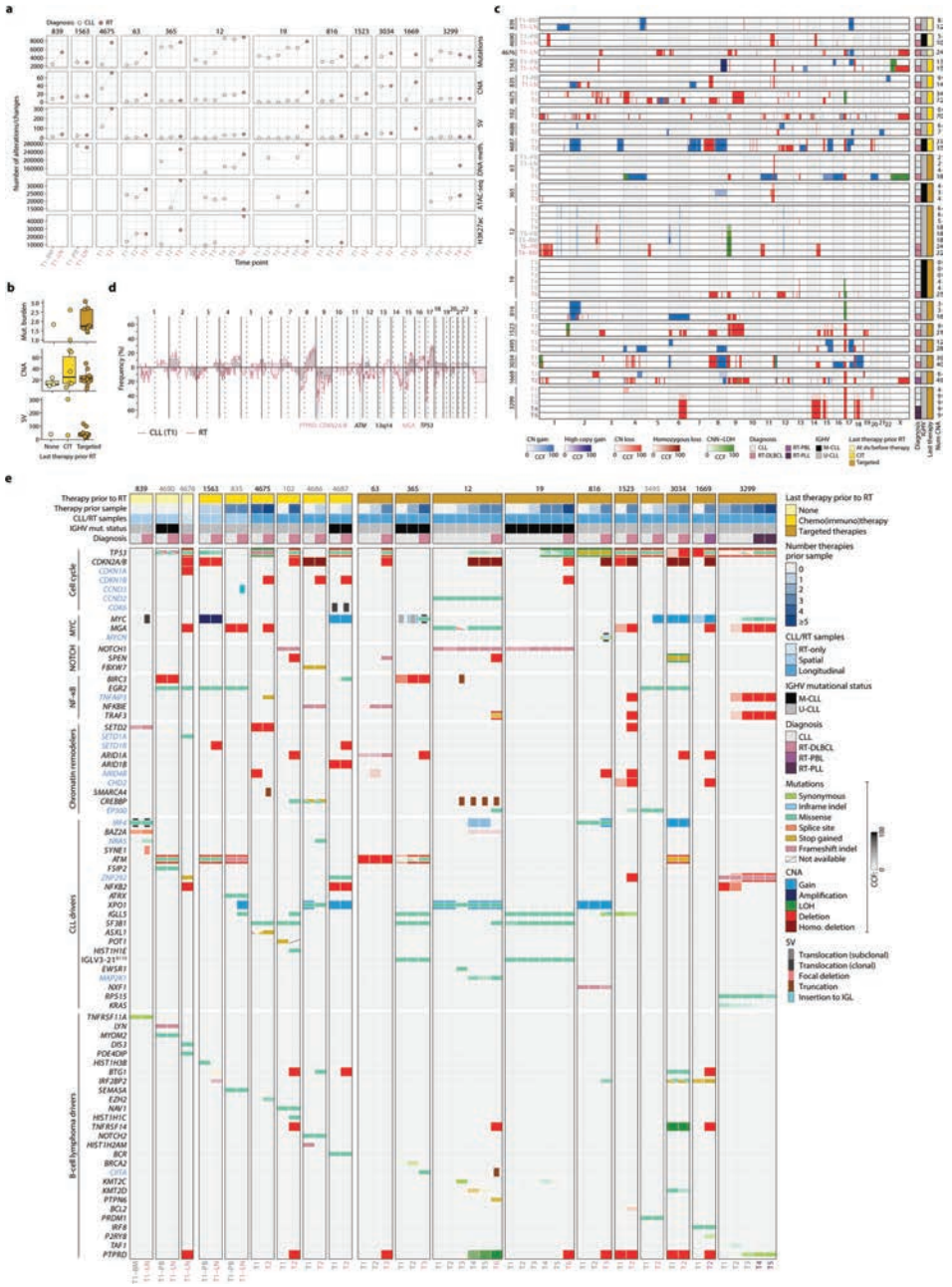
Correspondence and requests for materials should be addressed to Ferran Nadeu or Elías Campo.

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.



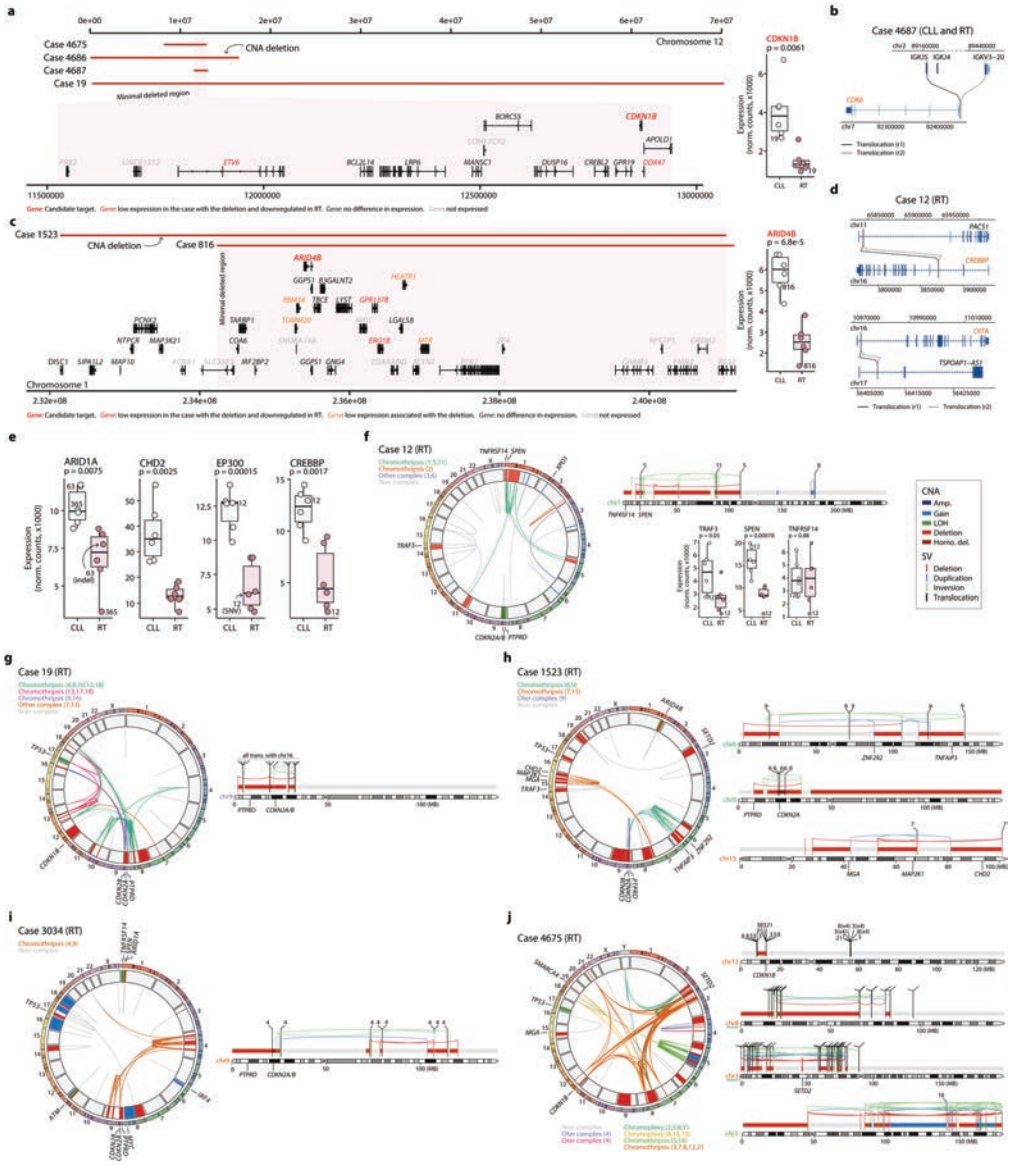
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Cohort studied and types of Richter transformation. a. Representation of the disease course of the patients included in the study. Each sample analyzed, treatment and date of RT are depicted. Patients labeled in gray lacked germline DNA. Patient 4676 also lacked DNA from the previous CLL sample. Patients are grouped based on the last line of therapy received before RT in three groups: patients developing RT before any treatment, after chemo(immuno)therapy, and after targeted therapy. The type of transformation (RT-DLBCL, diffuse large B cell lymphoma type; RT-PLL, prolymphocytic transformation; RT-PBL, plasmablastic transformation) and IGHV mutational status are also shown. Additional molecular studies conducted in each case are also depicted. Abbreviations: Ale: alemtuzumab; AlloSCT: allogenic stem-cell transplantation; AutoSCT: autologous stem-cell transplantation; B: bendamustine; Burkimab: rituximab, methotrexate, dexametasone, ifosfamide, vincristine, etoposide, cytarabine, doxorubicin and vindesine; C: cyclophosphamide; CHOP: cyclophosphamide, doxorubicin, vincristine and prednisone; CLB: chlorambucil; CLB-R: chlorambucil and rituximab; CP: cyclophosphamide and prednisone; F: fludarabine; FCM: fludarabine, cyclophosphamide and mitoxantrone; G-GemOx: rituximab, gemcitabine, and oxaliplatin; LR-ESHAP: lenalidomide, rituximab, etoposide, methyl-prednisolone, cytarabine and cisplatin; M: mitoxantrone; Prd: prednisone; R: rituximab; R-B: rituximab and bendamustine; R-CHOP: rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone; R-CVP: rituximab, cyclophosphamide, vincristine and prednisone; R-DHAP: rituximab, dexamethasone, cytarabine and cisplatin; R-ESHAP: rituximab, etoposide, methyl-prednisolone, cytarabine and cisplatin; RFC: fludarabine, cyclophosphamide and rituximab; RFCM: rituximab, fludarabine, cyclophosphamide and mitoxantrone; R-ICE: rituximab, ifosfamide, carboplatin and etoposide; TBI: total body irradiation. **b**. Morphology of the RT-DLBCL of patient 63 (hematoxylin-eosin, H&E, staining). **c**. Morphology of the RT-DLBCL of patient 365 and Ki67 staining showing high proliferative index. **d**. Morphology of the RT-DLBCL of patient 816. **e**. Morphology of the RT-PLL of patient 3299. **f**. Morphology of the RT-PBL of patient 1669 (H&E staining), which was negative for CD20 and PAX5, while positive for MUM1/IRF4. Each experiment for **b-f** was repeated twice. The scale bars in **b-f** represents 20 μm.



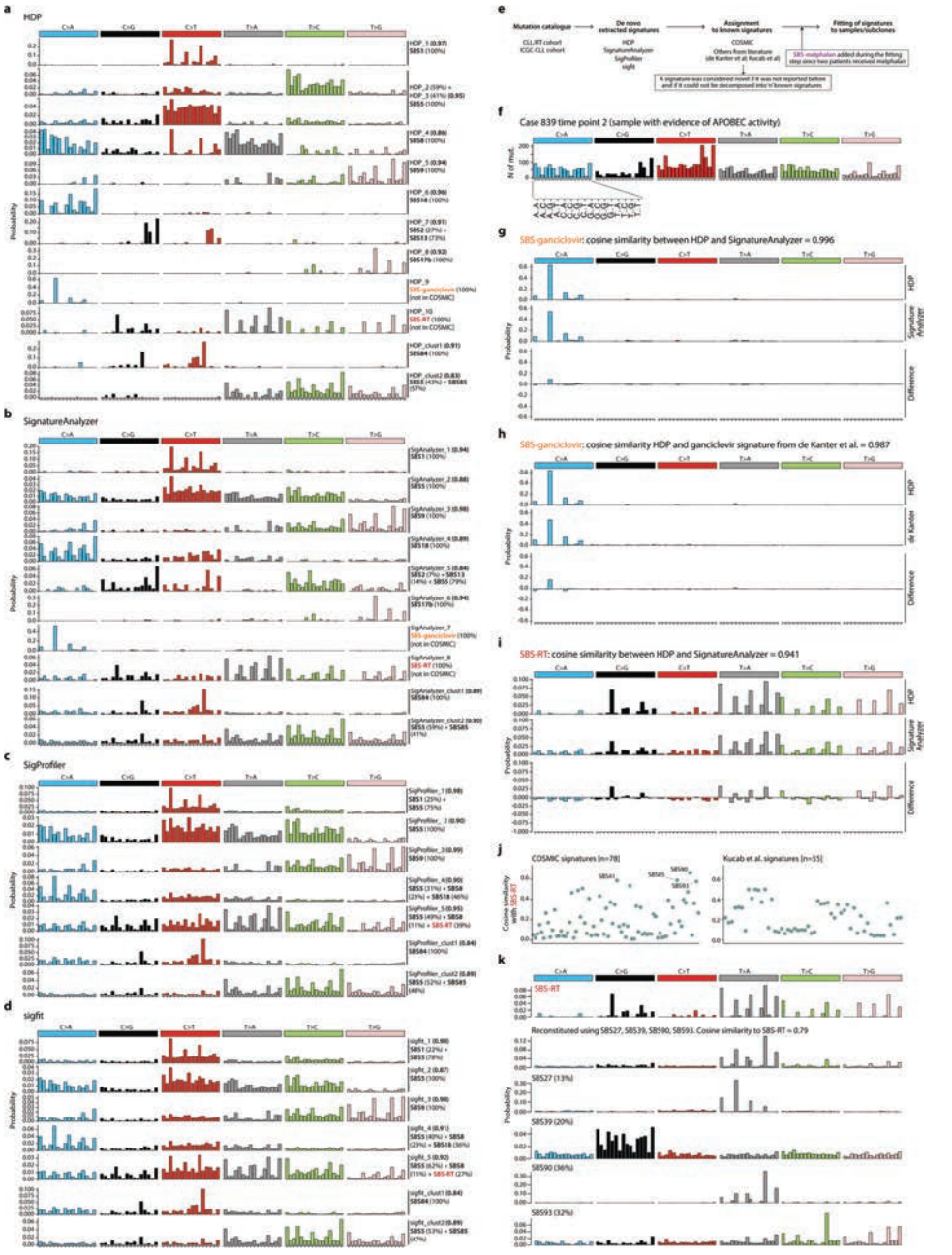
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Genetic and epigenetic changes from CLL to RT, CNA profiles, and landscape of driver alterations. a. Number of somatic genetic alterations and epigenetic changes compared to normal counterparts along the course of the disease. Cases/time points with no grid lines correspond to unavailable data. **b**. Mutational burden, number of CNAs and number of SVs found in RT stratified according to the last therapy prior transformation. Targeted, targeted therapies. center line, median; box limits, upper/lower quartiles; whiskers, 1.5×interquartile range; points, individual samples. **c**. Copy number landscape of the studied cohort grouped by patient. The diagnosis, IGHV mutational status, last therapy prior RT, and total number of CNAs are indicated for each time point. **d**. Aggregated copy number profile of RT vs CLL. The first CLL samples (time point 1, T1) were considered. The plot shows the percentage of samples with gains (up) and losses (down). Among recurrent alterations found either in CLL or RT samples ($n \geq 5$), deletions of 9p (*PTPRD* and *CDKN2A/B*) and deletions of 15q (*MGA*) were enriched in RT whereas deletions of *ATM* (11q), *TP53* (17p), and 13q14 were found at similar frequencies in CLL and RT. **e**. Oncoprint of putative driver alterations. Samples, grouped by patient (patient id at the top), are represented by columns while genes in rows. Novel drivers in RT are labeled in blue. Genes are grouped according to their biological function or if they were previously described as potential driver genes in CLL and/or mature B cell lymphomas. Metadata including the type of therapy before RT, number of treatment lines before each sample, the spatial/longitudinal nature of the CLL/RT samples analyzed, IGHV mutational status, and diagnosis is detailed in the upper rows. In the main plot, mutations (SNVs and indels) are depicted with horizontal rectangles, CNAs using the background color of each cell, and SVs with vertical rectangles. The transparency of the color of mutations and CNAs indicates the cancer cell fraction (CCF). For patients lacking the germline sample (patient id indicated in gray), the CCF of the alterations could not be inferred and a CCF of 100% was used for illustrative purposes.



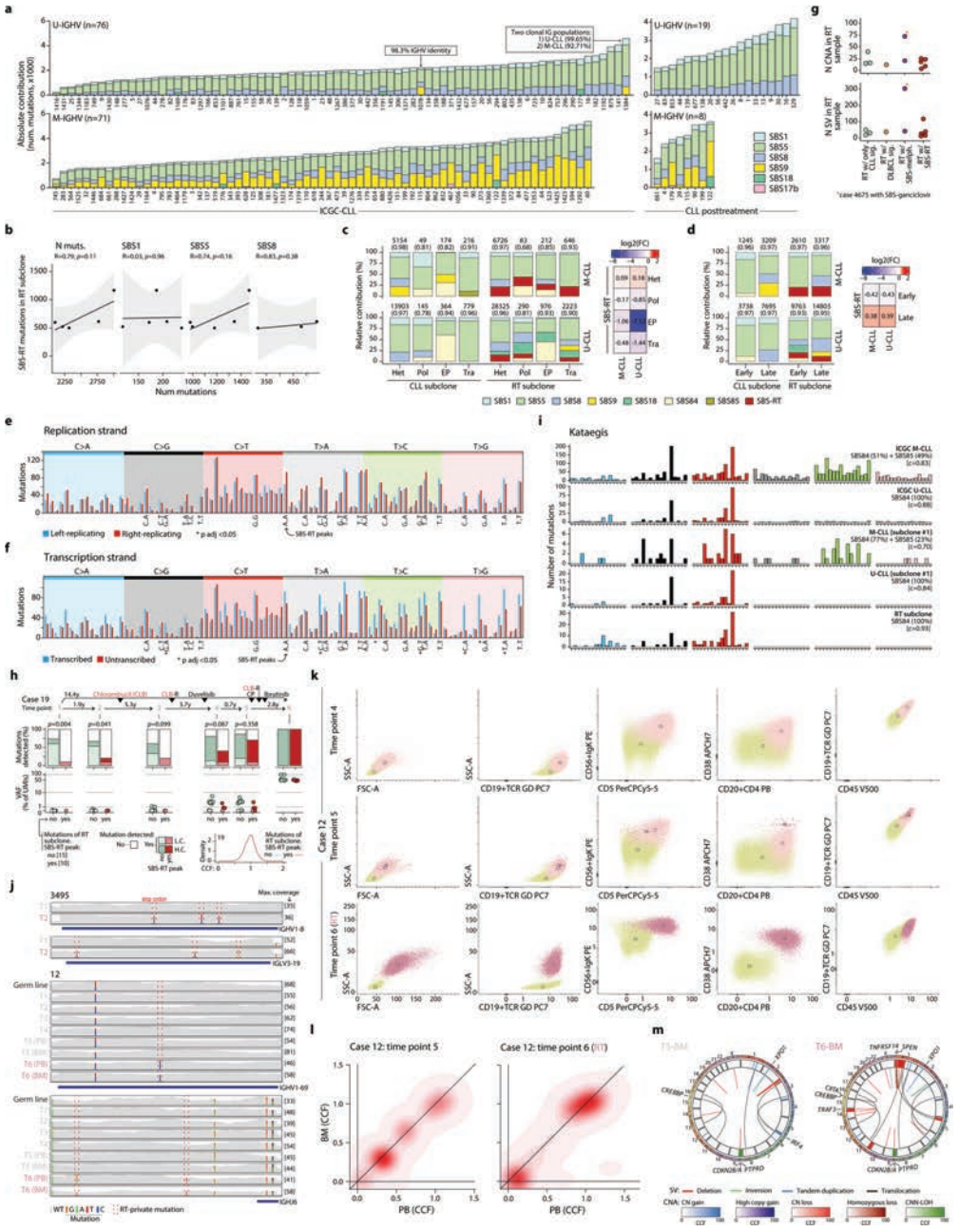
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Complex genomic rearrangements affecting driver genes. a. Deletions in chr12 identified in four cases with the minimal deleted region affecting *CDKN1B*, which expression in CLL and RT sample pairs is shown on the right. The case carrying the deletion at time of RT is labeled in the boxplot. **b**. Reciprocal translocation juxtaposing *CDK6* next to IGKJ5 in patient 4687. **c**. Deletion in chr1 affecting two cases with the minimal deleted region targeting *ARID4B*. Its expression in CLL and RT sample pairs is shown in the boxplot on the right. **d**. Reciprocal translocations truncating *CREBBP* and *CIITA* in the RT sample of patient 12. **e**. Expression levels of known and novel RT-driver genes in CLL and RT paired samples. Cases carrying deletions/mutations at time of RT are labeled. **f-j**. Complex genomic rearrangements affecting driver genes in five selected RT samples. The circos plots show the SVs (inner links) and CNAs (middle circle) found in each sample. SVs are colored based on whether they are part of a complex event, while CNAs are painted according to their type. Chromosome-specific plots on the right show the main chromosomes affected by complex events targeting driver genes (annotated at the bottom). In these chromosome-specific plots, the color of both CNAs and SVs indicates their type. For patient 12 (f), the expression levels of three genes affected by simple (*TRAF3*) and complex (*SPEN* and *TNFRS14*) chromosomal alterations are shown. For patient 4675 (j), the partner of the translocations found in chr3 and chr8 are not specified for simplicity due to the high number of clustered structural events. All boxplots: center line, median; box limits, upper/lower quartiles; whiskers, 1.5×interquartile range; points, individual samples. All p values are from two-sided T tests.



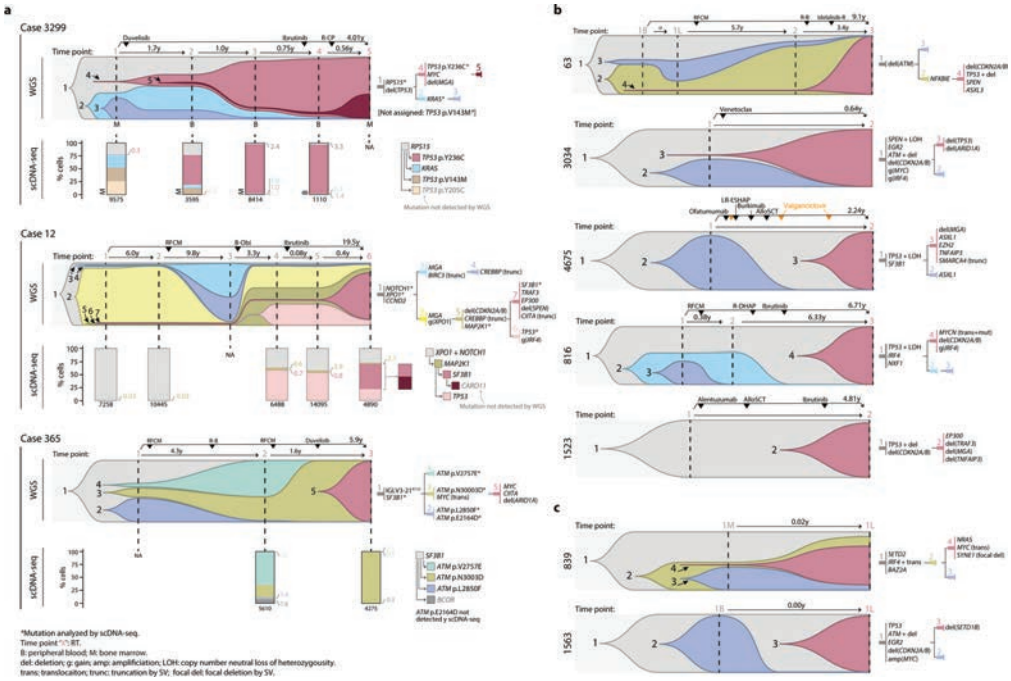
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Extraction and assignment of mutational signatures. a–d. Signatures extracted by the Hierarchical Dirichlet Process (HDP) (a), SignatureAnalyzer (b), SigProfiler (c), and sigfit (d). COSMIC signatures needed to reconstruct the extracted signatures are shown together with their contribution (in percentage). The cosine similarities between the extracted and reconstructed signatures are shown in brackets. **e**. Workflow of the mutational signature analysis. **f**. The 96-mutation profile of the RT sample of patient 839 (time point 2), which had marked evidence of APOBEC activity (SBS2 and SBS13). **g**. Comparison of the SBS-ganciclovir extracted by HDP and SignatureAnalyzer. Based on the high cosine similarity (0.996), we considered that both signatures represented the same mutational process and selected the one extracted by HDP for downstream analyses. **h**. Comparison of the SBS-ganciclovir extracted by HDP and the ganciclovir signature reported by de Kanter et al.[35]. **i**. Comparison of the SBS-RT extracted by HDP and SignatureAnalyzer. Based on the high cosine similarity (0.941), we considered that both signatures represented the same mutational process and selected the one extracted by HDP for downstream analyses. **j**. Pairwise comparisons of the SBS-RT with known signatures from COSMIC and Kucab et al.[33]. **k**. Decomposition of the SBS-RT in "n" known signatures using an expectation maximization approach. The low cosine similarity (<0.85) between SBS-RT and the best reconstituted signature obtained using any combination of known signatures suggests that SBS-RT represents a novel mutational signature.

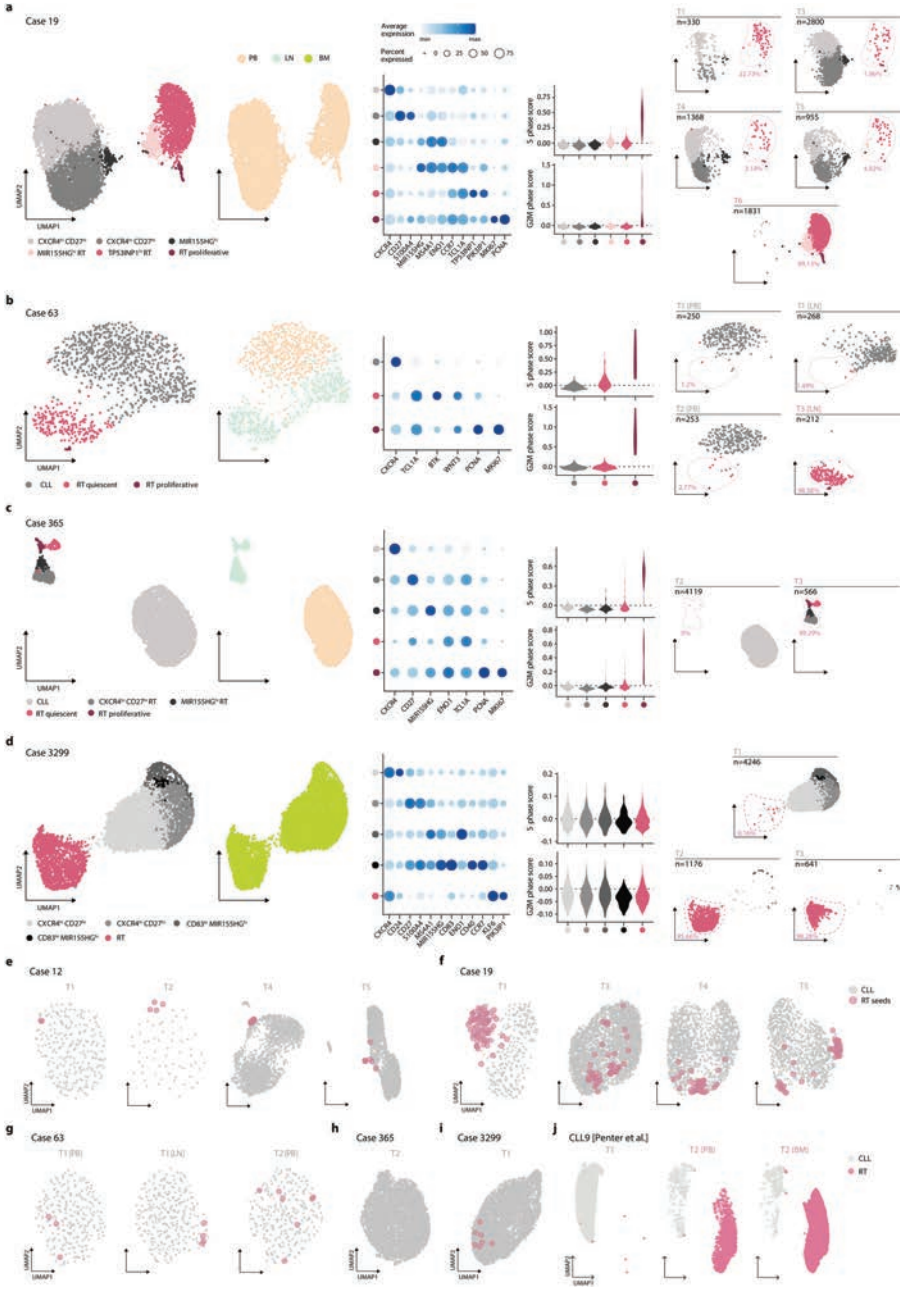


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Fitting of mutational signatures, characterization of SBS-RT, and co-occurrence of RT subclones. a. Mutational processes in ICGC-CLL (left) and post-treatment CLL (right) cohorts. **b**. Correlation of SBS-RT with the total number of SNVs and other mutational processes in RT subclones. Gray area, 95% confidence interval. **c**. Activity of the mutational processes identified in regulatory regions of the genome: heterochromatin (Het), polycomb (Pol), enhancer/promoter (EP), and transcription (Tra). The heat map (right) shows the log2-fold change of the observed vs expected number of SBS-RT mutations/region. **d**. Contribution of the mutational processes in early/late replication regions. **e-f**. Replication (e) and transcriptional (f) strand bias of the mutational profile of RT subclones with SBS-RT. The main peaks of the SBS-RT are indicated with their context on the x-axis. Significant asymmetries are indicated with asterisks (exact *p* values are listed in Supplementary Table 16). **g**. Number of CNAs and SVs in RT samples. **h**. Detection (top) and variant allele frequency (VAF) (bottom) of mutations assigned to the RT subclone during the disease course in patient 19 based on UMI-based NGS. Mutations are grouped according to the main peaks of SBS-RT. *P* values by Fisher's test. L.C., low confidence; H.C., high confidence. Density plot showing the distribution of the cancer cell fraction (CCF) of the SNVs assigned to the RT subclone by WGS (bottom right). **i**. Mutational profiles of kataegis in ICGC-CLL samples (row 1–2), CLL subclones from the present CLL/RT cohort (row 3–4), and RT subclones (all U-CLL) (row 5). Mutational processes identified are indicated together with its contribution and cosine similarity to the reconstructed profile. **j**. Immunoglobulin genes of two cases harboring RT-specific SNVs at time of RT (time points, T, highlighted in rose). PB, peripheral blood. BM, bone marrow. **k**. Complete flow cytometry analysis in case 12. Numbers along axes are divided by 1000. **l**. Density plot showing the comparison of the CCF of the SNVs of synchronous BM and PB samples analyzed in patient 12. **m**. Circos plots of the BM samples of patient 12 for comparison with the rearrangements observed at PB (Supplementary Fig. 1).

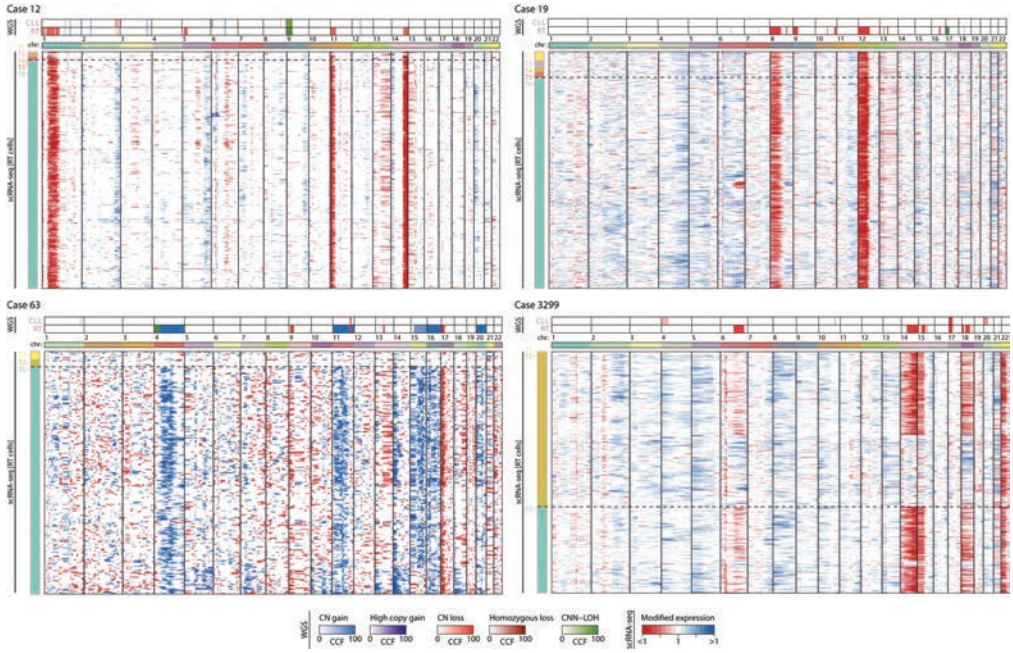


Extended Data Fig. 6 | Clonal dynamics from CLL to RT. a. Subclonal reconstruction and clonal evolution of three cases (3299, 12 and 365) with WGS and scDNA-seq data available. The upper fish plot shows the clonal evolution along the course of the disease inferred from WGS analyses. Each color represents a different subclone and their height is proportional to their cancer cell fraction (CCF) in each time point (vertical lines). The treatments that the patient received and the elapsed time (in years) between samples are indicated at the top. The tissue is indicated for samples of patient 3299 in which different tissues were analyzed by WGS and scDNA-seq in the same time point. The phylogeny of the subclones is depicted together with the main driver alterations (top right). The lower bar plots show the dynamics of the different subclones according to the scDNA-seq analyses. The total number of cells per sample is shown at the bottom. The number of cells assigned to each subclone can be found in Supplementary Table 20. The mutation tree inferred from scDNA-seq data is shown at the bottom-right part. **b-c**. Subclonal architecture and dynamics of six cases with longitudinal samples (b) and two cases with spatial samples (c) analyzed by WGS.

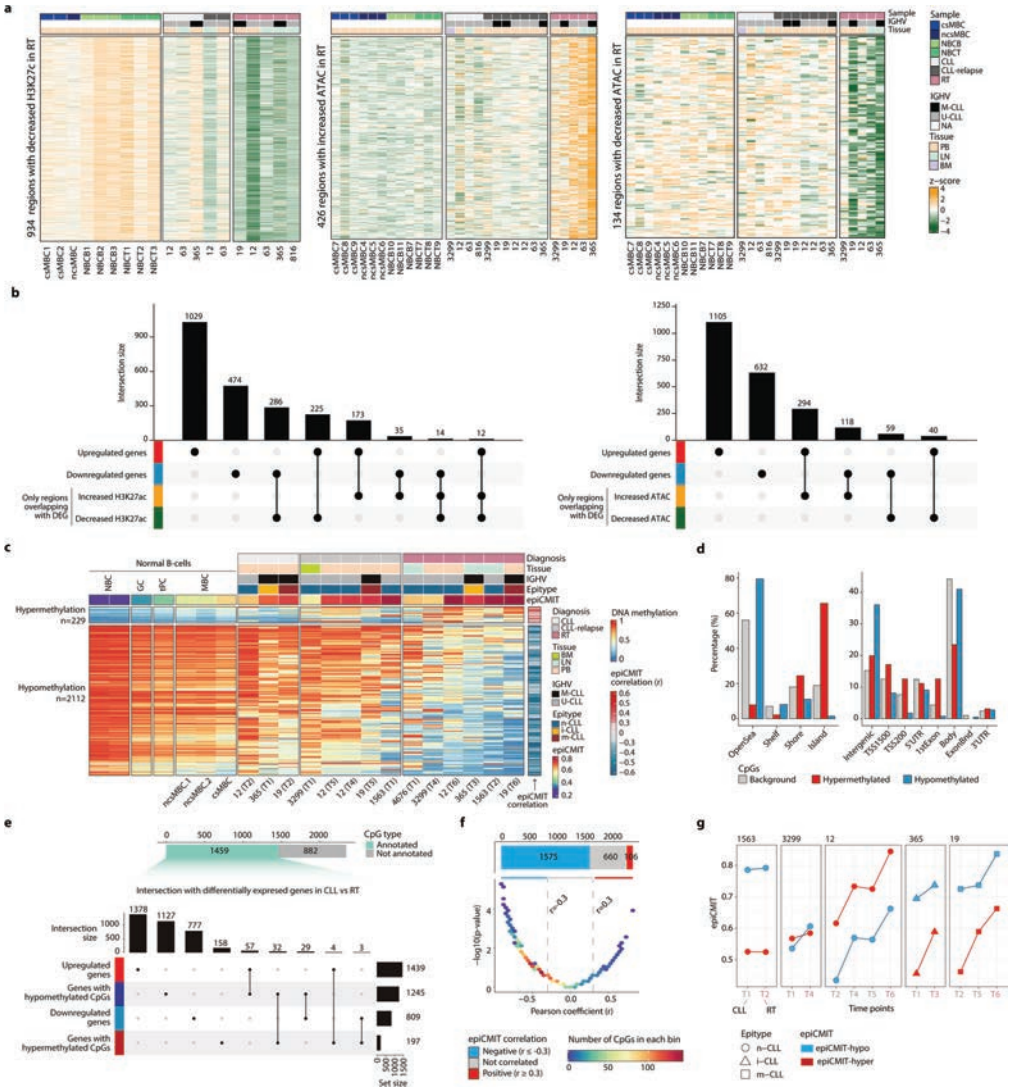


Extended Data Fig. 7 | See next page for caption.

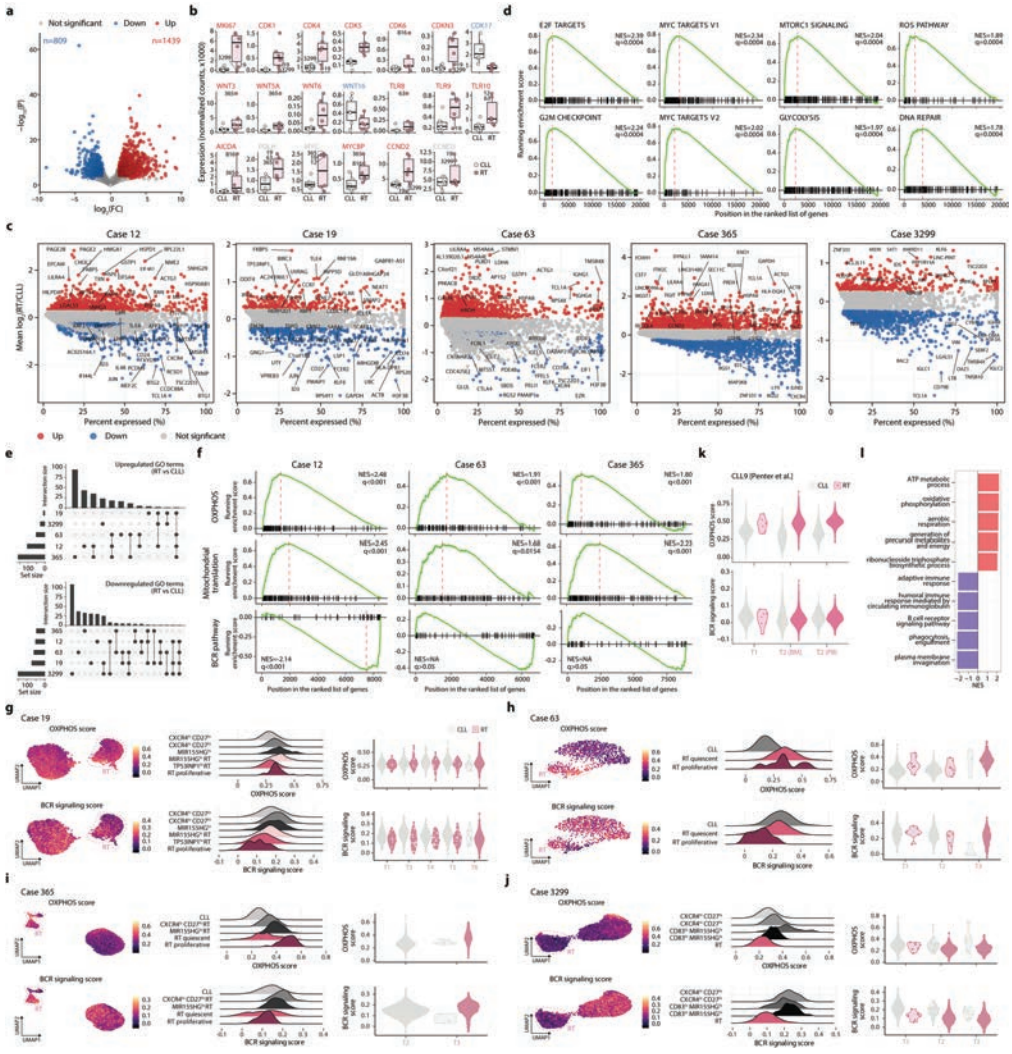
Extended Data Fig. 7 | scRNA-seq characterization of CLL and RT. a–d. UMAP visualization of tumor cells from all time points colored by annotation and tissue of origin. hi, high; lo, low; PB, peripheral blood; LN, lymph node; BM, bone marrow (left). Dot plot with the expression of key markers in each cluster. Color and size represent scaled mean expression and proportion of cells expressing each marker gene, respectively (middle-left). Violin plots showing the cell-cycle phase scores (S and G-to-M) for each cluster of cells (middle-right). UMAP visualization split by time point (right). 'n' refers to the total number of cells in that time point, and the percentage refers to the proportion of cells within RT clusters. **e-i**. Time point-specific UMAP visualizations for each case. RT seed cells are depicted in rose and with an increased size. **j**. UMAP visualization of case CLL9 from Penter et al.[43] split by time point. PB, peripheral blood; BM, bone marrow.



Extended Data Fig. 8 | CNA profile of RT cells by scRNA-seq. For each patient, the CNA profile of CLL and RT samples according to WGS is shown (top) together with the CNA profile of each individual RT cell based on scRNA-seq (bottom). For scRNA-seq, each row represents a RT cell and the horizontal dashed line separates the RT cells identified in the time points previous to the diagnosis of RT (that is, seed RT cells) from those present in the sample collected at time of diagnosis of RT. Note that CLL cells were used as reference for CNA analyses using scRNA-seq data.



Extended Data Fig. 9 | Epigenomic characterization of RT. a. Heatmaps showing the regions with decreased H3K27ac, increased ATAC, and decreased ATAC levels, respectively, in RT. **b**. Overlap of differentially expressed genes by bulk RNA-seq with regions with increased or decreased H3K27ac and ATAC levels, respectively. **c**. Heat map showing differentially methylated CpGs (DMC) between CLL and RT. Normal B cells, CLL, CLL at relapse, and RT samples are shown separately with different biological information on top. The correlation of each CpG with the epiCMIT is depicted on the right. To note, the epiCMIT is associated with the gain and loss of methylation upon cell division, but its transformation to 0-1 scale (for interpretability purposes) makes it anticorrelated with hypomethylation, as the epiCMIT=max{epiCMIT-hyper, epiCMIT-hypo}, being the epiCMIT-hyper=hypermethylation, and the epiCMIT-hypo=1-hypomethyaltion at relevant CpGs, as originally reported[49]. **d**. Genomic enrichment over the background for hyper- and hypomethylated CpGs in CLL vs RT. **e**. DMC distribution based on their genetic annotation and their intersection with differentially expressed genes by bulk RNA-seq analyses. **f**. DMC distribution based on the correlation of each CpG with the epiCMIT and their p values. CpGs were piled up in color-coded bins based on the number of CpGs in each bin to avoid overplotting. **g**. epiCMIT evolution in longitudinal CLL and RT samples, with the epiCMIT-hyper and epiCMIT-hypo scores depicted separately (RT samples being the last time point labeled in rose). The epiCMIT score used to compare among samples is the greater of the two (hyper and hypo).



Extended Data Fig. 10 | Transcriptomic characterization of RT. a. Volcano plot of the differential expression analysis (RT vs CLL, bulk RNA-seq). **b**. Expression levels of selected genes in CLL and RT according to bulk RNA-seq. center line, median; box limits, upper/lower quartiles; whiskers, 1.5xinterquartile range; points, individual samples. **c**. Differentially expressed genes (RT vs CLL) for each case by scRNA-seq. **d**. GSEA plots of selected hallmark gene sets according to bulk RNA-seq analyses. NES, normalized enrichment score. **e**. UpSet plots highlighting the intersections of the case-specific upregulated (top) and downregulated (bottom) GO terms in RT by scRNA-seq. **f**. GSEA plots for the terms oxidative phosphorylation (OXPHOS), mitochondrial translation, and BCR signaling pathway for cases 12, 63, and 365 based on scRNA-seq. **g-j**. scRNA-seq-derived UMAP visualization of tumor cells from all time points colored by OXPHOS and BCR signaling score (left). Ridge plots showing the same scores across clusters (middle). Violin plots displaying the same scores across time points, stratified by CLL and RT clusters (right). **k**. Violin plots displaying the OXPHOS and BCR signaling scores across time points, stratified by CLL and RT clusters, in case CLL9 from Penter et al[43]. **l**. GSEA between RT and CLL cells of patient CLL9 from Penter et al.[43].

CORRESPONDENCE OPEN

Check for updates

ATM germline variants in a young adult with chronic lymphocytic leukemia: 8 years of genomic evolution

Blood Cancer Journal (2022)12:90 ; https://doi.org/10.1038/s41408-022-00686-6

Chronic lymphocytic leukemia (CLL) is a disease commonly diagnosed in the elderly with a median age of ~70 years. However, CLL can also be detected in adolescent and young adults (AYA). According to different studies, 0.85–3.7% of patients with CLL are diagnosed in AYA and 3% of these patients had a first-degree relative with CLL [1]. Families with multiple individuals affected with CLL and other related B-cell tumors have been described with contradictory findings regarding their potential early age at diagnosis [2]. Despite these observations, our knowledge about the molecular profile and predisposing factors in AYA CLL is scarce [3, 4].

Comprehensive studies have dissected the (epi)genomic, and transcriptomic landscape of CLL [5]. Approximately 9–18% of CLL harbor del(11q) which occurs in younger patients with bulky disease and poor survival. These deletions are frequently associated with germline and acquired mutations of *ATM* [6]. Patients with the inherited disorder ataxia telangiectasia have biallelic alterations of the *ATM* gene and increased susceptibility to lymphoid malignancies [7]. Rare, protein-coding germline *ATM* variants are associated with CLL in adults [8]. However, *ATM* mutations are uncommon in familial CLL [9].

Here, we describe an 18-year-old woman diagnosed with CLL whose family history included a younger brother with B-cell acute lymphoblastic leukemia (B-ALL) and other family members carrying germline *ATM* mutations. A combination of whole-genome and single-cell characterization of this CLL at diagnosis and during the course of the disease provided an opportunity to understand the genomic profile of AYA CLL and the sequence of events driving its evolution.

An 18-year-old female was diagnosed with CLL, Binet-Rai stage AI, at another institution, in the study of a lymphocytosis detected in a routine blood test. She had a past medical history of anxiety-depressive syndrome during childhood and chronic headache, but no neurological symptoms were reported. The patient had a younger brother diagnosed with B-ALL when he was 3 years old, and was in complete remission 13 years later, and an older sister with epilepsy. Her parents were both healthy.

At the time of CLL diagnosis, the patient was asymptomatic with a normal physical exam. Her white blood cell count (WBC) was 9.08×10^9/L, with 75% lymphocytes. Hemoglobin and platelet count were normal. Peripheral blood smear showed small atypical lymphocytes consistent with CLL, which phenotype was $CD5^+$, $CD23^+$, $CD43^+$, $CD200^+$, $CD10^-$, CD20 and CD22 weakly positive with weak kappa light chain restriction. The fluorescence in situ hybridization (FISH) analysis for *ATM* (11q22), *D12Z3* (cen 12), *DLEU* (13q14.3), *LAMP1* (13q34), and *TP53* (17p13) were normal. One year after diagnosis, the patient received two cycles of rituximab plus fludarabine and cyclophosphamide (FCR) due to progressive disease, achieving a complete remission. The patient was then referred to our hospital. Physical examination was normal without evidence of lymphadenopathy or splenomegaly. WBC count was 2×10^9/L with 10% lymphocytes, hemoglobin 117 g/L, and normal platelet count. Watchful waiting was recommended. Five years later, the CLL progressed with increased lymphocytosis, inguinal, axillary, and laterocervical lymphadenopathy (2–3 cm) and splenomegaly of 4 cm below the costal margin. At that time, the karyotype was 46,XX,del[13](q12q21)[6]/46,XX[10] and a heterozygous del(13q14.3) was detected by FISH in 92% of nuclei. FISH for *ATM*, *D12Z3*, and *TP53* were normal and no *TP53* mutations were observed. The sequence of the IGHV genes showed a clonal rearrangement of the IGHV3-21 with 100% homology to the germline, not belonging to any major stereotype subset (Supplementary Tables 1, 2). Due to CLL progression, ibrutinib 420 mg per day was started and the patient achieved a partial response. However, after 20 months, ibrutinib had to be discontinued due to the severe diarrhea and acalabrutinib 100 mg every 12 h was started. Progression of CLL was observed after 13 months of treatment and rituximab and venetoclax were initiated (Fig. 1A).

The patient was included in the CLL program of the International Cancer Genome Consortium and the whole genomes of the germline and tumor sample at diagnosis were sequenced [5]. No somatically-acquired driver alterations were detected but three germline *ATM* mutations were identified, including a pathogenic 28-base frameshift deletion (p.N3003Dfs*6) and two missense single nucleotide variants (p.K2204M and p.Y1961C). Although the p.K2204M missense variant has not been identified in previous studies, the p.Y1961C has been reported in a CLL patient and its modeling showed reduced ATM kinase activity [10]. Based on this result, we studied the segregation of these mutations in the family members by Sanger sequencing. The mother harbored the frameshift deletion, while the father and the sister carried the two missense variants. Both the patient and her brother with B-ALL inherited all three variants (Fig. 1B, Supplementary Tables 3, 4). A milder ataxia telangiectasia phenotype, where the disease progresses at a slower pace, has been observed in patients with reduced levels of ATM kinase activity [11]. At time of last follow-up the two siblings (28 and 16 years old) had not developed neurological symptoms.

To better unfold the contribution of somatic alterations during the evolution of the disease, whole-genome sequencing (WGS) was performed at 3 additional time points over a period of 8 years and complemented with single-cell DNA-sequencing (Fig. 1A, Supplementary Table 1). Using a longitudinal sample-aware mutation calling pipeline that increases sensitivity, we identified 689 genome-wide and 7 non-synonymous variants in the WGS at diagnosis, increasing up to 1779 genome-wide and 18 non-synonymous at the latest sample analyzed. Among them, four mutations were found in CLL driver genes over the course of the

Received: 11 May 2022 Revised: 13 May 2022 Accepted: 25 May 2022
Published online: 07 June 2022

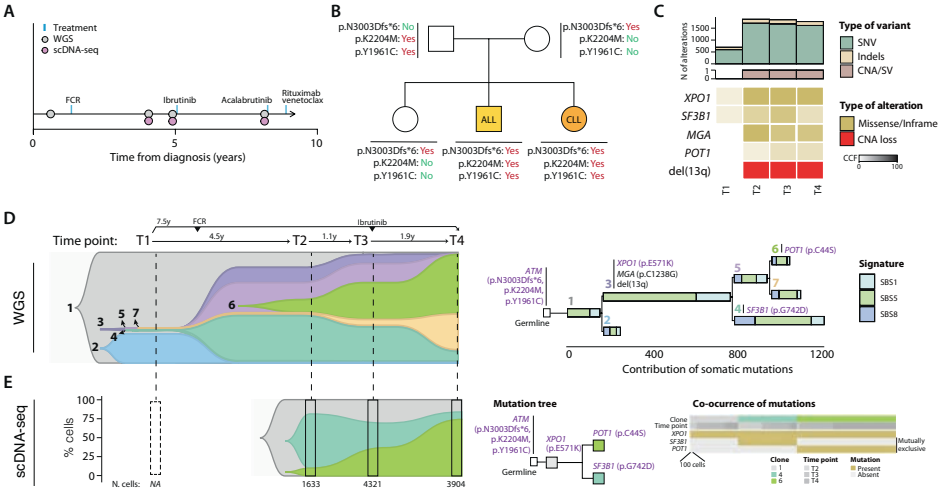


Fig. 1 Clinical course and genomic characterization. A Clinical course and samples analyzed. **B** Pedigree tree of germline variants in *ATM*. The two missense variants carried by the mother and the frameshift variant from the father were inherited by the chronic lymphocytic leukemia (CLL) case studied and her brother that developed acute lymphoblastic leukemia (ALL). **C** The upper barplots show the number of mutations [single nucleotide variants (SNV) and short insertions and deletions (indels)] and copy number alterations (CNA) or structural variants (SV) at each time point. The lower oncoprint shows the driver alterations, the transparency of the color is proportional to the cancer cell fraction (CCF). **D** The fishplot [left] depicts the subclonal architecture and clonal dynamics inferred from WGS. Each vertical line represents a time point analyzed. Each subclone is painted in a different color, and its height is proportional to the CCF at each time point. The upper-right tree shows the phylogeny of the tumor cell subpopulations, the length of the branches is proportional to the number of acquired SNV, and they are colored by contribution of mutational signatures identified in CLL [right]. The clock-like signatures SBS1 and SBS5 contributed most of the mutations acquired. **E** The fishplot (left) shows the clonal dynamics measured by single-cell analysis. For each available time point, the integrated barplot shows the proportion of cells harboring each specific combination of alterations in the driver genes illustrated on the "Mutation tree" (middle). The total number of analyzed cells at each analyzed sample is shown at the bottom. The "Co-occurrence of mutations" plot (right) indicates the presence or absence of mutations in each cell. For illustrative purposes, cells have been merged in bins of 100.

disease: *XPO1* (p.E571K), *SF3B1* (p.G742D), *MGA* (p.C1238G), and *POT1* (p.C44S). The mutations in *XPO1* and *SF3B1* were already present at diagnosis but were missed in our previous study [5] due to their very low frequencies. After 4 years (time point 2), their clonal size expanded, and the remaining two driver mutations in *MGA* and *POT1* were detected. Regarding structural alterations, only del(13q) was clonally detected at the second time point and onwards (Fig. 1C, Supplementary Methods, Supplementary Tables 5, 8).

Somatic driver alterations were present at different allele frequencies through the disease course, suggesting an ongoing clonal evolution driving the pre- and post-treatment progression of the disease. To dissect the underlying clonal evolution, we reconstructed the subclonal evolution and explored the mutational processes active during the CLL course (Fig. 1D, Supplementary Methods, Supplementary Tables 9, 10). This analysis revealed a branching pattern of evolution in which the founding CLL clone did not carry any recognized driver alteration beyond the *ATM* germline variants. Additionally, two minor subclones were already present at diagnosis: subclone #3 carrying del(13q), *XPO1* and *MGA*, and subclone #4 which originated from subclone #3 and acquired the *SF3B1* mutation (Fig. 1D). These lineage trajectories are in line with previous literature in which *ATM* loss preceded del(13q) in a familial CLL study [12] and with a recently described combinatorial effect of *ATM* loss and *SF3B1* mutation [13]. Intriguingly, these small subclones at diagnosis expanded after treatment with FCR, that, on the other hand, reduced or eliminated the initial subclones #1 and #2, with no additional CLL drivers, suggesting that decreased competition allowed the expansion of subclones carrying potent drivers. Of note, subclone #4 carrying the *SF3B1* mutation represented the largest subpopulation of cells at relapse post-treatment with FCR (time point 2), in line with the poor prognosis of *SF3B1* mutated cases under FCR therapy [14]. Nonetheless, this subclone slightly diminished at time point 3 and was virtually eradicated at time point 4 after treatment with ibrutinib, which is in line with the higher sensitivity of *SF3B1* mutated CLL cells to BCR inhibition in vitro [13]. Additional diversification was observed in subclone #3 at time point 2 which led to the emergence of subclone #6 harboring the *POT1* mutation. This subclone expanded under ibrutinib treatment and accounted for 54% at the last time point analyzed 3 years after its detection (Fig. 1D). To confirm these evolutionary trajectories, we performed single-cell DNA-sequencing of 32 CLL driver genes and identified the reported mutations in *XPO1*, *SF3B1*, and *POT1* [note that *MGA* was not included in the commercial gene panel used]. This single-cell analysis confirmed the timing of acquisition of these driver mutations and the clonal dynamics inferred from WGS (Fig. 1E, Supplementary Methods, Supplementary Tables 11, 14).

Here we have reported the 8-year genomic evolution of a CLL diagnosed in a young patient that inherited three *ATM* variants, two of them previously reported to inactivate or reduce *ATM* activity (Supplementary Table 4) [10]. The combination of these three germline *ATM* variants predisposed to two distinct B-cell neoplasm in two siblings. These *ATM* variants represented the only recognized driver events in the founding CLL clone, suggesting that *ATM* inactivation might be a genomic factor contributing to CLL initiation. Tumor evolution and disease progression was dictated by the acquisition of secondary driver alterations, which could be detected in small subclones years before their expansion,

and by different types of treatment that influenced subsequent clonal dynamics. Of note, this patient responded well to initial FCR therapy and later to ibrutinib treatment when *ATM* inactivation was accompanied by an *SF3B1* mutation, which is in line with the favorable clinical behavior of del(11q) CLL under BTK inhibitors [15]. Altogether, the lack of somatically-acquired, genetic driver alterations in the founding CLL of this patient emphasizes the need to study the germline as well as non-genetic aspects of the tumors to further understand the mechanisms leading to CLL.

ACCESSION NUMBER

The WGS and single-cell DNA-sequencing data have been deposited to the European Genome-phenome Archive (EGA) under the accession code EGAS00001006268.

Romina Royo[1,10], Laura Magnano[2,3,4,10], Julio Delgado[2,3,4,5], Sara Ruiz-Gil[6], Josep Ll. Gelpi[1,5], Holger Heyn[6,7], Malcom A. Taylor[8], Tatjana Stankovic[8], Xose S. Puente[3,9], Ferran Nadeu[2,3,11] and Elías Campo[2,3,4,5,11] ✉

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [2]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [3]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. [4]Hospital Clínic de Barcelona, Barcelona, Spain. [5]Universitat de Barcelona, Barcelona, Spain. [6]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [7]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [8]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, UK. [9]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. [10]These authors contributed equally: Romina Royo, Laura Magnano. [11]These authors jointly supervised this work: Ferran Nadeu, Elías Campo.
✉*email: ecampo@clinic.cat*

REFERENCES

1. Cherng HJ, Jammal N, Paul S, Wang X, Sasaki K, Thompson P, et al. Clinical and molecular characteristics and treatment patterns of adolescent and young adult patients with chronic lymphocytic leukaemia. Br J Haematol. 2021;194:61–68.
2. Goldin LR, Bjorkholm M, Kristinsson SY, Turesson I, Landgren O. Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. Haematologica. 2009;94:647–53.
3. Luskin M, Wertheim G, Morrissette J, Daber R, Biegel J, Wilmoth D, et al. CLL/SLL diagnosed in an adolescent. Pediatr Blood Cancer. 2014;61:1107–10.
4. Nassereddine S, Dunleavy K. A case of chronic lymphocytic leukemia in an AYA patient. Clin Lymphoma Myeloma Leuk. 2019;19:S280.
5. 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;526:519–24.
6. Skowronska A, Austen B, Powell JE, Weston V, Oscier DG, Dyer MJS, et al. ATM germline heterozygosity does not play a role in chronic lymphocytic leukemia initiation but influences rapid disease progression through loss of the remaining ATM allele. Haematologica. 2012;97:142–6.
7. Reiman A, Srinivasan V, Barone G, Last JI, Wootton LL, Davies EG, et al. Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. Br J Cancer. 2011;105:586–91.
8. Tiao G, Improgo MR, Kasar S, Poh W, Kamburov A, Landau DA, et al. Rare germline variants in ATM are associated with chronic lymphocytic leukemia. Leukemia. 2017;31:2244–7.
9. Yuille MR, Condie A, Hudson CD, Bradshaw PS, Stone EM, Matutes E, et al. ATM mutations are rare in familial chronic lymphocytic leukemia. Blood. 2002;100:603–9.
10. Barone G, Groom A, Reiman A, Srinivasan V, Byrd PJ, Taylor AMR. Modeling ATM mutant proteins from missense changes confirms retained kinase activity. Hum Mutat. 2009;30:1222–30.
11. Stewart GS, Last JIK, Stankovic T, Haites N, Kidd AMJ, Byrd PJ, et al. Residual ataxia telangiectasia mutated protein function in cells from ataxia telangiectasia patients, with 5762ins137 and 7271T→G mutations, showing a less severe phenotype. J Biol Chem. 2001;276:30133–41.
12. Kostopoulos IV, Tsakiridou AA, Pavlidis D, Megalakaki A, Papadhimitriou SI. Familial chronic lymphocytic leukemia in two siblings with ATM/13q14 deletion and a similar pattern of clonal evolution. Blood Cancer J. 2015;5:e322–e322.
13. Yin S, Gambe RG, Sun J, Martinez AZ, Cartun ZJ, Regis FFD, et al. A murine model of chronic lymphocytic leukemia based on B cell-restricted expression of Sf3b1 mutation and Atm deletion. Cancer Cell. 2019;35:283–296.e5.
14. Stilgenbauer S, Schnaiter A, Paschka P, Zenz T, Rossi M, Döhner K, et al. Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. Blood. 2014;123:3247–54.
15. Kipps TJ, Hillmen P, Demirkan F, Grosicki S, Coutre SE, Barrientos JC, et al. 11q deletion (del11q) is not a prognostic factor for adverse outcomes for patients with chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL) treated with ibrutinib: pooled data from 3 randomized phase 3 studies. Blood. 2016;128:2042–2042.

ACKNOWLEDGEMENTS

The authors thank the Hematopathology Collection registered at the Biobank of Hospital Clínic—Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) as well as Sílvia Martín for the technical support. This study was supported by the "la Caixa" Foundation (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221, to EC), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (810287, BCLLatlas, to EC, and HH), CERCA Program/Generalitat de Catalunya, Generalitat de Catalunya Suport Grups de Recerca AGAUR 2017-SGR-1142 (to EC), CIBERONC (CB16/12/00225 to EC), Ministerio de Ciencia e Innovación PID2020-117185RB-I00 (to XSP), FEDER: European Regional Development Fund "Una manera de hacer Europa", and Fundación Asociación Española Contra el Cáncer FUNCAR-PRYGN211258SUÁR (to XSP). The authors thankfully acknowledge the computer resources at MareNostrum4 and the technical support provided by Barcelona Supercomputing Center (RES activity BCV-2018-3-0001). FN acknowledge research support from the American Association for Cancer Research (2021 AACR-Amgen Fellowship in Clinical/Translational Cancer Research, Grant Number 21-40-11-NADE), the European Hematology Association (EHA Junior Research Grant 2021, Grant Number RG-202012-00245), and the Lady Tata Memorial Trust (International Award for Research in Leukemia 2021–2022, Grant Number LADY_TATA_21_3223). EC is an Academia Researcher of the "Institució Catalana de Recerca i Estudis Avançats" (ICREA) of the Generalitat de Catalunya. This work was partially developed at the Centre Esther Koplowitz (CEK, Barcelona, Spain).

AUTHOR CONTRIBUTIONS

RR collected data, analyzed data, and wrote the manuscript. LM and JD collected samples and clinical data, and wrote the manuscript. SR-G and HH performed single-cell experiments. MAT and TS interpreted data. JLIG and XSP analyzed and interpreted data. FN and EC designed the study, collected and analyzed data, wrote the manuscript, and supervised the research. All authors reviewed and approved the manuscript.

COMPETING INTERESTS

HH is co-founder of Omniscope and consultant to MiRXES. XSP is co-founder of and holds an equity stake in DREAMgenics. FN has received honoraria from Janssen for speaking at educational activities. EC 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

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41408-022-00686-6.

Correspondence and requests for materials should be addressed to Elías Campo.

Reprints and permission information is available at http://www.nature.com/reprints

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