The evolution of T-cell acute lymphoblastic leukemia in adult patients under treatment

Inés Sentís Carreras

Directors de tesi

Dra. Núria López-Bigas i Dr. Abel González-Pérez

TESI DOCTORAL UPF / 2020

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCE

Universitat Pompeu Fabra Barcelona

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

Acknowledgments

First of all, I would like to thank Dr. Núria López-Bigas for giving me the opportunity to do my PhD in her laboratory. Secondly, I also want to acknowledge her and Dr. Abel González-Pérez for the guidance, supervision and support along this 4-year journey. I want to extend this to all the people in the lab that contributed to the investigation and helped me complete my PhD. First, to Dr. Santi González for his patience and coaching since the very first moment he jumped in the project and for being a great colleague. Also to Dr. Ferran Muiños and Dr. Erika López-Arribillaga who helped me in anything they could. In addition, I want to thank the collaborators of the project; the lab of Dr. Anna Bigas and the lab of Dr. Josep Maria Ribera and all the patients that, in the most difficult times, agreed to collaborate in science.

I want to thank my lab PhD mates for being such good friends. Oriol for always listening to my worries and giving me such good advice, Pepe for being supportive (even in the distance) and keeping the joking environment, Hania for being such a caring person and for always suggesting great music and Claudia for always being so thoughtful and cheerful to all of us. I am also very grateful to my friends and former lab colleagues, Dr. Joan Frigola and Dr. Carlota Rubio-Pérez with whom I have shared great bonding beer times and for always encouraging parties. I hope it will continue like that so I can finally make it to the party ;)

I greatly appreciate the support and help from all the people in the lab (current mates and former ones). To Jordi, Iker and Loris for their patience and to help me improve as a bioinformatician. To the rest of the great post-docs that I have encountered along the way: Fran, Sabari and David. To the rest of the members of the wet lab not mentioned already: Victor, Núria, Mari, Ramón and Meritxell. Also to the talented master students: Winona, Inés, Laia, Carla and Morena and other students and visitors of the lab. Also, special thanks to Martina who helped me get all the project and uni paperwork done on time. Finally, to the recent members of the BBGLab.

También quiero dar las gracias a los míos. Primero quiero empezar por mis padres. A mi padre Ernest Sentís y a mi madre la Dra. Pilar Carreras ;) por apoyar mis decisiones, cuidar de mí siempre y estar orgullosos de mi haga lo que haga. Sin duda los mejores padres que una puede desear. Al amor de mi vida, mi marido Carlos, por empujarme a ser valiente y creer en mí siempre. Si he llegado hasta aquí es gracias a él. A mis hermanas Maria, Helena y Cris porque son pilares esenciales de mi vida que me han escuchado, aconsejado y animado desde siempre. A mis abuelos Mª Pilar, Ramón, Gloria y Albert por ser, con sus palabras, mis mayores "fans". También a mis primos, sobrinos, tíos, cuñados y suegros que forman parte de lo más importante en la vida: mi familia.

Por último quiero agradecer el apoyo de mis amigos. A mis amigas del cole con las que he reído, llorado y que hacen el esfuerzo de intentar entender mi investigación. También agradecer el apoyo de mis amigos y amigas de la Cerdanya, por todos los momentos vividos

y por su capacidad de hacerme desconectar de mis preocupaciones cada vez que nos vemos. Me siento muy afortunada de tenerlos a todos ellos en mi vida.

Abstract

Acute lymphoblastic leukemia (ALL) is a blood cancer characterized by a high proliferation and maturation arrest of the lymphoid precursors which can either be from B or T-cell lineage. In adult patients, this type of cancer is considered a rare disease and the outcome is worse than children, especially for those presenting the T-Cell ALL (T-ALL) type. In order to get insights on the evolution of adult T-ALL under therapy, we have whole genome sequenced leukemic samples at diagnosis and relapse of 19 adult patients with T-ALL who relapsed after standard treatment. We report the somatic driver alterations and active mutational process and compared them to other ALL cohorts. We pinpoint candidates of therapy resistance by looking at relapse-enriched alterations (e.g. genes NT5C2, ABCB1 and SMARCA4). In most cases, the relapse clone is estimated to diverge from the primary the previous year to the diagnosis, by which time, the relapse-fated subpopulation size ranges from few to millions of cells. We have also simulated different scenarios of primary and relapse leukemias and concluded that the relapsed leukemias of the sequenced cohort are driven by genetic resistance. In this project we provide an integrated vision of the mutational evolution of T-ALL adult cases and highlight the relevance of finding cancer driver genes of resistance. In line with that, we have also generated a compendium of mutational cancer driver genes across different cancer types through the analysis of thousands of tumors with a whole new framework for driver gene discovery (IntOGen).

Resum

La leucèmia limfoblàstica aguda (LLA) és un càncer de sang que es caracteritza per una altra proliferació i arrest en la maduració dels precursors limfoblàstics que poden ser del llinatge B o T. En pacients adults, aquest tipus de càncer és considerat una malaltia rara i presenten pitjor pronòstic que els pacients pediàtrics en especial en aquells adults del tipus T-LLA. Per tal de conèixer millor l'evolució de la T-LLA en adults en tractament, hem seqüenciat el genoma sencer de mostres a diagnòstic i recaiguda de 19 pacients adults amb T-LLA que van recaure després de rebre el tractament estàndard. Reportem les alteracions somàtiques *driver* i els processos mutationals actius en comparació amb d'altres cohorts de LLA. També assenyalem candidats de resistència al tractament tot mirant les alteracions abundants en recaiguda (per exemple als gens NT5C2, ABCB1 i SMARCA4). En la majoria dels casos, el clon de recaiguda s'estima que va divergir del clon primari l'any previ a la diagnosi, moment pel qual, les cèl·lules destinades a fer la recurrència constitueixen una subpoblació cel·lular que va de poques a milions de cèl·lules. Mitjançant simulacions de diferents escenaris de leucèmies primàries i de recaiguda, concloem que les leucèmies de recaiguda d'aquesta cohort seqüenciada es deuen a una resistència genètica. En aquest projecte donem una visió integrada de l' evolució mutacional de les T-LLA en casos adults i resaltem la rellevància de trobar gens *driver* de resistència. En aquesta línia, també hem generat un compendi de gens *driver* mutacionals de diferents tipus càncer a través de l'anàlisi de milers de tumors amb una nova plataforma de detecció de gens *driver* (IntOGen).

Table of contents

1. INTRODUCTION

1.1 Cancer is an evolutionary process

Cancer is a term that comes from the greek word for crab «karkinos» and comprises a set of diseases that present abnormal cells that uncontrollably divide and invade the proximal tissues and/or spread to other parts of the body [1,2]. It presents a high heterogeneity among its different forms (more than 100 different ones) with particular risk factors and epidemiology [3]. However, global numbers point out that cancer is the second leading cause of death worldwide and approximately one third of the cancer deceases are due to the following risk factors: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco and alcohol use [4].

In general, cancer is also defined as a genetic disease as it is caused by changes in the genome that triggers the loss of division and growth control of the cells. The genome is formed by deoxyribonucleic acid (DNA) which is a molecule in the shape of a double helix of polynucleotide[1] chains (strands). Alterations can appear at different levels of the genome: from the sequence of a gene[2] to each one of the packing levels of the DNA (see Figure 1). The necessary information for the cell to develop, function, grow and reproduce is stored in this molecule in the cellular nucleus. Unrepaired alterations from damaged DNA can cause cancer if it affects specific cellular functions that lead to abnormal division of the cells and the

[1] The monomeric units of DNA are called nucleotides which are formed by a desoxyribose, a phosphate group and a nitrogenous base. There are 4 possible nitrogenous bases of two types: pyrimidines (thymine or T and cytosine or C) and purines (guanine or G and adenine or A; see Figure 1).
[2] Def. gene: DNA sequence of fixed position (locus) with the basic physical unit of inheritance.

formation of a tumoral mass or neoplasm. The acquisition of the "malignancy" of the cancerous cell is a progressive inner-process called tumorigenesis or carcinogenesis.

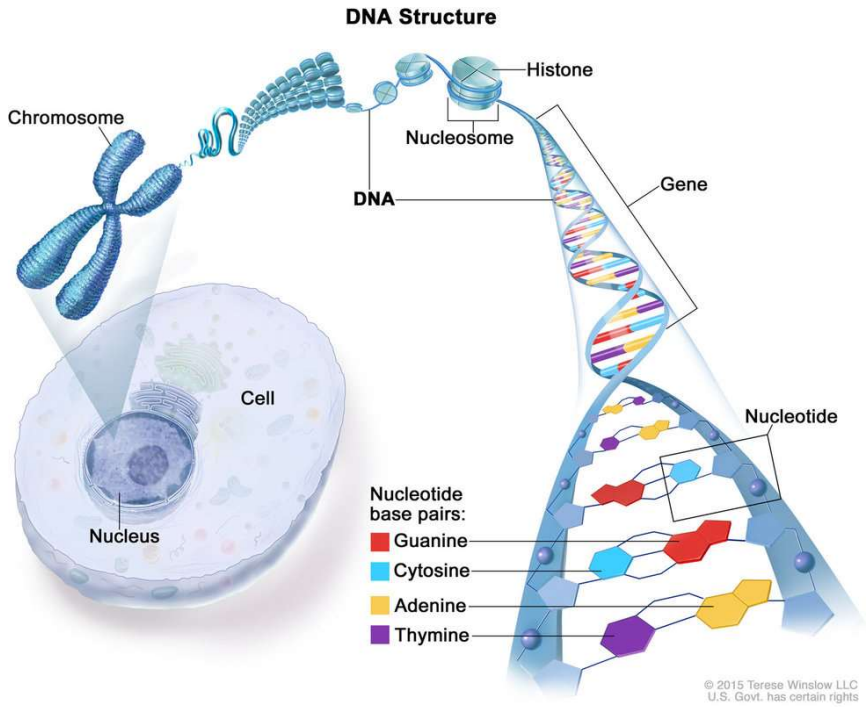


Figure 1. Illustration by Terese Winslow. DNA structure consists of different levels of DNA wrapping and packing.

Our knowledge of this disease has grown as a result of the advances in genetics. The first notorious approaches towards the comprehension of cancer disease started in the late nineteenth century [5]. David von Hansemann who was a pathologist, noticed that some tumor cells presented multipolar mitosis which resulted in abnormal chromosomal numbers in daughter cells [6,7]. Contemporary to his work, the german zoologist Theodor Boveri, observed an unequal number of chromosome distribution in the daughter cells of a double-sperm fertilized sea-urchin eggs [8]. The phenotypic differences between the chromosomal imbalanced daughter cells drove him to the hypothesis that cancer is a cellular disease. He

proposed that tumors also arise as a consequence of abnormal segregation of chromosomes that provide the capacity of unrestricted growth to the daughter cells [9]. These german scientists were the first ones to link aberrant chromosomal distributions of daughter cells to the aetiology of cancer.

In the following decades, experimental advances were made in the field of chemical carcinogenesis which allowed the connection between exposure to chemical agents and the cause of cancer in humans [10].

In the 70s and 80s the first cancer-causing genes were described. Bishop and Varmus observed how normal avian cells turned malignant with the presence of transferred Rous Sarcoma Virus (*src*) sequences [11]. They discovered that the *src* sequences were already present in the normal avian genome thus realizing that viral cancer-causing genes were altered sequences of already existing genes of the normal cells. This finding introduced the concept that cancer might emerge from mutated versions of genes. A few years later, this was consolidated by the description of a single mutation at codon 12 which was able to activate the oncogenicity of *HRAS* gene human bladder cancer [12].

1.1.1 Hallmarks and ecological features of cancer

The somatic mutation theory [13] (SMT) presents tumorigenesis in humans as a multi-step process in which the accumulation of defects in the regulatory circuits that rule the normal cell disrupt their homeostasis to transform it into its malignant counterpart [14]. There are multiple ways for a cell to acquire a cancerous state. Unfortunately, this implies that cancer is a complex disease in which each patient has a unique tumoral manifestation of it (analogously coined as “malignant snowflakes” since ultimately they

are all different [15]). Paradoxically, zooming out of its complexity there are six well-defined traits that characterize malignant cells called the "Hallmarks of cancer" by Hanahan and Weinberg [14,16] (see Figure 2).

1. Evading growth suppressors

   This means to gain insensitivity to growth-inhibitory (antigrowth) signals. This hallmark reunites tumor suppressor gene discoveries. Two of the most notorious examples are RB protein which controls cell-cycle progression and TP53 which acts as a sensor of aberrant cell functionality and can halt cell-cycle or even trigger apoptosis if needed.

2. Sustaining proliferative signaling

   There are many ways in which a cell can maintain its proliferative capacity. Cancer cells can stimulate their surrounding environment into the production of proliferative ligands. Another possibility is the autocrine way in which tumoral cells produce growth factors and the corresponding receptors themselves. Furthermore, cells can just increase the expression of growth factor receptors or modify their structure to make them active and ligand-independent. Alternatively, a cell can become ligand-independent by acquisition of mutations in downstream effectors of proliferation related pathways. For example, activating mutations in Ras protein break the intrinsic negative feedback-loop that regulates the Ras GTPase activity. Another example are loss-of-function PTEN mutations that prevents phosphatidylinositol (3,4,5) trisphosphate (PIP3) from degradation and constitutively PI3K signaling activates PI3K proliferative signaling.

3. Activating invasion and metastasis of the tissues

One of the most studied cancer dissemination processes is the "epithelial-mesenchymal transition" (EMT) mechanism. This is a developmental regulatory program that apart from being involved in embryonic morphology it also acts in the transformation towards malignancy of epithelial cells in cancer. The transcriptional factors such as Snail, Slug, Twist and Zeb1/2 are the players of this process in which they modify the cell into an invasive phenotype by making it matrix in-adherent, creating a fibroblastic morphology, increasing its capacity for motility and resistance to apoptosis. Another important element related to this hallmark, is the disruption of normal signaling between cancer cells and the surrounding stromal cells. For example, in some occasions, macrophages can supply with metalloproteases and cysteine cathepsin proteases to degrade the matrix bindings and promote cell invasion. Apart from EMT, there are other forms of dispersion described like nodules of cancer cells invading in mass ("collective invasion") or cancer cells acquiring ameboid motility to slight thought the tissue.

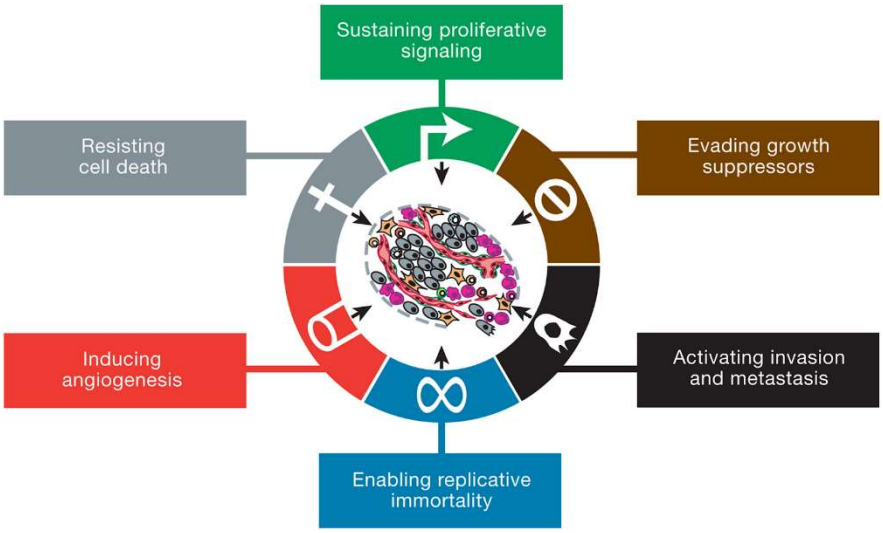


Figure 2. The Hallmarks of Cancer reprinted from *Hallmarks of cancer: The next generation* (v.144, p.647) by Hanahan and Weinberg, 2011 Cell.

4. Enabling replicative immortality

That is limitless replicative potential which is acquired by overcoming two proliferative barriers: senescence and crisis (cell death). The first term means that cells enter a nonproliferative but viable state and the second one means overcoming an apoptotic crisis phase of senescent cell population and becoming immortal. Most of the cell lines used in cancer research are immortalized and by studying them it has been discovered that telomeres (the protective ends of chromosomes) play a central role in this enduring cell state. Telomeres are eroded along the cell descendant generations giving a finite replicative potential to it. In fact, most of the immortalized cells have functional levels of expression of telomerase (the specific DNA polymerase that adds the repetitive telomeric segments). Apart from this function, there are evidences of other involvements of telomerase in tumorigenesis that are

telomere-independent. The protein subunit TERT has been associated with cell proliferation, apoptotic resistance and DNA-damage repair involvement.

5. Inducing angiogenesis and sustain it

   Angiogenesis is the physiologic process of creating new blood vessels from existing ones. It is very active during embryogenesis and transiently activated in adults in concrete processes like wound healing or female reproductive cycle. The purpose of angiogenesis is to ensure good tissue irrigation so that cell metabolic exchange of nutrients and waste is guaranteed. In order to sustain highly demanding neoplastic growth, it has been observed that there is an angiogenic switch during tumor progression. However, neovascularization patterns in tumors is highly variable among tumor types, some of them being hypovascularized (e.g. pancreatic ductal adenocarcinomas) and others hypervascularized (e.g. pancreatic neuroendocrine carcinomas). Some immune innate system[3] cells are associated with the angiogenic switch contributing to tumor growth and local invasion.

6. Resisting cell death

   Apoptosis or "cell suicide" is a programmed cell death that serves as a normal mechanism to eliminate damaged cells and maintain tissue homeostasis. When it is impaired, there is a loss of control of cell proliferation and, therefore, it contributes to tumor progress. There are two circuits to trigger apoptosis, one cell extrinsic and

[3] Immune innate system: immunity mechanism against pathogens. It is phylogenetically conserved among multicellular organisms. Cell members of it are macrophages, neutrophils, mast cells, and myeloid progenitors.

the other with cell intrinsic origin. The first one starts by the activation of an external receptor (tumor necrosis factor receptor superfamily) and the intrinsic one is normally triggered by inner cell stress (e.g DNA damage). Both stimulate the caspase enzymes (caspase 8 and 9 respectively) which provokes the interaction of those with apoptotic inhibitors and the Bcl-2 family members which some are pro- and some anti-apoptotic regulators. This ends up lising the outer membrane of the mitochondria and releasing cytocrom c which in turn activates other caspases that initiates proteolytic activities to induce disassembly of the cell. One way to resist apoptosis is by the overexpression of anti-apoptotic proteins like Survivin or Bcl-2. On the contrary, deactivating mutations in pro-apoptotic regulators also contributes to inhibit apoptosis. Another notorious example is the loss of function of TP53 which implies a disruption of a critical damage sensor within the intrinsic apoptotic circuit.

The 6 original hallmarks were defined in 2000 and were embraced as a research guidance by the scientific community. However, as some pointed out some years later [17], there is not much difference between benign tumor mass and a malignant one in terms of the the features described in the seminal paper by Hanahan and Weinberg meaning that except for the tissue invasion and metastasis hallmark, most of them are shared between the two. In 2011 the hallmarks were re-defined and updated to a past decade of research. In addition, a new approach was presented with the distinction of two concepts surrounding cancer: “enabling characteristics" and “hallmarks of cancer”. The acquisition of hallmarks is possible by two consequential characteristics of neoplasias (i.e. enabling characteristics): genome instability and mutability and tumor promoting-inflammation. The first one refers to the successive acquisition of genomic alterations that

provide the characteristics of the hallmarks to the cells. Examples of these alterations are gains and losses of copy number and/or genome rearrangements favoring dysregulation of cell homeostasis or inactivating mutations in key players of genome integrity maintenance. This settles a wide mutational space for the cell to explore and acquire favorably mutagenic genotypes.

The second enabling characteristic implies the observed infiltration of innate and adaptive immune system cell members in tumors. Inflammation associated to the immune response can foster tumor progression and contribute to the hallmark capabilities of cancer since it releases signaling factors for the acquisition of them.

Other novelties of the reviewed seminal paper is the addition of two hallmarks emerged in line with the recent scientific advances such as deregulating cellular energetics and avoiding immune destruction. The first one comes from observed altered energy metabolism in many different cancer types. Some cancer cells switch to an "aerobic glycolysis" in which they take energy prioritizing glycolysis only instead of mitochondrial oxidative phosphorylation. It is believed that this preference provides the cell with lots of glycolysis intermediates that can serve to fuel biosynthesis pathways. Some of the related alterations with this energy switch are upregulation of GLUT1 (glucose transporter) and activation of oncogenes like RAS that among other things upregulates glycolysis. Regarding the second new hallmark (avoiding immune destruction), immune surveillance acts as the natural barrier against tumorigenesis and cancer progression so some solid tumors have managed to avoid detection and therefore destruction by the immune system. Transplantation experiments with immunodeficient mice have shown that cancer cells arising on those are inefficient when injected in immunocompetent hosts.

Some other lines of criticisms have emerged regarding this summarized view of cancer by Hanahan and Weinberg accusing them of considering cancer only as a cell-based disease caused by alterations in the DNA and without taking into account other points of view [18]. These critical voices pointed out the ignored evolutionary view of cancer that defines it more like a tissue-based disease. Among other arguments, they specifically call in question whether proliferation is an acquired cancer cell characteristic and therefore quiescence seems to be the base state of normal cells (as it has been suggested in the seminal paper). Instead, they claim that carcinogenesis is caused by a faulty interaction of the cells and their environment (other cells, extracellular matrix) which, in their opinion, is the real regulator against a default proliferative state of all cells which, in addition, resembles to what happens in organogenesis. This perspective of cancer model is collected in the tissue organization field theory (TOFT) which has not been as widely accepted as SMT over the past years of cancer research. However, the debate that surrounded both serves to remind the research community that cancer is more complex than it seems and multidisciplinary efforts must join to elucidate the biology behind and eventually find suitable cures [19,20].

Following this reasoning it seemed necessary that, despite the effort in the comprehension of the genetics and cellular biology of cancer, other aspects like the dynamics along time and space of tumors must also be taken into account for the clinical battle against it. Therefore, some years later, new ways to characterize tumors have emerged. With the advances in knowledge of cancer progression and tumor adaptability plus the technological improvements that allowed to retrieve several layers of information from tumors, some cancer researchers with evolutionary perspectives proposed a two dimensional framework to classify cancer according to the genetic, environmental and kinetics main characteristics of

it [21]. The two components of this classification are and Evo-index to capture the evolvability[4] of the tumors and the Eco-index that measures the environmental viability of the neoplastic cells.

This system provides up to 16 different categories when splitting each index into two subdivisions. The Evo-index rounds up the heterogeneity of neoplasms in space and time by relying on the concepts of diversity (D) and change over time (Δ). On the other hand, the Eco-index can be summarized into the hazards (H) or the deathly hurdles that cells must face and resources (R) that are fundamental for the cell maintenance (see Figure 3). For example, a tumor with low diversity (D) among its tumor cells, with a low mutation rate or genome instability (Δ) that ensures homogeneity, suffering from an hypoxic situation (H) that attracts immune response and limited resources to keep the growth rate (R), has little capacity to evolve and seems easy to eradicate. Contrary, the worst possible scenario would be a tumor that evolves rapidly (high D and Δ) and has plenty of resources (high R) which is highly adaptive to changes in the environment or any other affecting interventions (H: like immune evasion or therapy).

[4] The concept of a neoplasm as an evolving system is extended in the next section (1.1.2). Here, the definitions are limited to provide a general understanding of the classification

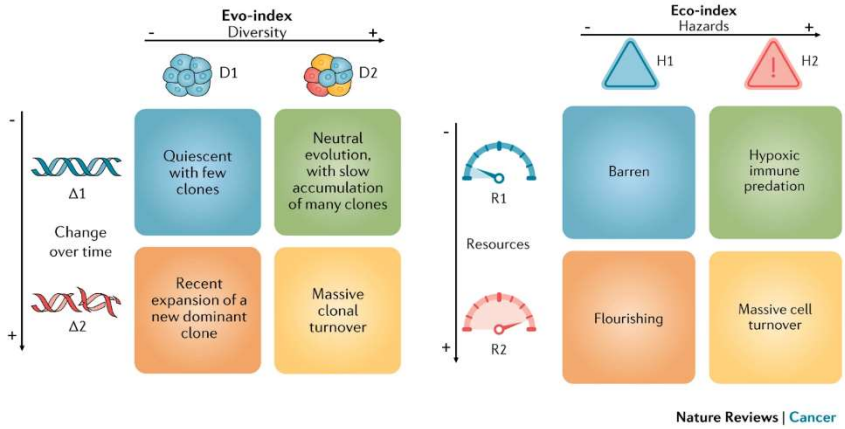


Figure 3. The Eco and Evo-index adapted and reprinted from *Classifying the evolutionary and ecological features of neoplasms* (v.17, p.605-619) by Maley et al., 2017 Nature Reviews Cancer

Both the hallmarks and the Eco-evo indexes emerge from the perspective of the cancer genomics research field. These initiatives define cancer disease as they provide ways to characterize them. However, there are other classifications (sometimes more specific) based on other criteria such as histological origin, histological stage and well defined biomarkers. Those are addressed below to guide where acute lymphoblastic leukemia settles. The concept of evolvability of this disease is relevant to this piece of work and is extended in the following section.

1.1.2 Darwinian evolutionary theory in cancer

The Darwinian evolutionary theory describes how populations of organisms change over time due to variation and the effect of natural

selection[5] on the heritable traits that influence the fitness[6] of individuals. The same rationale can be applied to neoplasms. In 1976, Peter Nowell [23] analogously characterized cancer development as an evolutionary process. In other words, a tumor can be understood as a population of individuals (cells) that accumulate changes (alterations) in the genome which are then subjected to the pressures of selection. The unrepaired genome variability can be advantageous to the cell and create a clonal expansion. That is, the cocktail of alterations is heritable to the daughter cells and is transmitted to the following generations making a quick population growth that outcompetes the rest of the cells. Various beneficial alterations carried by different mutant clones present a dynamic competitive scenario between cellular populations that we call intratumoral heterogeneity [24] (ITH or diversity as mentioned in the previous section). This diversity does not only imply changes in functional parts of the genome (i.e coding regions). There are increasing lines of evidence that epigenetic changes such as DNA methylation, chromatin remodeling and post-translational modification of histones are also sources of ITH [25].

There are several little clonal expansions that can create tissue mosaicism or benign forms of cell growth in certain normal tissues with particular constraints. However, occasionally, a cell can accumulate sufficient mutagenic load to become malignant to proliferate and invade. Heterogeneity can serve to reveal the tumor's life history as it is explained

[5] Charles Darwin observed that species have changed overtime as it is evidenced in the fossil records. He also noticed that the offspring of some species presents variation that makes them more suitable to survive "the struggles of existence" increasing their chances to reproduce and transmit their advantageous characteristics to the next generation. Therefore, he inferred that nature is able to select the favoral variability of the individuals and called that natural selection [22]. According to him and Alfred Russel Wallace, evolution happens by natural selection.

[6] Darwinian fitness is defined as the ability of an individual to survive and have fertile offspring

in the coming sections. Not only does it open the possibility to explore the past trajectory of the tumor but can also provide hinds for forecasting its progression and most likely outcome. In other words, ITH provides the tumor with high capacity of adaptability which usually challenges effectiveness of treatment and can result in a therapy-resistant tumor form.

1.1.3 Molecular cancer data

In 1971, the U.S government declared the "war on cancer" which stated a commitment to support research to reduce the incidence, morbidity and mortality from cancer [26]. A few years later, in 1986 it became apparent the need to obtain the full sequence of the cancer genome to systematically detect the mutated genes that cause it [27]. From 1990 to 2003 scientific efforts resulted in the sequencing of the human genome as part of the Human Genome Project (HGP) [28] and inspired other initiatives to sequence tumor genomes to reveal the basic cancer mechanisms setting the bases of the cancer genomics research field. A summarized definition of it is the following: "Cancer genomics is the study of the totality of DNA sequence and gene expression differences between tumor cells and normal host cells. It aims to understand the genetic basis of tumor cell proliferation and the evolution of the cancer genome under mutation and selection by the body environment, the immune system and therapeutic interventions."[29]

1.1.3.1 The revolution of Next Generation Sequencing in Cancer Genomics

The increment in knowledge due to cancer genomes research initiatives fostered the improvement of sequencing technologies and vice versa. It started with Sanger Sequencing, then continued by identifying mutations with capillary-based sequencing in exons that were individually amplified and then sequenced and has evolved to large-scale analysis of hundreds of

cancer genomes with massively parallel sequencing (MPS) (see Figure 4, [30]). Even within MPS technologies there has been a great improvement from 1 gigabase (GB) in a single run to more than 600 gigabases per run around 2012 [31]. Not only it provides higher-throughput but also it is possible to cover the whole-genome with a much more reasonable price, thus, increasing the chances to systematically apply genome sequencing to the clinics (see the sequencing cost of a human genome through years compared to Moore's Law in here [32]).

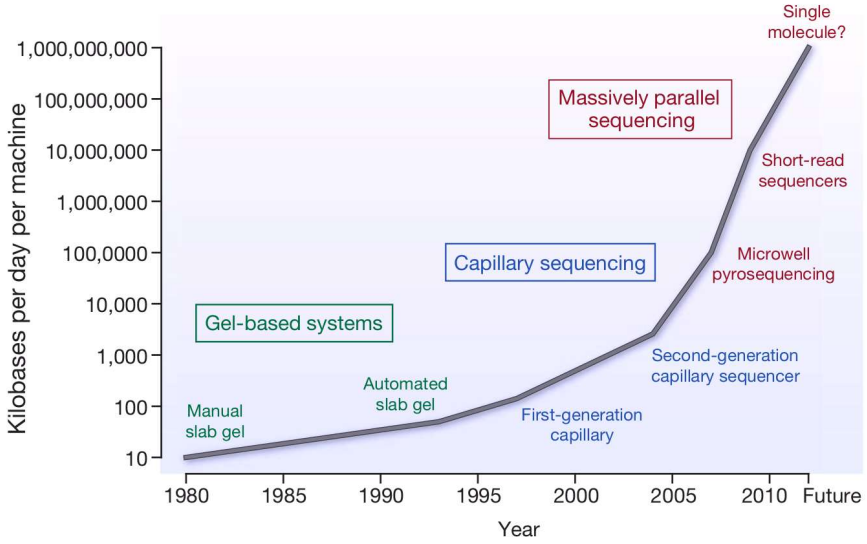


Figure 4. Improvements in the rate of DNA sequencing over the past 30 years and into the future reprinted from *The cancer genome* (v. 458, p.719-724) by Stratton et al., 2009 Nature.

First generation sequencing

Maxam-Gilbert and Sanger sequencing technologies are considered the "First Generation Sequencing". Maxam and Gilbert used radiolabeled DNA treated with chemicals to break the chain at specific bases and determine the position of the specific nucleotides by the length of the cleaved fragments in a polyacrylamide gel [33]. However, in 1977 Frederik Sanger

presented the “chain termination” method (also known as Sanger Sequencing; [34]) becoming one of the major breakthroughs in the history of biology and medicine. The key of this approach is that it mixes dye-labelled normal deoxynucleotides (dNTPs) and dideoxy-modified dNTPs (ddNTPs). The last ones are analogs of the first ones that are unable to bind to the next dNTP and halts DNA extension. Doing 4 of these reactions (1 per each ddNTPs) on 4 lanes in a gel generates fragments of elongated DNA. The shorted fragments migrate faster. The terminal base of those can be identified by autoradiography so the sequence can be inferred as there is a radioactive band in a given position of one of the specific lanes.

Several improvements were done to the Sanger Sequencing in the following years, especially introduction of capillary base electrophoresis. The first semi-automated sequencing machine was commercialized by Applied Biosystems in 1987 based on Leroy Hood improvement to Sanger Sequencing.

Second generation sequencing

The first commercialized “Next Generation Sequencing” (NGS) machine was Roche 454 sequencing system that used pyrosequencing method to provide mass parallelisation of sequencing reactions. This method differed from the past technologies because it does not use radio- or fluorescently-labelled dNTPs. Instead it infers the sequence by measuring the amount of pyrophosphate produced when a base is incorporated. The release of the pyrophosphate triggers an enzyme reaction that involves luciferase and so producing light [35].

Another important landmark was the Solexa Genome Analyzer, the first “short-read” sequencing platform that was launched and commercialized

by Solexa and lately acquired by Illumina [36]. This technology does also sequencing-by-synthesis[7] using reversible dye-terminators chemistry. The following years, Illumina developed the HiSeq platforms (models 2500/2000/1500/1000) which currently dominate the sequencing services and facilities. Short-read sequencing implied a methodological switch from chain-termination and electrophoresis to fragmented DNA, clonally amplified, loaded in newly developed microchips with improved chemistries that allows massively parallel sequencing and, therefore, reduces sequencing time and costs.

Another short-sequencing platform is SOLiD (Supported Oligonucleotide Ligation and Detection) System. As its name suggests sequencing is based on DNA fragments for ligation to oligonucleotide adapters using DNA ligase and not by synthesis like in sequencing-by-synthesis. Another notorious sequence-by-ligation (SBL) is Complete Genomics "DNA nanoballs" technique. Even though Illumina technology is more spread and used, both SBL technologies remain competitive [33].

Third and fourth generation sequencing

There is a diffuse boundary that separates second and third generation sequencing but here [33], the latter is defined as those technologies that are able to sequence single molecules and avoid DNA amplification. There are two main commercialized technologies worth mentioning here: single molecule real time (SMRT) platform from PacificBioscience (PacBio machines) and nanopore sequencing performed in GridION and MinION platforms by Oxford Nanopore Technologies (ONT). However, there are great differences between these two. The PacBio uses DNA polymerase

[7] Sequencing-by-synthesis refers to methods that need DNA-polymerase during sequencing. We can distinguish two subcategories: cyclic reversible termination and single nucleotide addition [35].

attached to the bottom of a "well-structure" called zero-mode waveguide (ZMW) with a small diameter. Fluorescent nucleotides are incorporated inside the ZMW which provide real-time bursts of fluorescent signal without interference of signals of other nucleotides. On the other hand, nanopore sequencing first requires the denaturalization of the strands of the DNA so that one strand enters the nanopore, a protein channel pore embedded in a synthetic membrane. The sequence is inferred as each base that enters the pore creates a different membrane current. Both technologies provide sequencing results of long-reads which are very useful in the *novo* assembly of genomes or to gain resolution to determine the breakpoints that define structural variants.

The fourth generation sequencing is known as "in situ" sequencing. It adds a new layer of information since the distribution of the reads coming from RNA are a reflection of the heterogeneity of the tissue sequenced [37].

Other sequencing technologies such as Single-Cell, Hi-C or ATAC-seq are being used to study cancer and cover other aspects to understand tumors (increase the resolution to the level of individual tumoral cells or analyzing chromatin interactions and accessibility).

Sequencing analysis and bioinformatics

With all the contemporary advances and market variability that came with nucleotide sequencing it also became apparent the need to standardize the huge amount of raw data that these technologies were producing (and that currently still produce). Since the Solexa Genome Analyzer II (GAII) platforms by Illumina have proven to be the ones with highest penetration in the market and the main representatives of NGS, the up-coming section is mainly focused in the data management and analysis of the genomic data of this technology. The processing of the great amount of sequencing data

consolidated a discrete but already existing field: bioinformatics. Also known as computational biology, it is defined as the research area that involves the storage, organization and analysis of biological, medical and health information using computers, databases, maths and statistics [38].

One of the first things that IT departments and/or bioinformaticians faced was storage. In one run of a GAII platform, 115,200 Tiff formatted files are produced per run, each at about 8 megabytes (MB) in size which sum-up to 1 terabyte (TB) [39]. The intensities of the images are processed to provide base calls in BCL format file (Binary Base Call). Those are then converted into FASTQ format which is a text-based sequencing data file that contains both raw sequence data and quality scores [40]. When the reads in the FASTQ format are aligned or mapped to the human reference genome, the output file can vary from 8 to 150 and also up to 300 GB depending on the read length (36-250 base pairs or bp), the depth and the breadth of the coverage that the reads provide [39,40]. There are different formats to represent the alignment files: SAM, BAM, CRAM format. The first one refers to Sequence Alignment Map (SAM) which is the human-readable form of the alignments. Each SAM file starts with a header followed by a row for every read together with 11 tab-delimited fields describing that read. BAM and CRAM files are compressed versions of SAM files. BAM files are the most widely used format since most of the processing algorithms take BAMs as the default input.

After some years of sequencing projects, there is now one gold standard pipeline for doing alignments and creating BAMs: “GATK Best Practices”. These are a series of workflows (each adapted to a particular experiment design) of the best way to use the Genomic Analysis ToolKit (GATK) which has been developed and maintained by the Broad Institute [41]. These workflows are the result of many years accumulating knowledge of

how to analyze high-throughput sequencing data (HTS). There are different experimental HTS designs: whole-genome sequencing (WGS), whole-exome sequencing (WXS), targeted or panel of genes sequencing (TGS) and RNA-sequencing. Regarding the DNA-seq methods each one of them is more adequate to help answer a particular research question and/or clinical necessity (see Table 1) depending on the genomic region(s) of interest as well as the project budget. For example, sequencing of a panel of genes targeting specific alterations is often used for accurate diagnostics of the (sub)type of cancer contributing to a better precision medicine whereas sequencing the whole-genome is more used for research purposes.

Platform	Cost (per sample, USD)	Sites	Region size (bp)	Depth	Data size
WGS	$1000–$3000	All coding and non-coding regions	~3 x 10^9	30-60x	Depending on coverage ~60-350 GB
WXS	$500–$2000	Exonic regions	~6 x 10^7	150-200x	Depending on coverage ~5-20 GB
TGS	$300–$1000	Specifically targeted regions	Varies by panel size ~1 x 10^5 - 1 x 10^7	200-1000 x	Varies by panel size and coverage ~100 MB–5 GB

Table 1. Different types of Next Generation Sequencing for genomics reprinted from *Applications and analysis of targeted genomic sequencing in cancer studies* (v.17, p.1348) by Bewicke-Copley et al., 2019 Computational and Structural Biotechnology Journal

Once the alignments are done, alterations can be detected by using *variant callers*. These are algorithms that report the variability of the sample genome and that is why it is said that they "call" variants. Many callers have been developed in the past few years [42–45]. The first caller distinction is

whether they differentiate between germline and somatic variants. According to the glossary added in Vogelstein et al., 2013 [46]:

- "Germline variants: Variations in sequences observed in different individuals. Two randomly chosen individuals differ by ~20,000 genetic variations distributed through-out the exome."
- "Somatic mutations: Mutations that occur in any non-germ cell of the body after conception, such as those that initiate tumorigenesis."

In most cancer sequencing projects, two samples are taken from each patient: one of normal tissue (or control) and another of the tumor mass. By sequencing, aligning and comparing both we can differentiate the germline variants as those present in both the normal and tumoral samples that are different from the reference genome from the somatic variants which are those exclusive to the tumoral sample. Even though we tend to associate tumor initiating alterations with a somatic acquisition process, there are many inherited germline variants that predispose to cancer. One of the most notorious cases is the inheritance of one alterated copy of *BRCA1* and *BRCA2* genes which increases the risk of developing various types of cancer [47]. One of the most widely used germline callers is *HaplotypeCaller* developed by the Broad Institute and distributed as part of the GATK.

Usually, callers are specific of one, or maximum two, types of alterations

- Single Nucleotide Variants (SNVs): A single nucleotide change in the sequence.
- Small Insertions or Deletions (InDels): Small gains (insertions) or losses (deletions) in the DNA sequence (from 1 to 100 bp).

- Copy Number Variants (CNVs): Sometimes also considered as intermediate SV but often distinguished and referred to as gains and losses of DNA fragments greater than 1 kilobase (kb) but less than 5 megabase (mb) [48]. In other words, can be summarized as detecting more or less copies of a DNA region from the expected two copies of a human diploid genome. Therefore, the gain in a DNA fragment does not refer to a *de novo* inserted sequence but to increase of a copy(ies) of a fragment.
- Structural Variants (SVs): A region of DNA that suffers a change in copy number (deletions, insertions and duplication), orientation (inversions) or chromosomal location (translations) [49]. Can be also understood as rearrangements of DNA sections, thus, some people consider whole-genome duplications and chromosomal aneuploidies as CNVs but not SVs.

In the current manuscript, "mutation" refers to SNVs and InDels but it is also used as a synonym of alteration by the community. In addition, as shown above, there is a distinction of CNVs from the rest of SVs as a different genomic alteration category. There are many callers that are specific for CNVs only. This is the reason why sometimes they are treated separately.

Before diving into the great variety of variant callers, there are other highly used sequencing techniques for detecting CNVs and SVs apart from short-read sequencing. The first one is BAC array-comparative genomic hybridization (array-CGH). This technique is widely used especially in clinical diagnostics. It provides detection of imbalances but lacks accuracy to provide absolute copy numbers. A contemporary method that also provides SV analysis is representational oligonucleotide microarray

analysis (ROMA). Another common method is the usage of SNP[8] array data to infer copy number variants. All of them have specific software to estimate copy numbers. However, with the coming up of NGS, especially of paired-end sequencing[9], it has been possible to detect SVs with a better resolution [49].

With all the available variant callers one can find themselves a bit lost when choosing which one is more appropriate. In that case, searching for benchmark publications helps in deciding. One of the best guidance when picking aligners and variant callers is the work of Alioto et al., 2015 where they used tumor-normal pair samples that were publicly available to compare sequencing methods, pipelines and validation methods to call variants [42]. According to their paper, Strelka obtained some of the highest precision and recall measures when tested with different datasets and using BWA as an aligner. The highest precision score was obtained when intersecting MuTect2 [41] and Strelka. Others have also benchmarked different callers and conclude that MuTect2 and Strelka seem to be performing better [51].

As mentioned above, there are certain algorithms specialized only on detecting copy number changes and contrary, there are others that have been designed to detect all (or most) of the SVs. Table 2 is divided into two, the upper one describes computational tools specified in CNVs detection whereas the bottom one lists some of the most used SVs callers. Columns are the same except for the one which differentiates:

[8] SNP: Single Nucleotide Polymorphism. A polymorphism, it has to occur in at least one in 100 people [50].
[9] Paired-end sequencing: Both ends of the DNA fragments are sequenced which provides better alignment and increases the quality of the sequencing.

- Segmentation algorithm used in CNV detection. These are changepoint algorithms that serve to define transition boundaries to localize and quantify copy number changes [52].
- Signal method used to detect breakpoints in SV analysis. There are 4 different signals to detect them [45]:
    - Read-depth: uses changes in read depth to identify regional rearrangements.
    - Paired-end: uses the abnormally mapped pair of reads (such as unexpected distance and direction) of the DNA segments to infer a SV event
    - Split-read: it works by splitting the short-reads in smaller fragments and then re-mapping them separately to the reference genome. The location of the breakpoint is revealed by the orientation and location of the splitted-remapped reads.
    - Local-assembly: it is used along with any of the above signals. It re-assembles reads that are already aligned to provide a better resolution of the breakpoint.

Software	Implementation	OS	Input	Sequencing	Variants	Segmentation	URL
FACETS	R, C++	Linux, Mac OS, Windows	BAM, snp-pileup	TGS, WXS, WGS	CNV	Joint segmentation (extended CBS*)	https://github.com/mskcc/facets
ASCAT	R	Linux, Mac OS, Windows	BAF and LogR values from BAM	TGS, WXS, WGS	CNV	allele-specific PCF^	https://www.crick.ac.uk/research/labs/peter-van-loo/software
Battenberg	R	Linux, Mac OS, Windows	BAM	WGS	CNV	allele-specific PCF	https://github.com/Wedge-lab/battenberg
Sequenza	Python, R	Linux, Mac OS, Windows	BAM	WXS	CNV	PCF	https://cran.r-project.org/web/packages/sequenza/vignettes/sequenza.html
Control-FREEC	C++	Linux, Windows (no longer supported)	SAM, BAM, SAMtools pileup	TGS, WXS, WGS	CNV	LASSO and dynamic programming	http://boevalab.inf.ethz.ch/FREEC/

PURPLE	Java,R	Linux	BAF and LogR values from BAM and SV VCF	WGS	CNV	PCF	https://github.com/hartwigmedical/hmftools/blob/master/purity-ploidy-estimator/README.md
VarScan(2)	Java	Linux, Mac OS, Windows	BAM, SAMtools pileup	WXS	CNV	CBS	http://varscan.sourceforge.net/
CONTRA	Python, R	Linux, Mac OS	BAM, SAM, BED	TGS, WXS	CNV	CBS	https://sourceforge.net/projects/contra-cnv/
ExomeCNV	R	Linux, Mac OS, Windows	BAM, Pileup, GTF	WXS	CNV	CBS	https://bioinformaticshome.com/tools/cnv/descriptions/ExomeCNV.html
Software	Implementation	OS	Input	Sequencing	Variants	Signals+	URL
BreakDancer	Perl, C++	Linux, Mac OS	BAM	WGS	SV	RP	https://gmt.genome.wustl.edu/packages/breakdancer/
Manta	Python, C++	Linux, Mac OS	BAM, CRAM	WGS	SV	RP+SR+LA	https://github.com/Illumina/manta

Delly	C++	Linux, Mac OS	BAM	WGS	SV	RP+SR	https://github.com/dellytools/delly
LUMPY	Python, C++	Linux	BAM	WGS	SV	RP+SR	https://github.com/arq5x/lumpy-sv
GRIDSS	Java,R	Linux	SAM, BAM, CRAM	WGS	SV	RP+SR+LA	https://github.com/PapenfussLab/gridss

Table 2. List of Structural variant (SV) and copy number variant (CNV) callers.
* CBS: Circular Binary Segmentation; ^ PCF: Piecewise Constant Fitting
+ Signals: RP Read-Pair; SR Split-Read, LA: Local-Assembly

The great variety of variant callers and the specificity of those to certain genomic alterations are a reflection of the amount of attention driven to the sequencing of tumors and the will for a precise characterization of those.

Pan-Cancer Initiatives and precision medicine

Due to the expansion of NGS, large-scale studies sequencing tumor samples of patient cohorts of different cancer types have been possible. These initiatives have aimed to uncover the main somatic alterations driving tumors with the ultimate goal of providing knowledge for more effective precision medicine approaches.

One of this first pan-cancer initiatives is The Cancer Genome Atlas (TCGA; [53]) which started in 2005 and it is a joint effort from multiple institutes. It began with a pilot project of only 3 cancer types and extended into two phases and ended up molecularly characterizing over 20,000 primary tumor samples from 33 different cancer types. It comprises not only genomic but also epigenomic, transcriptomic and proteomic data. With these data, it was possible to have the first genomic broad overview of different cancer types (12 at that moment) [54]. Furthermore, as a result of the analysis of lung squamous cell carcinoma cohort of TCGA a new clinical trial for lung cancer was launched (Lung-MAP; [55]) inspired by the results of the study [56]. The TCGA project has evolved to the Pan-Cancer Atlas [57,58] which is a resource that covers all the relevant findings from the published work derived from it.

As a consequence of TCGA and other large cancer sequencing projects (such as the Cancer Genome Project from the Wellcome Trust Sanger Institute), in 2008 the International Cancer Genome Consortium (ICGC; [59]) was built with the aim of coordinating large-scale cancer genome

studies. Under the umbrella of the ICGC, new advances in the understanding of tumor evolution of breast cancer revealed that the most recent common ancestor within the cancer cell populations appeared early in time so that there is a lot of subclonal diversification before diagnosis [60].

Most of the data collected in the ICGC data portal is focused on coding regions of the genome. In order to explore the non-coding parts of the genome and to study common patterns of mutations The Pan-Cancer Analysis of Whole Genomes (PCAWG) was launched. This large-scale project comprises more than 2600 cancer whole-genomes sequenced [61] and its analysis has revealed non-coding alterations relevant for cancer [62] as well as brought genomic analysis of somatic alterations closer to precision medicine [63] among other things.

Precision medicine also referred to as *personalized medicine* are used interchangeably but are indeed different things. Any medical appointment and any clinical decision made is personalized to the particular patient. In other words, physicians operate in an individualized way for each one of their patients and, therefore, medicine is intrinsically personalized. Having clarified that, precision medicine means applying medical procedures based on genetic, environmental, and lifestyle factors of the patient for better treatment efficiency [64]. Cancer genomics field boosted with the NGS outbreak anticipated to revolutionize oncology by identifying cancer specific events that can guide clinical decision-making [31,65]. Currently, targeted sequencing panels of cancer genes for diagnosis, prognostic and prediction of drug-response outcome are being used by clinicians to adjust treatments (e.g MSK-IMPACT panel from the Memorial Sloan Kettering Cancer Center [66]). However, there are still several bottlenecks between the huge amount of data generated from all the large and medium-scale

cancer studies and the delivered information that finally is evaluated for clinical decision-making. Concretely, the major bottleneck is the interpretation of the clinical significance of the genomic events which is very well summarized in Good et al., 2014 (Figure 5).

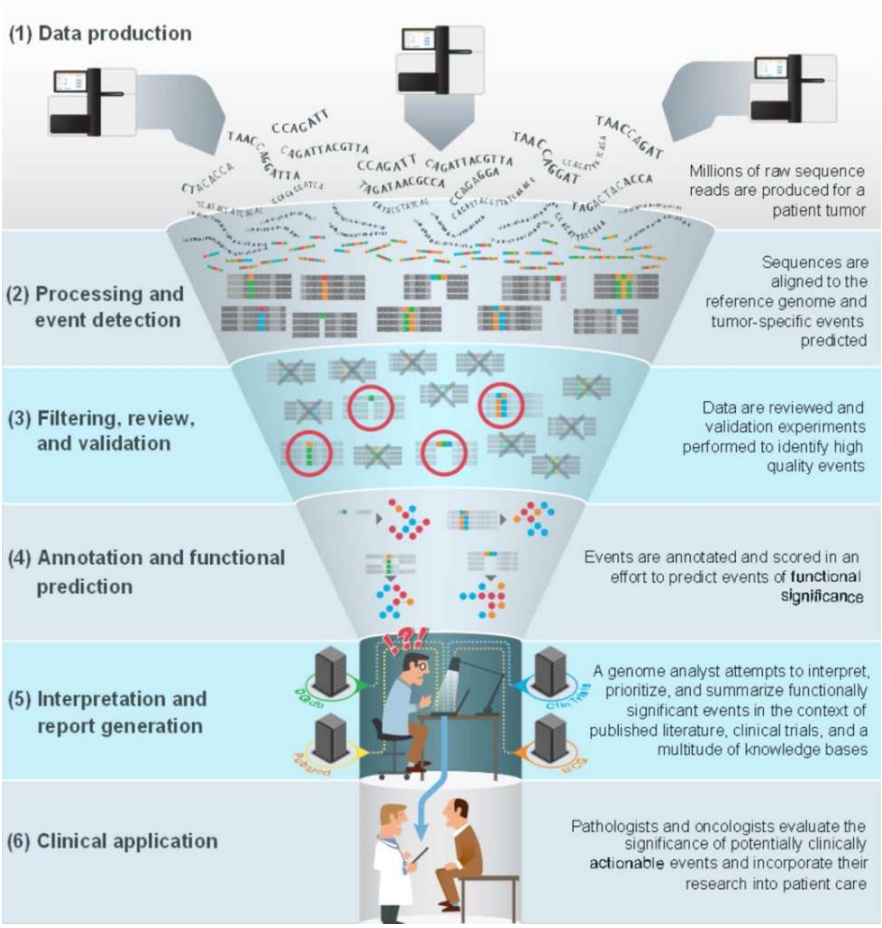


Figure 5. The interpretation bottleneck of personalized medicine reprinted from *Organizing knowledge to enable personalization of medicine in cancer* (v. 15, p.1-9) by Good et al., 2014 Genome Biology.

1.1.3.2 Acquisition of somatic alterations

Cancer genomics has been studying the acquisition of somatic alterations:

1) to understand the relevant somatic events that can be tackled to stop the cancer growth, spread potential and therapy resistance
2) from an evolutionary perspective to comprehend the cancer progression
3) to learn about the mutational processes that the tumoral cells are going through

Driver and passenger mutations

As mentioned before, it is of common knowledge that cancer arises due to the accumulation of genomic abnormalities. However, noticing the large mutation burden per sample of different cancer types it becomes evident that not all the alterations (ranging from 41 and 2.5 million point mutations per genome according to Radhakrishnan and Pich et al., 2017 [63]; see Figure 6) are responsible for tumorigenesis and cancer progression.
Therefore, one of the main goals of cancer genomic researchers has been to identify the alterations that are truly driving carcinogenesis. From here, we define as "driver" alterations those genomic events that confer growth advantage to the cells harboring them and that have been positively selected during the evolution of the tumor. The rest of non-driver alterations are called "passengers" which are not contributing to the cancer development and provide no functional consequence [30]. Passenger alterations are carried along in the clonal expansion derived from selection upon drivers. From Figure 6, one can see that in all cancers, passenger mutations outnumber drivers making their identification more challenging. The general focus has been to identify "cancer genes" which by definition are those carrying driver mutations (hence, also called driver genes).

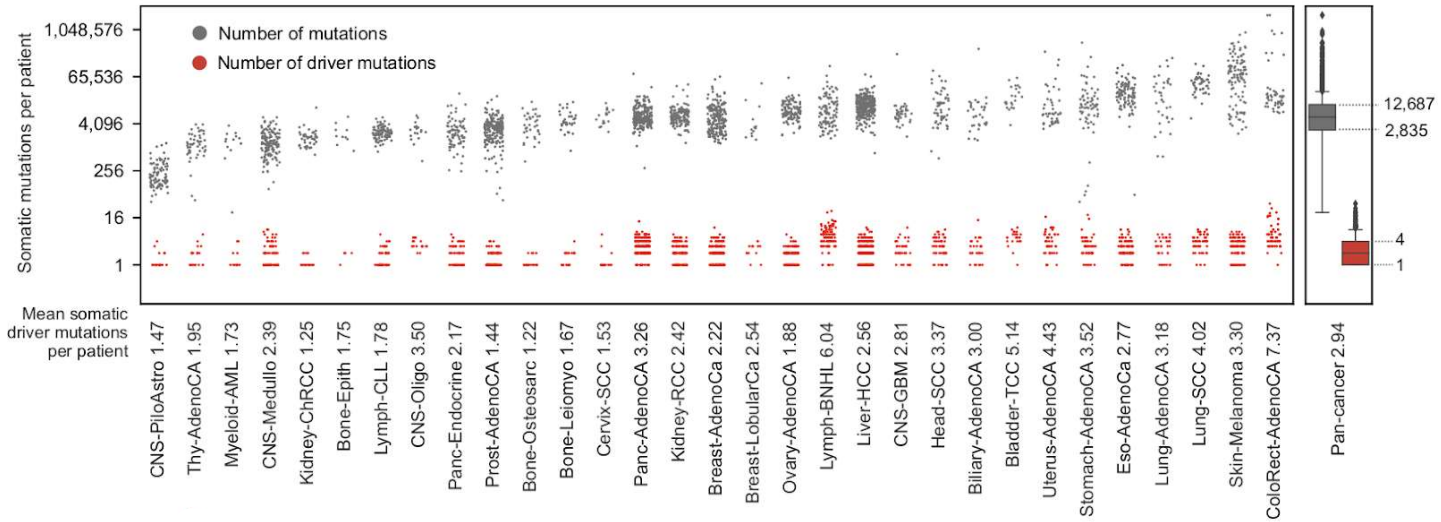


Figure 6. Mutation burden from whole-genome sequenced tumors of PCAWG reprinted from *The whole-genome panorama of cancer drivers* by Radhakrishnan and Pich et al., 2017 bioRxiv.

The main approach to systematically find cancer genes has been to detect signals of positive selection by analyzing the somatic mutations from sequenced tumors of large cohorts of cancer patients (e.g. TCGA/ICGC). The general procedure is to compare the mutational patterns with their expectation under neutral mutagenesis. During the last few years, many genomic research groups have developed algorithms and computational tools that search for the following characteristics of the mutational pattern of genes that constitute signals of positive selection:

- Mutational recurrence of genes. That is, genes that are found recurrently mutated across many cancer patients in a cohort which points towards a relevant role of the gene in the tumorigenesis of that particular cancer type. To computationally identify those, most algorithms search for genes mutated more frequently than the expected background level (understood as the background mutation rate). Examples of that are MuSiC [67] and MutSigCV [68]. Searching for recurrent mutated genes was the first step to find candidate driver genes [69]. It has also dragged with it some controversy since a good estimation of the background mutation rate is the key to avoid false positives and false negatives [70–72]. Currently, new approaches have emerged accounting for many confounders to improve the modeling of the background mutation rate to accurately measure the excess of mutations in genes. These approaches assert gene-specific positive and negative selection by measuring mutation count bias while correcting for covariates (regional genomic characteristics, mutational processes and consequence type). Methods that apply this are also quite recent like dNdScv [73] and cBaSE [74].
- Detection of functional impact (FI) bias. Identification of genes with a bias towards the accumulation of somatic mutations with

high FI. A score is given to each mutation. Some of the methods that assess the FI of non-synonymous mutations[10] are SIFT, PolyPhen2, MutationAssessor or CADD framework. Those provide scores that have been used in OncodriveFM [75] (and its evolved version: OncodriveFML [76]) to compute the bias and identify cancer drivers.

- Uncovering of mutational clusters. Find genes that have mutations clustered in particular regions of the sequence affecting specific amino acids of the protein. There are sequence-based clustering approaches such as OncodriveCLUST [77] (and its evolved version: OncodriveCLUSTL [78]), protein three-dimensional clustering approaches such as HotMAPS [79] and clustered mutations in protein-domains like smRegions [80]
- Detection of tri-nucleotide specific bias. Taking into account the number of mutations and nucleotides context (5' and 3' flanking bases) of point mutations (thus,"tri"-nucleotide) helps to identify cancer genes. This is a very recent approach [81] that takes into account the differential probabilities of each nucleotide context.

All these approaches are complementary to each other so one gene can show more than 1 signal of positive selection. In fact, the accumulation of evidence of positive selection helps to accurately define a list of candidate driver genes. One of the leading initiatives that particularly focuses on providing the most complete list of cancer genes is IntOGen (Integrative OncoGenomics) platform. It is a framework that automatically identifies

[10] SNVs are also known as point mutations or single-base substitutions (SBS) which can be classified as synonymous or non-synonymous. A synonymous mutation (also referred as silent mutations) means that the nucleotide change results into a codon which translates for the same amino acid (AA) as the original sequence. Non-synonymous therefore, means that there is a change in the AA residue of the protein and are further classified according to their consequence type in: missense (if it changes the AA) or nonsense (if it creates stop codon and as consequence the protein translation is prematurely terminated).

and characterizes cancer genes. It has a pipeline implemented which runs 7 driver discovery algorithms (dNdScv, CBaSE, MutPanning, OncodriveCLUSTL, HotMAPS, smRegions, OncodriveFML and dNdScv) and combines their results to create a compendium of driver genes and a repository of the mutational features associated to them that help to explain their mechanism of action (see Figure 7). The workflow is available in the Web platform (https://www.intogen.org/search) together with the results from the analysis of 28000 sequenced tumors from many projects and genomic dataset repositories such as cBioPortal, pediatric cBioPortal, ICGC , TCGA, PCAWG, Hartwig Medical Foundation, TARGET and St. Jude Cloud. Relevant results from the analysis of all these mentioned datasets can be found in Martínez-Jiménez et al., 2020 [82] as well as a historical revision of the identification of cancer genes.

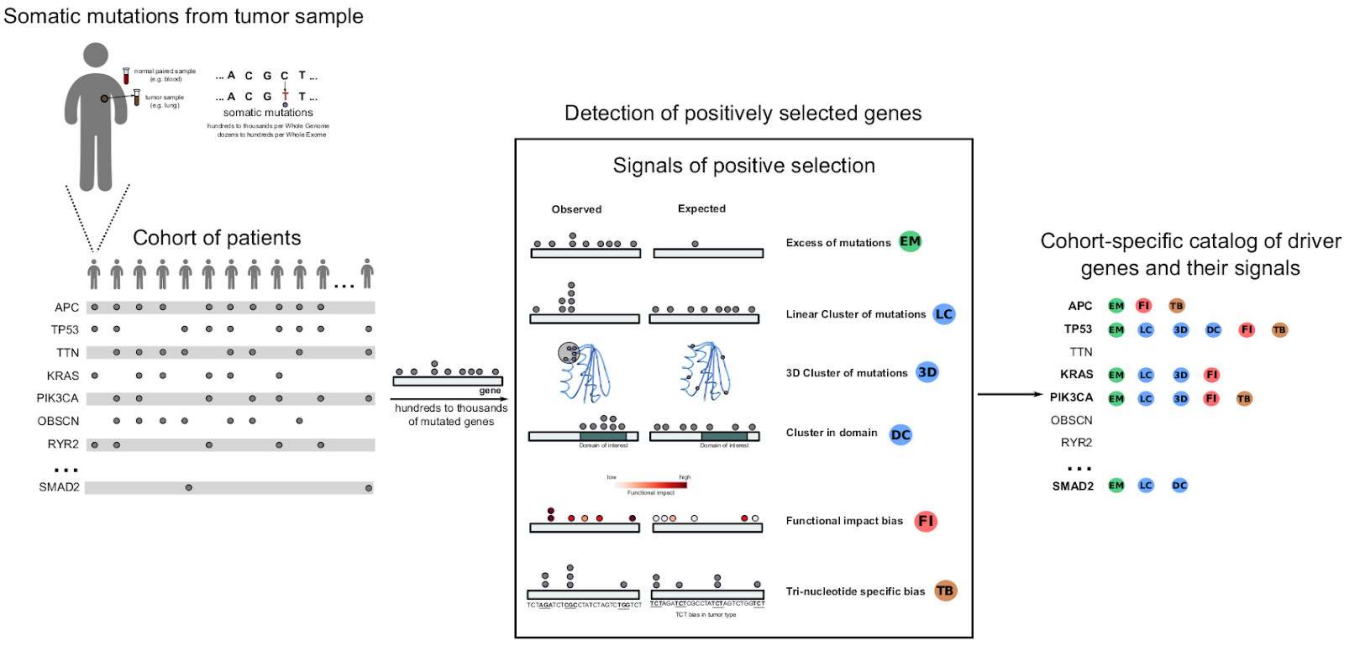


Figure 7. Systematic identification of cancer genes reprinted from *A Compendium of Mutational Cancer Driver Genes* by Martínez-Jiménez et al., 2020 Nature Reviews Cancer.

As mentioned before, cancer genes can be classified according to their mode (or mechanism) of action. Vogelstein et la., 2013 define cancer genes as [46]:

- Oncogene: A gene that, when activated by mutation, increases the selective growth advantage of the cell in which it resides
- Tumor suppressor gene: A gene that, when inactivated by mutation, increases the selective growth advantage of the cell in which it resides.

Previously mentioned, oncogenes were first identified in cancer-causing retroviruses [11]. Normal cell genes involved in relevant processes that, when altered are activated and have oncogenic potential, are called *proto-oncogenes*. We call this type of activating changes *gain-of-function* mutations which normally affect specific protein AA residues that confer them with a constitutively activated protein form. Since these are very specific places in the sequence, oncogenes tend to accumulate missense mutations in particular positions. Other alterations affecting oncogenes are duplications that lead to an overexpression of the gene or translocations that bring genes under the control of a different promoter or enhancers causing its overexpression [83].

On the other hand, mutations affecting tumor suppressors are called *loss-of-function* mutations which, as the name explicitly reveals, create a nonfunctional protein. *Nonsense* mutations creating truncated versions of the proteins are abundant among tumor suppressors. This type of cancer genes normally encode for relevant regulators of the cell such as checkpoint-control proteins of the cell cycle, enzymes of DNA repair or cascade members of apoptosis for example. The pattern of *loss-of-function*

mutations in tumor suppressor genes show more positional variability along the sequence compared to oncogenes [46].

The most reliable source of cancer genes together with related information such as their cancer incidence and their mode of action is the Cancer Gene Census (CGC; [84,85]). This is a catalog of genes (approximately 700 drivers) curated from the literature. Most of the tools mentioned above have tested their accuracy based on their ability to recover CGC genes. As a result of a driver discovery exercise one ends up with a list of candidate driver genes. There are also methods to infer their mechanism of action.

Endogenous and exogenous mutational processes

The great amount of data coming from the somatic mutational catalogs that the global cancer sequencing initiatives have produced, opens a new opportunity to understand the mutational processes that the tumor cells have suffered.

More than a decade ago, Gerd P. Pfeifer showed the presence of certain mutation patterns in *TP53* sequence in lung and skin tumors which coincide with the mutation types observed experimentally after the exposure to particular carcinogens. Him and colleagues reported that G>T/C>A transversions were more abundant in smoking patients than nonsmokers and these transversions are also more abundant in lung cancers compared to other cancer types pointing towards a mutagenic signature caused by tobacco carcinogens such as benzo(a)pyrene damaging the DNA [86,87]. Also, he described the abundance of C>T and CC>TT mutations as the result of the replication errors of the polymerase when encountering pyrimidine dimers in dipyrimidine sites caused by the exposure to UV light in skin cancers [88]. These observations were limited to a tumor suppressor

gene under positive selection. However, similar observations were made also in more comprehensive somatic catalogs of malignant melanoma and lung cancer in the following years [89,90].

It became apparent the need to explore mutational patterns of other cancer types with less evident associative relation (tobacco smoke → lung; UV light → melanoma) to understand the underlying mutational processes operating. Many considerations must be taken into account to design an approach to detect these types of mutational imprints. First of all, the cell has different mechanisms of repair that attenuate and, consequently, shape the signals left by the DNA damaging agent. For example, in the above mentioned studies they associate the presence of transcriptional strand bias in lung cancer as a reflection of the past activity of transcription-coupled nucleotide excision repair. Second, several exogenous and endogenous carcinogens might be acting in the same tumor development, therefore, mixing the particular "signature" of each one. The unique combination of mutation types imprinted in the DNA by specific mutational processes are, indeed, called *mutational signatures*. Each one can be understood as a probability distribution of the 96 types of mutations (in this particular context, "mutation" refers to SBS). The different mutations types can be summarized into 6 different substitutions C>A, C>G, C>T,
T>A, T>C and T>G taking as reference the pyrimidines (C,T) and therefore adding the reverse complement counts to the corresponding type. These substitutions are usually referenced with their sequence context: the 5' and 3' nucleotides of the flankings from the substitution position. Thus, 96 comes 6 substitutions * 4 possible 5' nucleotide context * 4 possible 3' nucleotide context. The 96 substitution types are usually represented as a mutational profile (also called mutational *spectrum*) in which one can easily visualize the trinucleotide channels with the amount of substitutions showing the mutational process activity. Apart from signatures, any

somatic mutation catalog from a tumor sample can also be represented as a mutational spectrum. In fact, most of the computational frameworks to detect signatures take mutational spectra as the input data structure with the counts for each of the 96 different mutations of each sample. As highlighted above, the general idea behind all signature computational approaches is that the somatic mutational catalogue is a combination of different signatures reflecting the mutational processes in which the ones with the higher activity during the tumor development have more *weight* in the required decomposition exercise to detect them [91].

We can distinguish two ways to detect the mutational signatures present in a somatic catalogue of a sample or group of samples (such as a cohort of a particular cancer type):

- *de novo* extraction: This approach aims to discover novel signatures.

  The first and most notorious method to extract mutational signatures was first used here [92] as part of the Breast Cancer project within the ICGC and, after, fully described in the landmark paper of Ludmil B. Alexandrov at the beginning of 2013 [93]. This computational framework implemented in SigProfiler uses a decomposition algorithm called Non-negative Matrix Factorization (NMF). This algorithm requires the somatic catalog as an input matrix data and the number of mutational signatures to be deciphered. Thus, they also included a model selection approach to determine the number of signatures. The application of this method to 507 whole genomes and 6,535 from exome sequenced tumors from 30 different cancer types revealed the first mutational signature profiles [94]. With the data from PCAWG the list of mutational signatures was amplified [95]. All of them are publicly

available in COSMIC [96] which not only includes the SBS signatures but also doublet base substitutions signatures and small InDels ones.

Following the Alexandrov model there are other *de novo* extraction methods developed such as SignatureAnalyzer (https://github.com/broadinstitute/SignatureAnalyzer-GPU/).

- fitting of signatures: The goal in this approach is to infer which signatures are active in the somatic catalog from a set of reference signatures such as the ones in COSMIC. Two examples are deconstructSigs [97] which solves the fitting using multiple linear regression model and sigfit [98] which adopts a Bayesian NMF approach.

There are more than 60 signatures detected in cancer. Some of them are of unknown etiology and some of them are believed to be sequencing artifacts. For example, Figure 8 shows some of the mutational signatures of known etiology. Signature 4 is found mostly in lung cancers (lung adenocarcinoma and squamous cell carcinoma) and it has been associated with tobacco smoking. In addition, there are 4 different signatures, all called signature 7 (a,b,c,d), that have been related to UV light damage (example in signature 7a in Figure 8). Furthermore, signature 1 and signature 5 are called "clock-like" signatures because mutations attributable to these tend to accumulate at a constant rate over time and, therefore, are proportional to the chronological age of the sequenced patient [99]. In fact, not only have they been detected in tumor samples but are also found in normal tissue [100,101]. Signature 1 shows high abundance of C>T mutations in (N)CG trinucleotide contexts which can be attributable to an endogenous mutational process associated with age. It has been observed that

spontaneous and/or enzymatic deamination of 5-methylcytosine to thymine generates mismatches in the double helix that if unrepaired at the time of replication become fixed as a C to T substitution which match with the observed pattern in signature 1. Regarding signature 5, there is no clear etiology. The number of mutations attributable to this signature correlates well with the age of the individuals. However the rates of signature 5 acquisition differ between cancer types without presenting a correlation with the stem cell division of each particular tissue of origin.

Figure 8. Mutational profiles of SBS signatures. Bar plots taken from COSMIC. The x-axis has each one of the 96 mutation types given as trinucleotide contexts. The y-axis represents the percentage of them. From above to below profiles of signature 4, signature7a, signature 5 and signature 1.

Currently, there is special interest in sequencing healthy tissue and pre-malignant neoplastic forms [102–105]. Some of these studies have shown that known mutational signatures of cancers can be recovered from healthy tissues (especially signature 1 and 5) [106,103,107,108]. Moreover, novel signatures are being detected in healthy cells which might reflect endogenous mutational processes specific to the tissue. One example of that is as a characteristic mutational profile detected in hematopoietic stem-cell

population and acute myeloid leukemia samples [101,109] which seem to be characteristic of the hematopoietic cell lineage.

Other interests regarding the analysis of mutational patterns involves the study of the consequences of chemotherapies. These treatments affect both cancerous and healthy cells. The damage in normal tissue may eventually translate into fixed mutations which may contribute to long-term secondary effects. There have already been described some therapy-related signatures such as signature 11 which is related to temozolomide, signature 31 which is associated with platinum-based drugs or signature 32 detected in post-treatment samples with azathioprine [94,95,110]. The detection of novel signatures seems to be going into this direction with very interesting recent projects in which they describe several footprints of chemotherapies in metastatic tumors including the novel capecitabine and fluorouracil (5-FU) signature [111] or another novel signature related to thiopurine treatment in pediatric relapse samples of acute lymphoblastic leukemia (now included as signature 87 in COSMIC; [112]).

1.1.3.3 Positive selection in cancer vs Neutral tumor evolution

During the 90's and early 2000's some speculative controversies took-off among cancer researchers. On the one hand, due to the first genomic characterization of tumor samples (especially from colorectal cancers) it started to prevail the idea that carcinogenesis might require genome instability and the acquisition of a hypermutator phenotype to develop and establish [113,114]. On the other hand, others argued against it since: (1) many tumors do not show chromosomal instability nor present alterations in key pathways that enhance mutability of the genome such as DNA repair pathways and (2) stressed the power of selection, instead of increased mutation rate, to force sporadic tumors to appear due to clonal selection and

expansion of cells harboring drivers [115,116]. As pointed out in here [117], these opposing views are not necessarily mutually exclusive. For some cancer types such as colorectal and endometrial cancers with defects in members of the DNA mismatch repair pathway, the mutator phenotype hypothesis adjusts well. The increase in the mutation rate provides a wider mutational spectrum upon which selection can act. For some other cancers, it seems that the normal mutation rate might be sufficient to account for the observed tumor development characteristics without the need of hypermutation [30].

Another debate dominated the past few years. The dNdScv framework [73] is not just a driver discovery method but it has provided an approach to estimate positive selection beyond the sequence of a gene. It can also be applied to a group of genes or even in the entire exome of the tumors revealing genome-wide measures of selection in cancer. The core of the approach is that it applies the normalized ratio of non-synonymous to synonymous mutations (dN/dS) which has historically been used in species evolution studies, to quantify selection (positive and negative) in cancer genomes. This apparent simplistic model has been refined to account for covariates to provide a good estimate of the background mutation rate at different scales (both locally and globally). The application of the method to PCAWG data showed no evidence of negative selection on coding point mutations. Also the analysis of gene sets revealed no clear signals of purifying selection. Therefore, except for driver mutations all the rest of coding somatic substitutions (~99%) seem to accumulate neutrally. The absence of negative selection contrasts with other studies applying also dn/ds (with slight differences) in which essential cellular genes turn to be negatively selected along with detected negative signals in the proteins controlling peptide exposition and the immunopeptidome itself [118].

Some years later, a paper came out describing that some colorectal tumors came from a single expansion and the cell subpopulations observed had driver alterations private to each one of them which appeared at very early stages of the tumor growth. This cancer growth behavior was named the "Big Bang" model [119]. A few years ago, a new approach to measure selection came out inspired by the Big Bang model and hypothesized that this mixture of subclones might be explained by neutral tumor growth dynamics. As a summary, they argue that mutations arising during the neutral expansion follow an accumulation distribution of 1/f power-law where f represents the allelic frequency. Then, adjusting a R^2 goodness-of-fit to the cumulative distribution of 1/f they indicated a threshold of $R^2 \geq 0.98$ to call for neutrality [120]. A proportion of tumors sequenced (approximately ⅓) across cancers presented neutrality under the model, especially gastric and colon cancers.

The relative simplicity of the model raised criticism among The PCAWG Evolution and Heterogeneity Working Group which argue that excessive assumptions were taken. Among the points stressed in the reply note [121], they run dN/dS for subclonal mutations previously considered as neutral which resulted in signs of significant positive selection. In the reply to the reply note, the authors of the original paper re-analyzed the data with dN/dS (excluding some problematic patients) and stressed that the neutrality detected in the original paper was confirmed by non-significant positive selection result [122].

All in all, the interest to detect natural selection and to better understand the evolution of tumor population cells is a cancer genomics hot topic which reflects its relevance towards studying driver alterations within the intra tumoral heterogeneity of neoplasias to avoid therapy resistance. In addition, the combination of methods should shed light for a better comprehension

of cancer rather than cloud over it. The scalability of methods designed for illuminating specific questioned areas must be taken with caution.

1.1.3.4 Evolution patterns through space and time

The somatic catalog of sequenced cancer cells represents a snapshot in the evolution of the tumor. The identification of drivers and the deciphering of the mutational processes acting, are approaches to a better understanding of tumorigenesis. However, inferring the ITH and how this mixture of populations of cells have evolved is fundamental to have the whole picture. The relevance of it lies in avoiding treatment failure. Therapy can eliminate most of the tumoral cells and therefore reduce the competence of the resistant ones that can progress again and cause recurrent cancer. When it happens, it is sometimes classified as advanced cancer or stage IV in the clinics. The majority of the cancer-related deaths happen at this point.

Treatment resistance, relapse and metastasis

The failure of the therapy to completely eliminate tumoral cells can manifest as the patient undergoing a second malignancy in a post-treatment period after the primary one was believed to be eradicated. This is called a relapse. When a second neoplasia appears in a different region (tissue/organ) from the primary site of the first cancer is called metastasis. Metastases diagnosed after the treatment of the first primary tumor are also called relapsed metastasis. In contrast, *in situ* cancer recurrence is sometimes called just relapse.

Metastases occur due to a multistep process which involves the dissemination of cancer cells to anatomically distant organs followed by the adaptation to the new particularities of the tissue microenvironment. This is called invasion-metastasis cascade and can be summarized in the following steps [123,124]:

1) locally invade the adjacent tissue through the surrounding extracellular matrix (ECM) and stromal cell layers
2) intravasation into lymphatic system and/or bloodstream
3) survive the circulation and vasculature and stop at the capillary system of a distant organ
4) extravasation into the new tissue location
5) colonize by overcoming the microenvironment hazards
6) generating a viable niche to grow

The development of the metastasis implies genetic and epigenetic changes that settle an heterogeneous scenario for selection to favor some traits under the pressure of these successive bottlenecks. Some of these changes have been well studied. For example, in carcinomas (epithelial derived cancers), cells undergo some phenotypic changes in which they lose intercellular adhesion and polarization and acquire motility and invasiveness characteristics called epithelial-mesenchymal transition (EMT) with a similar cellular program that in embryonic development.

It is thought that a lot of "seed" cells of the primary tumor die during this process especially during colonization [124]. Since usually, this is the critical point, it has also been observed that some cells that reach a distant tissue (micrometastasis) undergo dormancy state and sprout when the conditions are favorable to create a macrometastasis. The circulating tumor cells (CTCs) found in the bloodstream of patients have been observed to travel isolated alone as well as in clusters which may give rise to polyclonal metastatic seeding [125]. The genetic and epigenetic commonalities and differences observed when comparing primary and metastatic samples of the same individual can provide hinds of the seeding process. Intriguingly, there are primary tumors with "preferences" for certain metastatic sites such

as breast cancer which frequently metastasize in lung, bones, liver and brain.

The origin of such tumor heterogeneity which contributes to treatment failure and recurrence has motivated different explanations. One of those is the cancer stem cell model (CSC) which describes how self-renewing malignant stem cells maintain the clonality of the tumor. It is conceived as hierarchical organization of the tumor cell populations being the CSC at the top of it and therefore multipotent. Their tumor-initiating and clonal maintenance capacity have been observed through repopulation assays either by serial transplantation in recipients or *in situ* tracking studies [126,127]. When the first papers supporting this model came out, there was some controversy as it was seen as an opponent explanation to the clonal evolution model [128]. First, the CSC model explains ITH by an aberrant differentiation program whereas the clonal evolution model relies on competition among neighboring subpopulations of cells to produce such mixture. Second, the CSC model assumes that only a small pool of cells (CSC) contribute to tumor progression and therefore, are the ones that mutate and eventually become more aggressive which differs from the clonal evolution model which supposes that any tumoral cell acquires mutations and has the potential to progress. Third, regarding therapeutic resistance, the clonal evolution model considers that there is selection towards tolerant clones whereas the CSC model presumes that CSC are drug-resistant.

Apart from these differences, there are also some commonalities. In both theories, the tumor originates from a single cell that accumulates alterations and acquies proliferative power. Moreover, stem-cell like property is something compatible with both since not only can be a characteristic from the tumor cell of origin but also can be an advantageous trait to be selected

[128]. In fact, there are some explanations to reconcile both models, for example, by clarifying the term "stemness" (gain and maintain a stem-cell state). Stemness is influenced by cancer genetic and epigenetic diversity and the tumor microenvironment [126]. In fact, there is increasing evidence that the niche of the CSCs plays also an important role in the division of these cells which has been observed not to be as an asymmetric mitotic process as it was believed [129] but rather a more dynamic model. Also, for some cancers it is difficult to distinguish CSC from non-CSC since it seems to be a stemness generalized among tumoral cells as well as there are cancers in which it has been observed a reversing transition process between stem and non-stem cell states [126,129]. Related to that, the term "cancer stem cell" has also brought confusion into its origin since not all the CSC derive from normal stem cells. The concept of "stemness" should be restricted to its cell functionality independently on whether it refers to normal or malignant ones [130]. Besides, it is also difficult to isolate CSC since there are just a few clear markers (e.g. CD44+ in some solid tumors or CD34+/CD38- in leukemia) among cancer types and there is high variability in their frequency between tumors [131]. As a consequence of the previous reasoning, CSC term should be found restricted to tumor-initiating cell (T-IC) or leukemia-initiating cell (L-IC) which are [126]: (1) able to create a xenograft that is representative of the original tumor (2) able to self-renew in assays of serial passage in xenografts with different clonal cell dosages, (3) able to give rise to daughter cells that can proliferate but unable to establish or maintain the tumor in serial passage assays.

The lab of John E. Dick has extensively studied the role of L-IC in the context of acute myeloid leukemia (AML). Performing transplantation assays into severe combined immune-deficient (SCID) mice, they were the first to isolate and characterize them as CD34+/CD38- and they proved their self-renewing capacity [132,133]. Among other contributions, in 2011 he

discovered that L-IC clones of B-cell acute lymphoblastic leukemia harbored distinct genetic alterations and demonstrated a branched multi-clonal evolution model of leukemogenesis, thus, linking CSC and genetic evolution models to describe tumor heterogeneity [134].

Independently of the origin of ITH, tumors that present high ITH have more probability of treatment failure [135]. In a neoplastic mass, the more diverse, the more chances of the presence of chemoresistant cells, which are the major cause of relapse. Resistance is classified into two types of resistance [136]:

- *intrinsic* resistance: the factors mediating resistance are already present in the bulk before administration of the treatment.
- *acquired* resistance: it is developed during treatment by diverse therapy-induced adaptive responses. For example, it can occur in initially sensitive tumors that acquire resistant alterations during treatment or by an upregulation of an alternative compensatory signalling pathway of the therapeutic target.

Apart from the starting point of the resistance, some common mechanisms of resistance have been described, some which are dependent on the treatment and others that are more general [136]:

1) Drug transport and metabolism
    - drug efflux: some cell membrane transporters have been related to chemotherapy resistance by pumping out drugs. The most notorious case is the ATP-binding cassette (ABC) transporter family. Especially, overexpression of the gene MRD1 (ABCB1) has been related to multi-drug resistance in various cancers such as lung [137], breast [138] and leukemia [139,140].

- drug activation and inactivation: some drugs are delivered as prodrugs and are activated by cellular enzymes, so in the absence of those the cell tolerates the drug. Contrary, some cellular metabolites can inactivate chemotherapeutic agents. For example, platinum drugs can be inactivated by thiol glutathione.

2) Alterations in drug targets: alterations and/or changes of expression of the target can also create resistance. A notorious case is to find mutations in gatekeepers residues of kinases. This is a conserved residue at the opening of the ATP-binding pocket. Examples of that are mutations found in specific residue of the BCR-ABL1 oncogenic kinase, formed by rearrangement in chronic myeloid leukemia (CML), which conferred resistance to imatinib [141].
3) DNA damage repair: the effectiveness of chemotherapeutic agents inducing DNA damage depends on the cell capacity to activate DNA repair pathways. Thus, mutations that affect the response to DNA damage and DNA repair may increase the chance of survival of cells with large quantities of DNA lesions, such as those exposed to certain chemotherapies.
4) Downstream resistance mechanisms: that is, even though the cell accumulates anticancer agents and the target is inhibited, which should ultimately induce cell death, the cancer cell finds ways to survive like generating deregulation of apoptosis.
5) Resistance-promoting adaptive responses: these responses can be listed as:
    - activation of prosurvival signaling
    - oncogenic bypass and pathway redundancy: this is also called kinome reprogramming. It has been observed that due to the treatment, which is effectively inhibiting the

target (i.e. EGCF), an alternative kinase becomes activated.

- undergoing EMT: it is also believed that resistance can happen in some tumors as cells present certain plasticity to acquire stem-cell-like properties (like in EMT)

6) Tumor Microenvironment: It has been observed that stroma-induced resistance to different therapeutic agents. These interactions can change the sensitivity of tumor cells to some drugs [142]. Some of the observed influences of it are changes in expression of integrins and cytokines and growth factors such as autocrine, paracrine and endocrine activation of oncogenic signaling by growth factors.
7) Cancer Stem Cells: There is increasing evidence that CSCs confer resistance to chemotherapies, which are described somewhere [143] and some of them summarized here. First, it has been studied that the CSC niche provides protection against the exposure to drugs. For example, it has been observed that CSCs are surrounded by hypoxic conditions, for example, it has been demonstrated that hypoxia-inducible factor-1 (HIF1alpha) is required for the maintenance of L-IC in CML mouse models [144]. Paradoxically, a perivascular niche has also been reported to be essential for maintaining CSCs in certain tumors.

   In addition, as chemotherapy and radiotherapy target fast proliferating cells, it has been observed that CSCs are slow-dividers and most of the time acquired quiescent states, thus, avoiding the attack of these therapies. Finally, it has also been reported a high drug efflux by ABC transporters in CSC as well as some other general resistant characteristics described above.

Inferences of clonal populations and their dynamics

ITH fosters tumor evolution and can engender drug resistance, thus, it has attracted a lot of attention due to its clinical relevance [24]. As a consequence, many research groups have focused on disentangling the architecture of tumor and its different subclonal populations as well as deciphering the history of the tumor progression.

There are different mathematical models aimed to infer population dynamics described here [145] and summarized as follows. One of the first ones described was the multistage theory which models the probability of developing cancer as a function of the age. The kinetics of tumor initiation and progression such as the number of rate-limiting steps of the cancer (transforming steps towards carcinogenesis) can be estimated from age-incidence curves. In fact, in general, 6 rate-limiting steps have been inferred in cancer development which is quite close to the average number of driver alterations estimated in tumors (4-5 drivers [61]). Other more sophisticated models of population genetics have been applied in cancer research to study the evolution of tumors. For example, Wright-Fisher model can be used to simulate cell populations with a specific number of generations and study the accumulation of driver mutations through clonal expansions. Applications of it allow creating multiple cell types representing genetically different subclones of the tumoral mass using multinomial sampling and also accounting for number of mutations, selection force and genetic instability.

Phylogenies of the tumoral populations of a neoplasm are also widely used to represent the inferred reconstruction of evolutionary cancer processes [146]. In these representations, the clonal subpopulations represent the taxa of the phylogenetic tree. Most of the computational tools developed to infer

these phylogenies rely on the detected somatic SNVs and/or CNVs. In the case of SNVs detected, trees are built upon the inference of clones from the allele frequencies of the mutations. These are calculated with the number of reads mapping to the reference genome that harbor the variant respect to the total aligned and must be corrected by copy number changes affecting the regions as well as the normal contamination of the sample. This way, we can obtain the number of cancer cells having each substitution (also called Cancer Cell Fraction or CCF). After, the SNVs can be clustered by common frequencies into sets of mutations as a proxy of the subpopulations of cells in the tumor. This process is commonly known as clonal deconvolution. Some tools are more oriented to quantify ITH and only perform clonal deconvolution without inferring phylogenies such as SciClone [147] and Pyclone [148] but are useful when combined with other phylogenetic methods to infer trees [149].

Most of the algorithms developed to build cancer phylogenetic trees borrow classic evolution methods such as maximum parsimony, neighbor joining, UPGMA or Bayesian probabilistic inference methods and also a combination of some of those. For example, for bulk sequencing sample data, both PhyloWGS [150] and Canopy [151] are based on probabilistic models using Markov chain Monte Carlo (MCMC) sampling to obtain phylogenies that are consistent with the mutation frequencies. In addition, as more projects use single-cell sequencing, similar probabilistic approaches but specific to this data are used such as SCITE [152]. Related to the above, the diversity of the data has led a great variety of methods that can be divided in [146]:

- cross-sectional methods: gives information about the common progression of a population building trees from many tumor samples of a cohort.

- regional bulk sequencing: builds trees from single-patient data using samples of different tumor sites. This particular group of methods are the ones prone to combine deconvolution of clones and phylogenies.
- single-cell methods: uses the detectable cell-to-cell variation to create phylogenetic trees. Not only there are algorithms based on single-cell sequencing of DNA but also some preceding the sequencing technique that use fluorescence *in situ* hybridization (FISH) markers.

Notably, a simplistic but accurate phylogenetic representation of samples was developed by Nik-Zainal and colleagues within the Peter Campbell's group at the Wellcome Trust Sanger Institute [60] which has been later applied to many other followed-up research projects [153,154]. They "manually" build trees based on a deductive reasoning approach using the mutational frequencies of SNVs and borrowing the concept of "the most-recent common ancestor" (MRCA) from population genetics. The approach makes the following assumptions which seem reasonable within the evolutionary cancer setting:

1) Mutations occurred only once during tumor development. In other words, a position cannot be mutated twice which is referred to as the "infinite sites assumption".
2) Mutations cannot be undone or lost. In other words, back-mutations do not happen.

From these, one can deduce that two clone subpopulations harboring the same mutation implies that they share a common ancestor clone that had acquired the mutation and had transferred it to the daughter cells. Their approach follows three steps:

1) Phasing with Battenberg[11]: map mutations with adjacent germline heterozygous SNPs[12] to the copy number events which allows to determine whether a mutation is on the retained or subclonally deleted parental copy of a chromosome.
2) Bayesian Dirichlet process to perform a clustering of subclonal subtitutions
3) Apply the Pigeonhole Principle (PHP): an easy and very well-explained example of the principle is given here [155] as follows. According to the PHP no sum of the subpopulations can exceed the CCF of the ancestor. Imagine a deconvolution of the CCF that gives three different subclusters of mutations at 100%, 80%, 40%. Then 100%+80% > 100% therefore the subclone 80% must be a descendant of the 100% subclone. On the other hand, 80%+40% > 100% as a consequence 40% subclone must be a descendant of 80% one.

Usually, in this type of studies they use multiregional samples of the tumor and the phylogenetic trees show a trunk that represents the clonal mutations that are shared in all tumoral cells in every sample and the branches are subclonal cluster mutations. The length of the trunk and the branches represents the number of mutations specific to each lineage. This type of studies revealed that in breast cancer, primary tumors had a subclone lineage representing a 50% of tumoral cells [60] and that clones seeding metastasis disseminated late from the primary neoplasms but still acquired private mutations with some clinical actionable potential [154]. In contrast, the patterns of subclonal composition among the cohort of patients with multi-sampled breast tumors showed great variability [153].

[11] copy number caller presented in the same paper [60]
[12] such as the list of SNPs derived from the 1000 Genomes Project.

Apart from studying ITH and evolution patterns, cancer phylogenies trees also drove the interest to decipher the temporal sequence of driver events. There are different approaches to study the order of genomic events [156]. One way is to compare the driver alterations at different tumor development stages. For example, private mutations of metastatic sites that are enriched in particular genes or pathways in a cohort inform about the final phases of tumor evolution. Similarly, others have looked at displasias [100,105,157] or even normal cells [103,158] of the tissue to detect precursor lesions in cancer. For example, positive selection of oncogenic mutations in drivers of cutaneous squamous cell carcinomas such as *NOTCH1* have been detected in normal skin [103].

The most obvious thing to do would be to take serial samples of each patient but, usually, this is impossible. In general, one can say that clonal mutations, meaning mutations with CCF closer to 1, correspond to relatively early events in tumor evolution, most likely happening previous or at the time of the most recent clonal expansion, whereas subclonal mutations are usually considered later events. A pan-cancer analysis of TCGA [159] with single-patient primary data revealed that known cancer genes have a tendency to be clonal within across cancer types and that APOBEC-mediated mutagenesis happens late in tumor evolution and contributes to the acquisition of subclonal driver mutations. Furthermore, as seen before, multiregional sampling of a single biopsy serves to study tumor evolution and also to time its main genomic alterations. Those can be ordered by checking the relationship between SNVs and the surrounding copy number gains. For example, mutations are called “early” when they are present in the two alleles of the duplicated region because they must have happened before the gain event whereas mutations are called “late” when those are detected in a single allele since they most likely occurred after the duplication. In a similar way other events can also help order

mutations such as whole-genome duplications (WGD) or copy neutral loss-of-heterozygosity. Applying this reasoning the PCAWG Evolution & Heterogeneity Working Group revealed that across cancer types, mutations in the driver tumor suppressor *TP53* are usually early events, WGD happen in intermediate phases of the tumor evolution and copy number changes are typically late being losses earlier than gains [160]. In addition, they also used age-related mutations, which accumulate at a constant rate through time, as a calibrated "molecular clock" to get chronological time estimates of the WGD resulting in genome doublings happening several years before diagnosis. Other contemporary multi-cancer studies have also revealed that for some patients with synchronously diagnosed metastases, the systemic seeding can happen very early (approximately 2-4 years before diagnosis of the primary; [161])

The more we learn about the occurrence and latency of driver genetic aberrations the closer to early detection and prevention of cancer recurrence.

1.2. Overview of Leukemia

1.2.1 What is leukemia?

Blood and plasma cells are created through a highly regulated process called hematopoiesis. In this process the hematopoietic stem cells (HSC) differentiate and mature to form erythrocytes, megakaryocytes, and immune cells of myeloid, lymphoid, or monocytic lineage (see Figure 9) in bone marrow or lymphatic tissues (spleen, thymus and lymph nodes) which are tissues and organs composing the hematopoietic system. Genetic and epigenetic aberrations affecting HSC can cause a maturation arrest and uncontrolled proliferation of immature cells. When these cells (of any

lineage and hematopoietic immature stage) create a clonal expansion in the bone marrow, infiltrate and circulate at elevated numbers in the bloodstream we call it leukemia [162]. In certain cases (lymphoid lineage), there is also abnormal proliferation of the lymphatic tissues. Contrary, when the lymphatic tissues present malignant masses of well-differentiated lymphoid cells (lymphocytes) the disease is called lymphoma. There are two types of lymphoid cells, the T-cell and B-cell lineages which, as shown below, are often used to classify related malignancies.

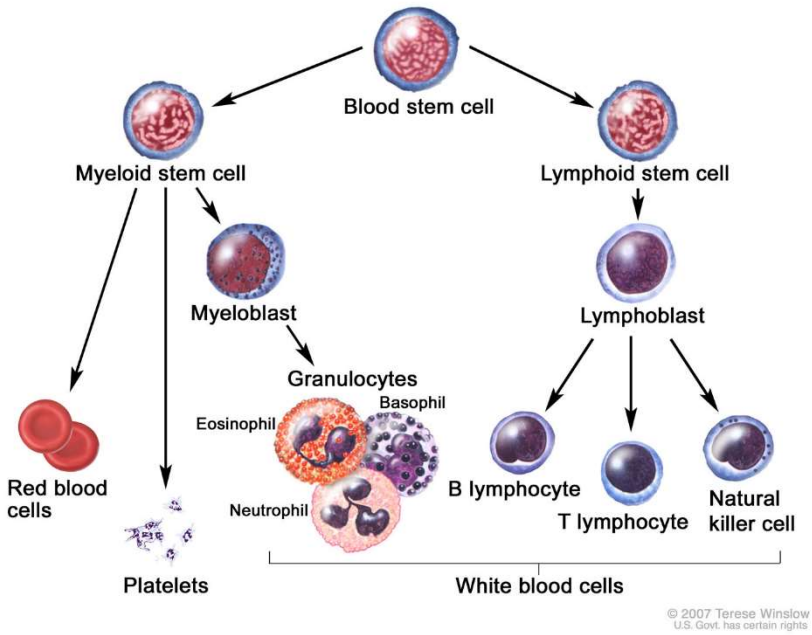


Figure 9. Illustration by Terese Winslow. Hematopoiesis process to produce blood cell types.

1.2.2 Cancer classification of leukemias

Different criteria allocates leukemia in distinct cancer groups. According to its cell of origin, leukemia belongs to the group of hematopoietic malignancies. However, due to its histological nature the different leukemia types are also classified as “liquid” cancers. Officially, the World Health

Organization (WHO) made the following general categories [163,164] in 2016 (on each group of the classification backbone, there are only written the malignant forms that are considered leukemias in bold):

Lymphoid malignancies:

- B Cell Neoplasm
    - Precursor B cell Neoplasms
        - **B-lymphoblastic leukemia/lymphoma** (with several subtypes)
    - Mature B cell Neoplasms
        - **Chronic lymphocytic leukemia/small lymphocytic lymphoma**
        - **B-cell prolymphocytic leukemia**
        - **Hairy cell leukemia**
- T Cell and NK Cell Neoplasms
    - Precursor T cell Neoplasms
        - **T-lymphoblastic leukemia/lymphoma**
    - Mature T cell Neoplasms
        - **T-cell prolymphocytic leukemia**
        - **T-cell large granular lymphocytic leukemia**
        - **Adult T-cell leukemia/lymphoma**
- Hodgkin's Lymphoma
- Posttransplant lymphoproliferative disorders (PTLD)
- Histiocytic and dendritic cell neoplasms

Myeloid malignancies:

- Myeloproliferative neoplasms
    - **Chronic myeloid leukemia (CML)**
    - **Chronic neutrophilic leukemia (CNL)**
    - **Chronic eosinophilic leukemia**

- Myelodysplastic and lymphoid neoplasms with eosinophilia and abnormalities of PDGFRA,PDGFRB and FGFR1
    - **Chronic myelomonocytic leukemia (CMML)**
    - **Atypical chronic myeloid leukemia (aCML), BCR-ABL12**
    - **Juvenile myelomonocytic leukemia (JMML)**
- Myelodysplastic and myeloproliferative neoplasms
- Myelodysplastic syndromes
- Acute myeloid leukemia and others
    - **Acute myeloid leukemia (AML)** (with several subtypes)
    - **Acute myelomonocytic leukemia**
    - **Acute monoblastic/monocytic leukemia**
    - **Pure erythroid leukemia**
    - **Acute megakaryoblastic leukemia**
    - **Acute basophilic leukemia**
    - **Myeloid leukemia associated with Down syndrome**

Acute leukemias of ambiguous lineage:

- **Acute undifferentiated leukemia**
- **Mixed phenotype acute leukemia**

Since the WHO classification is based on the cell of origin, some forms of leukemia and lymphomas are considered different manifestations of the same disease. For example, chronic lymphocytic leukemia (CLL) and small cell lymphoma (SLL) are both part of mature B-cell neoplasms or T-cell lymphoblastic leukemia and T-cell lymphoblastic lymphoma are joined under the same category (T-ALL/T-LBL). Sometimes, these types of cases are differentiated and separated in leukemias and lymphomas depending on whether the malignant cells prevail in bone marrow and blood or the lymph nodes and therefore distinguishing between “liquid” and “solid” hematopoietic malignancies. Therefore, clinically, a more used

classification for leukemias are lineage (myeloid or lymphoid) and condition (acute or chronic) which results in the four major types of leukemias: chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CML), acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML). Also lineage serves to distinguish solid hematopoietic cancers as myelomas and lymphomas. Usually, the latter are subdivided into Hodgkin Lymphomas (HL) and Non-Hodgkin Lymphoma (NHL) [50].

1.2.3 Epidemiology and etiology

The incidence of leukemias worldwide is of 6.1 per 100,000 in males compared to 4.3 per 100,000 for females. Mortality is also higher in males than females (4.2 per 100,000 vs 2.8 per 100,000 respectively) [162]. ALL is considered a rare cancer in adults whereas CLL and AML are the most frequent ones. In contrast, ALL is the most frequent leukemia among children (75% of all leukemias in pediatric compared to 12% in adults [50]). The rest of the leukemias written above (1.2.2) also have a low incidence and are different from the four major ones as they involve transformed cells. Acute leukemia forms are among the most common pediatric cancers therefore, the incidence of them by age follows a bimodal distribution. Contrary, chronic leukemia incidences increase with age (see Figure 10).

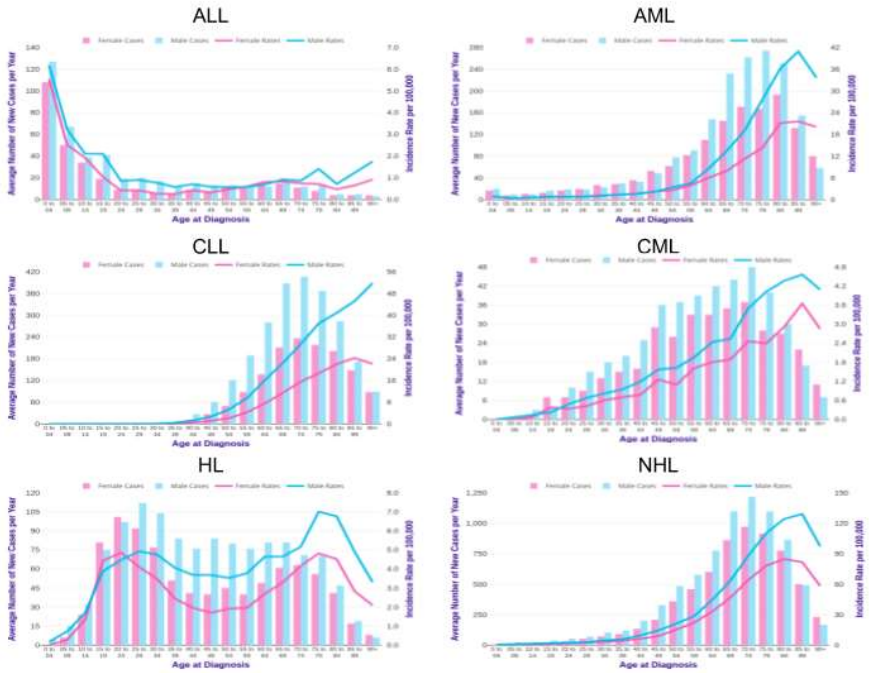


Figure 10. Hematopoietic cancer distributions of incidences by age between 2015-2017 in the UK. Female (pink) and male (blue) cases per 100,000 are differentiated. Data collected by the Cancer Research UK, entry August 2020 source: cruk.org/cancerstats

In hematopoietic cancers with pediatric incidence such as acute leukemias, mortality rate is higher in adults than in children. Concretely, the average mortality of age groups in the UK population according to the data of the Cancer Research UK is 0.517 per 100,000 in adults compared to 0.262 per 100,000 in pediatric patients in ALL and 9.84 per 100,000 vs 0.187 per 100,000 in children in AML. Similar population numbers are given for US cases [162]. However, in general, leukemia one-year diagnosed survival rates have increased from 34% in 1971-1972 to 68.5% in 2010-2011 as well as the five-year survival rates that have also improved 38.5 points [165] over the last years as reported by Cancer Research UK.

Among different populations, the US incidence data shows that the prevalence of leukemias is higher among White-Caucasian people than the rest except for ALL in pediatric patients which is higher among the Hipanic community [162].

The identified risk factors for developing leukemia are [162,166,167]:

- radiation (therapeutic, occupational and wartime-related). It has been observed that ionizing radiation among survivors of the atomic bombs, workers of nuclear plants and radiologists previous to the 50s manifested leukemia at higher rates than other population groups.
- chemical exposures (residential and occupational): high exposures to hydrocarbons in common products, industrial disinfectants and some building materials showed associations with leukemia as well as some pesticides (especially in children)
- chemotherapy: Therapy-related secondary AMLs can occur after chemotherapy employed as treatment of a primary malignancy.
- family history: e.g. people with CLL/SLL relatives have more chances to develop leukemia.
- genetic syndromes: some examples such as Down syndrome, Li-Fraumeni syndrome, Fanconi anemia and Bloom syndrome.
- Infections: There are different examples that support a causality by infection. For example DNA herpesvirus EBV in association with Burkitt's lymphoma or retrovirus human T-lymphotropic virus 1 (HTLV-1) in adult T-cell leukaemia/lymphoma. In addition, the proposed double hit theory by Mel Greaves also states a causal relation between infection and ALL. H e proposed that some ALL are driven by a first *in utero* event which creates pre-leukemic clones that together with a "delayed infection" can cause a second aberrant genetic event and trigger leukemia transformation [168].

1.2.4 Scientific and clinical advances in the history of leukemias

At the beginning of 1800, and during the following 50 years, several contemporary scientists diagnosed and defined leukemia for the first time [169]:

- In 1811 Peter Cullen describes a peculiar case of "milky" blood in a patient with "splenitis acutus" (acute hyperplasia of the spleen)
- In 1825 Alfred Velpeau reported a case of a patient with "pus-filled" blood and enlarged liver and spleen that presented swelling of the abdomen, fever, weakness, and urinary stones (first time description of leukemia symptoms)
- In 1844 Alfred Donné was the first physician to perform a microscopic examination and description of immaturance presented by the white blood cells
- In 1845 John Bennett was the first physician to realize that the accumulation of leukocytes was a primary systemic blood disorder and not a secondary manifestation of other diseases and call it leucocythemia. Two years later, Rudolf Virchow introduced the term leukämie to name a disease with unbalanced red and white cell quantity in blood.

It was not until 1869 that Ernst Neumann connected leukemia origin and the bone marrow as he was one of the first to realize that leukocytes were formed there [170]. There was a great advance in knowledge of this disease by the mid-20th century due to the discovery of several chromosomal abnormalities characteristic of different types of leukemias. Mel Greaves has summarized these advances into clinical and biological in the following timeline ([167]; see Figure 11).

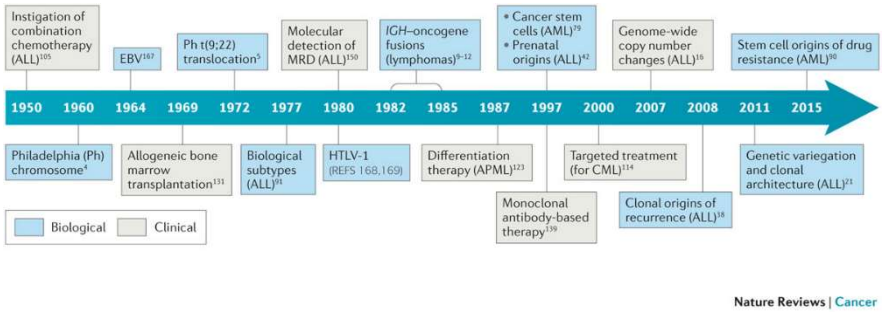


Figure 11. Scientific and clinical leukemia discoveries reprinted from *Leukaemia 'firsts' in cancer research and treatment* by Greaves 2016 Nature Reviews Cancer.

One of the most notorious discoveries was the observation of the Philadelphia (Ph) chromosome in CML by David Hungerford and Peter Nowell who at that time resided in that city [171]. Later, it was identified the translocation of the Ph aberration which involves chromosome 22 and 9 and, following, it was discovered that the fusion gene BCR-ABL1 resulting from the translocation has leukemorgenic power not only in CML but also in ALL. Another important medical advance as treatment for leukemia patients is the allogeneic bone marrow transplantation which implicates the administration of healthy hematopoietic stem cells from a compatible donor. In fact, Dr. E.D. Thomas and his medical team were awarded with the Nobel Prize (1990) for being pioneers in transplants for leukemia patients [167].

1.2.5 Hematopoiesis, lymphoid differentiation and maturation

HSCs are the common ancestors of all the blood cells. Those are rare and quiescent with the ability to self-renew and to differentiate into all blood cell lineages. Hematopoiesis happens during embryonic development at different stages and sites (e.g. HSCs are present in the fetal liver) but, right after birth, HSCs become resident in the bone marrow where hematopoiesis takes place during adulthood [172]. The first population derived from HSC are multipotent progenitors (MPPs) which have also been defined as lympho-myeloid-restricted multipotent progenitors (LMPP) in mouse

which can either give rise to common myeloid progenitor (CMP) or lymphoid progenitor cells [172,173]. Regarding the lymphoid lineage, the studied LMPP can differentiate into a population called early lymphoid progenitors (ELPs) which, in turn, differentiate into the thymic early T-cell-lineage progenitors (ETPs) or into common lymphoid progenitors (CLPs) of the bone-marrow. At this point of lymphopoiesis, B and T lineages differentiation pathways separate depending on stimuli on CLPs [174] (see Figure 12). However, a common biologic process between the two is the expression of recombinase activating gene proteins RAG1 and RAG2 which initiate rearrangement at the immunoglobulin heavy chain (IGH) locus in B and also triggers the T-cell receptor gene rearrangement which are necessary to create the diversity of Immunoglobulins (Igs) and T-cell receptors (TCR).

B-Cell Development

CLPs committed to B-lymphoid lineage are called pro-B-cell which start to express several markers of differentiation. In the next step, expression of CD19 marks the pre-BI-cell population which completes the gene recombination in the heavy chain locus. This locus is present in segments that code for the variable (V), diversity (D), joining (J), and constant (C) regions [174]. RAG1 and RAG2 are responsible for the cleave and shuffled bind of the VDJ genes to create the diversity of IGH. When expression of the RAG genes halts, the IGH assembles with the IG light chains (IGLs; also previously rearranged in a similar way) to form the pre-B-cell receptor (pre-BCR). The presentation of pre-BCR serves as a check-point for selection of those cells that have Ig functional. The signals generated by the pre-BCR triggers a clonal expansion (positive selection process) of the harboring cells now called pre-BII cells. Further rearrangements of the light chain follow to create a complete assembled BCR which are carried by the denominated immature B-cells [173,174]. Those undergo a second check-

point (negative selection process) in which if they bind to a self-antigen with the BCR are either killed or inactivated. All the maturation steps previously explained take place at the bone marrow (see Figure 12). The resulting cells are mature naive B-cells which migrate into the lymph node and spleen for further differentiation. In the inter-follicular region of those, they may develop into short-lived plasma cells or enter the germinal center (GC) within the follicle where somatic hypermutation and heavy chain class-switching takes place. These processes transform the cells into long-lived plasma cells and memory/marginal zone B cells [175] (see Figure 12).

T-Cell Development

Some CLPs migrate to the thymus and become ETPs. Even though it is widely accepted that lymphocytes come from lymphoid committed precursors characterized as the CLPs there is also evidence that ETPs retain myeloid potential [176,177]. In any case, these progenitors reach the cortex of the thymus lacking the mature T-cell markers CD4 and CD8 and start their differentiation process. They undergo 4 stages of differentiation (DN1,DN2,DN3,DN4) detected by the combination of expression of two different markers (CD25 and CD44) but still retaining the double-negative (DN) phenotype for CD4 and CD8 (see Figure 12). When cells reach DN3 stage, they initiate rearrangement of TCR loci and expression of pre-T cell receptor (pre-TCR) formed by an already rearranged β-chain and invariant/surrogate α-chain [172,178]. Pre-TCR signals initiate proliferation of DN4 and induce the co-expression to double-positive (DP) CD4/CD8 stage. After, the TCRA gene is finally rearranged to get a complete TCR. The TCRs are exposed to the major histocompatibility complex (MHC). Active interaction positively selects the T-cells into CD8 or the CD4 positive, depending on whether they recognize MHC class I or MHC class II respectively [178]. Another check-point (negative selection

process) regarding potential autoreactivity of thymocytes purifies the final TCR repertoire [172].

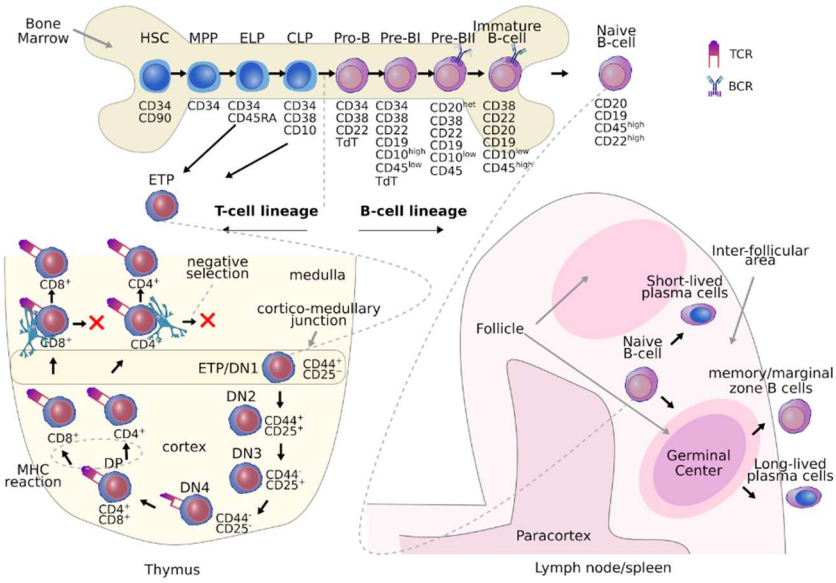


Figure 12. Lymphoid hematopoiesis. Figure inspired by The Biology of Acute Lymphoblastic Leukemia Carroll et al., 2011 Childhood Leukemia: A Practical Handbook and Lymphoid Hematopoiesis and Lymphocytes Differentiation and Maturation Cavalheiro et al., 2017 IntecOpen. Cell illustrations are taken from https://reactome.org/icon-lib

1.3 Acute lymphoblastic leukemia

ALL is a disease caused by a maturation arrest and high proliferation of the lymphoid progenitor/precursor cells also called lymphoblasts or just blasts in the bone marrow, blood and extramedullary sites [179]. As explained above (1.2.3), acute lymphoblastic leukemia is one of the four major leukemia types. In general, ALL accounts for 1.5% of all cancers [50]. It is more prevalent in children than in adults (75% children vs. 12% in adults of all leukemias [50]). Most pediatric patients respond well to treatment achieving a 5-year overall survival rate between 85% - 90% [180,181] whereas in adults it is 40% approximately [179,180]. Despite the

improvements in treatment, relapse ALL forms (15-20% of pediatric and 40-75% of adult patients recur; [182]) presents a discouraging prognosis to the point of becoming the second cause of cancer-related mortality among pediatric patients [183].
Clinically, ALL is differentiated from LBL when there are >20% blast cells in the initial diagnostic aspirate of the bone marrow [184]. For example, cases of T-LBL manifest with enlargement of mediastinum due to the thymus and little dissemination of T-lymphoblasts in blood whereas T-ALL present more than 20% of blasts cells infiltrated in bone marrow independently of the thymus involvement. In fact, approximately only ⅓ of the T-ALL present mediastinal masses, the rest of them lack evidence of it and normally correlate with increased circulating blasts in the bloodstream [185].

A first morphological inspection to the cells of the aspirate can differentiate ALL from AML but as explained hereunder other checks must be performed to fully characterize which type of ALL is presented.

1.3.1 Subclassification of the disease: B-cell ALL and T-cell ALL similarities and differences

There are two main types of lymphoblasts: T-cell and B-cell lineage, therefore, we can distinguish B-ALL and T-ALL disease forms. The first is more prevalent both in adults and children (75% B-ALL vs. 25% T-ALL in adults and 85-90% B-ALL vs. 10-15% T-ALL in pediatrics; [186,187]).

B-ALL cases commonly present fever, caused by neutropenia (low levels of neutrophils) and infection, fatigue due to anemia and bleeding at mucocutaneous as a result of thrombocytopenia (low levels of platelets) [188]. Other typical clinical manifestations are enlargement of lymph nodes

(lymphadenopathy), the spleen (splenomegaly) and the liver (hepatomegaly) due to the infiltration of lymphoblasts [189]. Furthermore, the infiltration and intramedullary growth of blasts in the bone marrow can cause bone pain. Other organs that can be affected by infiltration are meninges, testes and ovaries and the central nervous system. Patients diagnosed with T-ALL often suffer the same described symptoms as B-ALL plus, as mentioned before, mediastinal thymic masses and also tend to present lower degree of leukopenia (low white blood cell count) [188].

Accurate diagnosis of ALL implied standardized guidelines to classify it which comprehensively tackles different aspects such as morphology, immunophenotype, (cyto)genetics and genomics.

Morphology

The French-American-British (FAB) morphological classification of ALL is based on the following features of the blasts [186]: cell size, nuclear shape, prominence of nucleoli and the amount and appearance of cytoplasm (degree of basophilia, presence of cytoplasmic vacuolation). There are three groups:

- L1: small homogenous cells with a regular nuclear shape but its contents not clearly visible. It also presents a moderate basophilic cytoplasm. Most patients have lymphoblasts fitting this description, especially children.
- L2: large heterogeneous cells with irregular nuclear shape accompanied by a cleft in the nucleus. Large and prominent nucleoli. There is also heterogeneity in cytoplasm colours and moderate abundance of it. This appearance of the blasts is more common among old patients.
- L3: large cells but homogeneous in size. Also homogeneity in nucleus shape with oval-to-round form and a prominent nucleoli.

In addition, these cells present evident cytoplasmic basophilia and vacuolation. Usually, L3 blasts express mature B- lymphoid markers. These blasts are detected in patients with leukemia secondary to Burkitt's lymphoma.

Immunophenotype

The WHO further classifies ALL into different subgroups according to the markers (proteins or glycoproteins) that are either in the cell surface or cytoplasm of the lymphoblasts which are detectable when using flow cytometry. The monoclonal antibodies that bind to those markers and emit light signals during the immunophenotyping process have been grouped into Clusters of Differentiation (CD) [173]. This classification has been defined by European Group for the Immunological Characterization of Leukemia (EGIL) [184,190] (see Table 3).

Precursor B-cell leukemia (HLA-DR+, TdT+, CD19+, and/or CD79a+, and/or CD22+, and/or CD34+)	
Pro B-ALL (B-I)	CD19+ CD79a+ cCD22+ (comprises 10% of adult ALL patients)
Common ALL (B-II)	CD10+ (comprises 50% of adult ALL patients)
Pre B-ALL (B-III)	CD10+ cIg (comprises 10% of adult ALL patients)
Mature B-ALL (B-IV)	sIg+ kappa or lambda (4% of adult ALL patients)
Precursor T-lymphoblastic leukemia (TdT+,cCD3+ and CD34+)	
ETP T-ALL	CD5+ CD7+ CD117+ CD11b+ CD65+ HLA-DR (CD13 and CD33 myeloid markers). 15% of pediatric and 35% of adult patients of T-ALL cases
Pro T-ALL (T-I)	CD7+ (7% of adult ALL)
Pre T-ALL (T-II)	CD2+ CD5+ CD7+
Cortical T-ALL (T-III)	CD1a+ sCD3+ CD2+ CD5+ CD7+ CD4+ CD8+ (17% of adult ALL)
Mature T-ALL (T-IV)	sCD3+ CD2+ CD5+ CD7+ CD4+ CD8+ (1% of adult ALL)

Table 3. ALL immunophenotypes. There are only listed the positive expressed markers that characterize each type. A “c” means cytoplasmic and an “s” surface; if not indicated assume surface expressed marker. The table is a summary from Hoelzer et al., 2016 Annals of Oncology and Abdul-hamid et al., 2011 IntechOpen and Follini et al., 2019 International Journal of Molecular Sciences.

The immunophenotypes of the lymphoblasts reflect the developmental step of the maturation arrest of the lymphoid precursors (see Figure 12 again). For example, lymphoblasts in ETP T-ALL are double negative for CD4 and CD8 and show a very close transcriptional program with early T-lineage progenitors (ETP) [191].

(Cyto)genetics

Recurring gross chromosomal rearrangements that alters the regulation of key genes and aneuploidies[13] are common across ALL and define different subtypes (see Table 4). Also common transcriptional programs define these subtypes, sometimes triggered by the same genetic aberrations that characterize them. Furthermore, since some of the genes dysregulated are actually genes involved in lymphoid development there is a correlation between the maturation arrest and the aberrations presented [192] as well as some enrichment of certain leukemogenic driver alterations on each of these subtypes [193]. Mainly, in B-cell lineage, rearrangements cause chimeric fusion genes that involve transcription factors of hematopoietic development, epigenetic modifiers, tyrosine kinases and cytokine receptors which act as oncogenes whereas in T-cell lineage alteration in the expression of genes (such as those from the groups mentioned above) are caused by the influence of the resulting misplaces regulatory regions of TCR.

Lineage	Subtype	Characteristic Aberration	Genes affected	Frequency (%)	
				adult	children
B-ALL	Hyperdiploid	Hyperdiploidy with more than 50 chromosomes	-	7	20-30
	Hypodiploid	Hypodiploidy with less than 44 chromosomes	-	2	2-3

[13] Aneuploidies: gains and losses of entire chromosomes

	ETV6-RUNX1	t(12;21)(p13;q22) translocation	ETV6-RUNX1, TEL-AML1	2	15-25
	TCF3-PBX1	t(1;19)(q23;p13) translocation	TCF3-PBX1, E2A-PBX	3	2-6
	BCR-ABL1/Ph positive	t(9;22)(q34;q11.2) translocation	BCR-ABL1	25-30	2-5
	Ph-like	Ph- but same transcription profile		20-25	10-16
	CRLF2 rearrangements	CRLF2 rearrangements	IGH-CRLF2, P2RY8-CRLF2	10-12	5-7
	MLL rearrangement	t(4;11)(q21;q23) translocation	MLL-AF4	?	1-2
	MYC rearrangements	t(8;14)(q24;q32), t(2;8)(q12;q24), t(2;8)(q12;q24) translocations	MYC	4	2
	DUX4 rearrangements	t(4;21)(q35;q22) t(4;14)(q35;q32)	ERG-DUX4 IGH-DUX4	5.4	7
	PAX5 rearrangements	PAX5 rearrangements	PAX5	7-9	2-9
	iAMP21	Intrachromosomal amplification of chromosome 21	-	0.3-2.1	2.5
T-ALL	TAL1 dysregulation	t(1;7)(p32;q35) and t(1;14)(p32;q11) translocations, del(1)(p32p32), small insertion→ de novo enhancer	TAL1	12-25	15-18
	LMO2 dysregulation	t(11;14)(p15;q11) translocation and 5' LMO2 deletion	LMO2	1-6	10
	TLX1 dysregulation	t(10;14)(q24;q11) and t(7;10)(q35;q24) translocations	TLX1 [HOX11]	30	5-10
	TLX3 dysregulation	t(5;14)(q35;q32) and t(5;14)(q35;q11) tran slocations	TLX3 [HOX11L2]	5	20-25

	HOXA dysregulation	(10;11)(p13;q14) and t(11;19)(q23;p13) translocations del(9)(q34;q34)	PICALM-MLLT10, MLL-MLLT1, SET-NUP14	20	10
	ABL1 dysregulation	Episomal amplification of ABL1 9q34 amplification encoding t(9;12)(q34;p13) t(9;14)(q34;q32)	NUP214-ABL1 ETV6-ABL1 EML1-ABL1	5-6	6
	NKX2-1/NKX2-2 dysregulation	t(14;14)(q11;q13) translocation t(14;14)(q13;q32) translocation +others	NKX2-1 NKX2-2	6-8	8

Table 4. ALL subtypes defined by gene expression profiles and recurrent aberrations. Summary table inspired from from Hunger & Mullighan 2015, Blood, Ustwani et al., 2016 Critical Reviews in Oncology/Hematology, Girardi et la., 2017 Blood, Gu et al., 2019 Nature Genetics, Belver & Ferrando 2016, Nature Reviews Cancer, Van Vlierberghe 2012 J Clin Invest.

Apart from the summary of Table 4 there are some other worth mentioning characteristics of ALL subgroups [194]:

- Hyperdiploid: this is one of the subgroups with better prognosis and it also has more incidence in adolescents than in adults [195]. The ploidy gains usually happen in these chromosomes: X, 4, 6, 10, 14, 17, 18 and 21 being trisomies and tetrasomies the major aberrations (over 75 % of patients with hyperdiploid subtype). The project of Paulsson et al., 2015 from Lund University revealed recurrent mutations in Ras pathway genes as well as histone modifiers [196]. Interestingly, through the study of monozygotic twins the lab of Mel Greaves showed that the hyperdiploidy condition is acquired prenatally in Pre-B cell *in utero* [197].

- Hypodiploid: in this case, chromosomal loss is associated with a bad prognosis. Some [198], further differentiate patients into near-haploid (24-31 chromosomes) and low-hypodiploid (32-39 chromosomes). Both show activating mutations in Ras pathway and PI3K signalling too. Mutations in TP53 are also very common.
- ETV6-RUNX1 type: patients in this group have good prognosis. Both genes (ETV6 and RUNX1) are involved in normal hematopoiesis [199]. There is evidence, again revealed by Greaves lab, that the fusion has a pre-leukemic origin as it has been seen to happen prenatally in monozygotic twins that required the acquisition of other postnatal genetic alterations to develop ALL [200]. Evidence suggests that these second events cooperating with the fusion are a consequence of aberrant RAG recombinase activity [201]
- TCF3-PBX1 type: It is more prevalent among African-Americans. In general, this one has a good prognosis. It is associated with pre-B immunophenotype [202].
- BCR-ABL1 type: can also be found as Ph or Ph+ in the literature since it refers to having the Philadelphia chromosome. Patients with this translocation have a dismal prognosis but there are some improvements due to incorporation of tyrosine kinase inhibitors in treatments. Recurrent deletions of IKZF1, most likely coming from aberrant RAG-activity, are associated with this subgroup [203].
- Ph-like: the name of these was given when they discovered that there were patients with similar gene expression profile to Ph positive patients but lack BCR-ABL1 fusion [204]. It is considered a high-risk group [205].
- CRLF2 rearrangement: this category partially overlaps with Ph-like subtype. Almost 50% of the cases of Ph-like ALL have rearrangements in the cytokine receptor-like factor 2 (CRLF2). It

is also common among ALL patients with Down Syndrome. The majority of CRLF2-rearranged cases co-occur with mutations in the JAK-STAT pathway [195].

- MLL-rearrangement type: rearrangements in MLL gene (former KMT2A) such as the resulting MLL-AF4 is very common within infant ALL patients (60% of infant patients younger than one year), with special incidence on those with less than 6 months of age [193]. In general, has a poor prognosis.
- MYC rearrangement: dysregulation of MYC by rearrangements is also common in Burkitt cell leukemia/lymphoma and therefore, also correlates with L3 morphological type. Prognosis has been reported to be poor [206] and favorable [194].
- DUX4 rearrangements: approximately 7% of B-ALL cases have a distinct gene expression profile that includes DUX4 rearrangements. Among these cases 50% to 70% have focal deletions in ERG too [207]. In addition, transcriptional deregulation of ERG in this subtype can happen due to the expression of an ERG isoform (ERGalt) that inhibits the wild-type ERG function [208]. Despite having recurrent alterations in IKZF1, this group presents good prognostic.
- PAX5 rearrangements: PAX5 gene plays a role in both lymphoid lineages. In B-cells, it is a key player for the cells to commit to the lineage. There are many different types of alterations affecting PAX5 but the chromosomal translocation ones involve a great variety of gene partners. A recent study [209] has redefined B-ALL subtypes an differentiates two groups of PAX5 alterations with different gene expression profiles: “PAX5alt” meaning PAX5 alterations (rearrangements, intragenic amplifications or mutations) and PAX5P80R referring to PAX5 aminoacid change p.Pro80Arg and biallelic PAX5 alterations. According to this

study, both subgroups of PAX5 in all patient ages (adults and children) presented intermediate to poor outcomes.

- iAMP21: This name refers to intrachromosomal amplification of chromosome 21. It is characterized by at least 3 gain copies of large regions of chromosome 21 creating dysregulation of the genes within such as RUNX1 [195]. Overall has a poor prognosis both in children and adults [210].
- Other B-ALL subtypes defined are: MEF2D rearrangements and ZNF384 rearrangements. The first usually partners with BCL9, HNRNPUL1, DAZAP1 and SS18 and overall is considered a high-risk new subgroup [211] whereas the second usually partners with EP300, TCF3 and TAF15 and the clinical prognosis seems to depend on the partners [212]. A very rare subtype that has also been described is characterized by IL3-IGH [207]. Recently, there have been detected some B-ALL cases with similar transcriptional profile as ETV6-RUNX1 but negative for the translocation that are a new subtype called ETV6-RUNX1-like [207].
- TAL1 dysregulation: TAL1 gene is a regulator of hematopoiesis. Aberrant expression of it is found in 60% of the T-ALLs in children and 45 % in adults but not for all of these cases is possible to find molecular evidence of causality [192,213] (e.g. 16%-30% rearrangement STIL-TAL1 and 3% translocation t(1;14)(p32;q11) in children). TAL1 increased levels are believed to dysregulate members of T-cell specific lineage so cell differentiation is halted [42]. Samples with TAL1 overexpression profile are associated with late cortical immunophenotype [192]. In general, the subgroup presents a good prognosis in children.
- LMO2 dysregulation: Aberrant expression of LMO2 sometimes overlaps with TAL1 overexpression too [192]. Apart from TAL1, LMO2 aberrant expression can co-occur with LYL1 dysregulation

and has also been considered another subgroup [214]. Although alterations affecting this gene are found in 10% of pediatric T-ALL, there are 45% of cases with LMO2 expression dysregulated suggesting different activating mechanisms [215]. In general, it has a favorable outcome.

- TLX1 dysregulation: Also known as HOX11, is upregulated when translocations place the control of the gene under TCR enhancers. It has been shown that downregulates the rearrangement and expression of the TCRA locus which ends up in thymocyte arrest in cortical development [178]. Genes associated with TLX1 are involved in cell growth and proliferation so, since most treatment drugs affect proliferation that might explain the favorable outcome of this subtype [192].
- TLX3 dysregulation: This gene is the former HOX11L2, its usual fusion partner is BCL11B but also its ectopic expression can be regulated by TCR enhancer. This subtype is associated with WT1 mutations and early cortical development phenotype [214]. It presents a dismal prognosis.
- HOXA dysregulation: This group involves different genomic aberrations that alter the expression of the HOXA genes, especially HOXA9 and HOXA10. Dysregulation of HOXA genes is common among ETP-T-ALL immunophenotypes [178].
- ABL1 dysregulation: Not only is involved in B-ALL but also in T-ALL. NUP214-ABL1 is the most common fusion affecting ABL1 expression in T-cell lineage, however, the mechanisms (amplified episomes and intrachromosomal amplification) to overexpress ABL1 are different than in Ph positive patients (translocation) [216]. Also the oncogenic power of NUP214-ABL1 is not sufficient to drive leukemogenesis so other alterations are required [214].

- NKX2-1/NKX2-2 dysregulation: In a study from the Netherlands in 2011 [217], they identified another subgroup of unclassified-subgroup samples which shared dysregulation of either NKX2-1 or its paralog NKX2-2 by different genetic aberrations. Both are believed to develop a similar oncogenic role in T-ALL and are associated with early cortical development arrest [215].
- Others: In that same study from the Netherlands mentioned above [217], they also described another group of samples with distinct expression signature. They observed deletions in del(5)(q14) that cause upregulation of MEF2C. This subtype is associated with immature stages of development so it is believed to play a role in the regulation of the genes in early stages of thymocyte differentiation. It also presents enrichment in CDKN1B deletions [218] and overall is associated with very poor prognosis [215]. Other subtypes less frequent are characterized by dysregulation of TAL2, LMO1, NKX2-5, MYC and MYB [215]. Since the immunophenotype group ETP T-ALL shows a different gene expression profile is also considered as a genetic subtype together with the ones mentioned above. In fact, ETP is characterized by high frequency of JAK mutations and low frequency of NOTCH1 mutations compared to the other subtypes. It is considered a high-risk group and is usually associated with treatment resistance [219].

1.3.2 Primary Genomics of ALL

Along with the dysregulation of gene expression due to the rearrangements, there are recurrent genes and pathways altered that contribute to leukemic transformation and proliferation. There are substantial differences between the B and T-ALL driver alterations. Some of those are enriched in particular

immunophenotypes and tend to co-occur with certain rearrangements (summary at Figure 13).

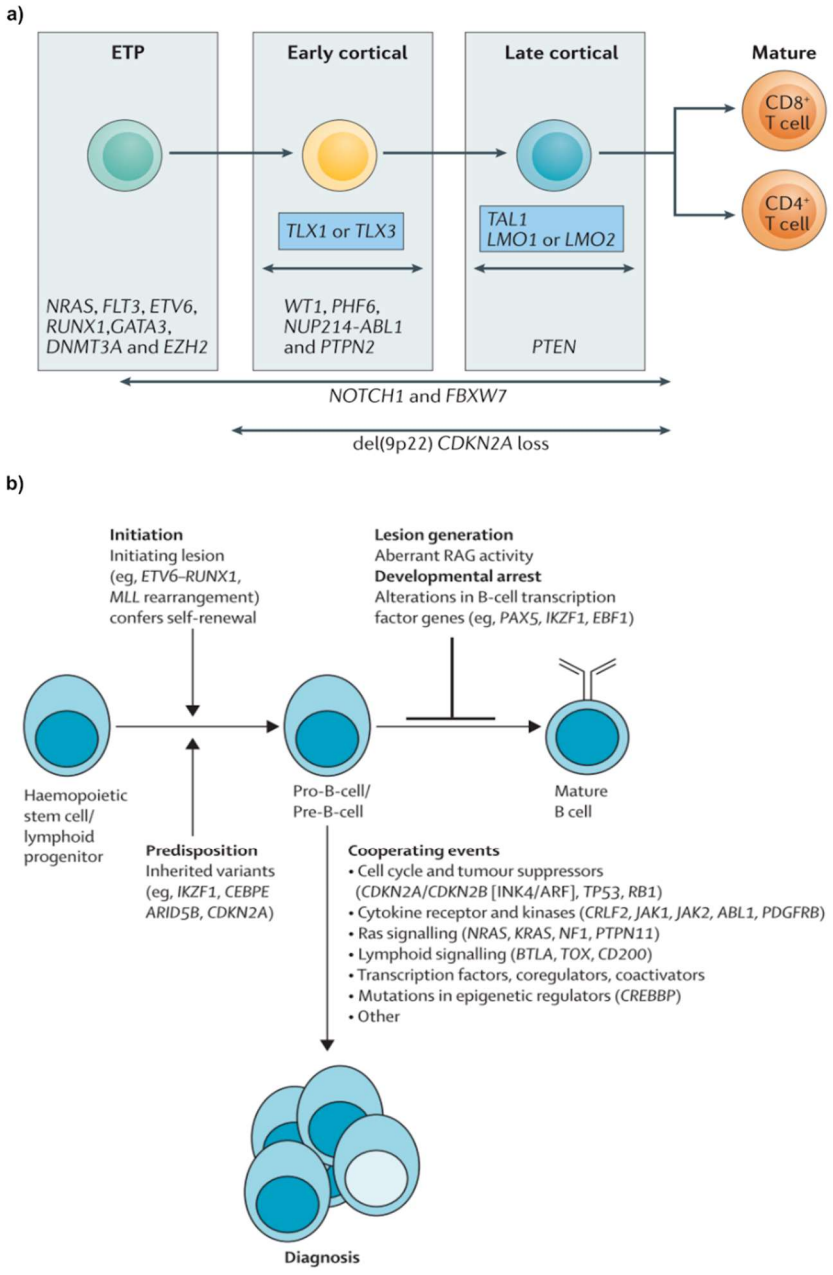


Figure 13. Prevalence of genomic alterations at each immunophenotype. a) T-cell lineage reprinted from *The genetics and mechanisms of T cell acute lymphoblastic*

leukaemia by Belver & Ferrando 2016 Nature Reviews Cancer and b) B-cell lineage adapted from *Acute lymphoblastic leukaemia* by Inaba et al., 2013 The Lancet.

1.3.2.1 B-ALL driver alterations

Transcriptional regulation of lymphoid development

IKZF1 (IKAROS Family Zinc Finger 1): The resulting protein from this gene is a transcription factor that acts as a chromatin remodeler. During lymphoid development, Ikaros protein allows chromatin accessibility which is necessary for V(D)J recombination and regulates the expression of B-cell-specific genes [173]. In a study of 2008 [203], they described recurrent deletions in this gene in B-ALL adult and pediatric patients (61 out of 304) especially in Ph positive ones which presented more than the 80% of this subtype cases with deletions. A closer look at the breakpoints of these deletions suggested an aberrant RAG-mediated recombination activity. The follow-up study of the same group, revealed a poorer prognosis outcome of the Ph positive and negative cases harboring *IKZF1* deletions [220]. It commonly co-occurs with CRLF2 rearrangements [194]. Some germline variants have been described to create predisposition to ALL [221].

Other members of the Ikaros family have also been found altered in B-ALL such as IKZF2 and IKZF3 especially in hypodiploid ALL subtype [198].

PAX5 (Paired Box 5): This gene is another regulator of B-cell lymphoid development which encodes for a transcription factor that represses necessary components for T-cell lineage and drives precursors to B-cell commitment, activates BCR signalling modulators and also plays a role in the maintenance of mature B-cell state [222]. It is altered in 30% of B-ALL [194] and mutated cases have recently been divided into two new ALL

subtypes [209] as referred above (1.3.1): PAX5alt and PAX5 P80R. The last refers to a group of patients that commonly present a different genetic profile and all share the same missense mutation p.Pro80Arg in the paired box DNA binding domain that constitute a hotspot in it whereas the first englobes different types of alterations such as rearrangements, sequence mutations and focal intragenic amplifications. According to the study, both types accounted for 9.7% of those cases that were previously unclassified. In addition, a recurrent germline mutation (p.Gly183Ser) has been described to confer ALL susceptibility [223].

EBF1 (EBF Transcription Factor 1 or Early B Cell Factor 1): This gene, together with TCF3 (E2A), regulates the expression of genes of the B-Cell lineage and it is essential for the rearrangements of the loci IgH and IgL [222]. PAX5 and EBF1 regulate each other in an auto-regulatory loop [222]. Focal deletions of this gene leading to haploinsufficiency arrests cells at pre-pro-B-cell stage and suggests a contribution to leukemogenesis [224].

RUNX1 (RUNX Family Transcription Factor 1) encodes for the protein known as acute myeloid leukemia 1 protein so, as the name suggests, it is involved in both acute leukemia forms plus other hematopoietic malignancies. In mice, it is expressed to trigger the transformation of vascular endothelium cells to primitive HSCs during embryonic development but it is not necessary for the maintenance of long-term HSC in adult hematopoiesis [222]. 25% of B-ALL cases present dysregulation of RUNX1 by the chimeric fusion with ETV6 due to the translocation between chromosome 12 and 21 [201]. Although this fusion is required but not sufficient for leukemic transformation [225], the oncogenic dysregulation of ETV6-RUNX1 happens as RUNX1 binds to its target sequences and the recruitment of ETV6 partners inhibits their transcription

[226]. ETV6 loss-of-function mutations are also common among B-ALL both in ETV6-RUNX1 and non-rearranged ETV6 [227,228].

Other leukemic driver transcription factors of this category are LEF1 (Lymphoid Enhancer Binding Factor 1) which is a mediator of the WNT signalling and has been associated with leukemic transformation [229] and ERG (ETS Transcription Factor ERG) previously mentioned at 1.3.1.

Tumor suppression and cell cycle regulation

TP53 which acts as tumor suppressor in many cancers [82] predicts a very dismal outcome in ALL patients harboring alterations in it [230]. The majority of the alterations (disruptive mutations and deletions) tend to locate in exons 7 and 8 and it is more prevalent (>90%) in hypodiploid subtype in both adult and pediatric patients [198]. In this B-ALL subtype, not only TP53 is intriguing recurrent but also RB1 [198]. This gene is called Retinoblastoma (RB) Transcriptional Corepressor 1 and encodes por pRB which is another tumor suppressor that acts as the key regulator of the entrance to cell cycle which is also widely altered in different cancers [82]. Deletions of the Cyclin Dependent Kinase Inhibitors 2 (CDKN2A and CDKN2B) can also be found in B-ALL [203] and they encode for p16INK, p14ARF and p15 which are involved in G(1)-S cell cycle transition as they inhibit CDK4 and CDK6 that inhibit pRB.

Cytokine receptors, kinases and RAS signaling

CRLF2 is part of the cytokine interleukin-7 (IL7) receptor which activates intracellular signalling through Janus kinases (JAK1-3) that consequently activate STAT transcription factors (JAK-STAT pathway) [231]. Cytokine receptors are linked not only to the JAK-STAT pathway but also to RAS pathway and PI3K-AKT pathways, all of them activating gene expression

programs [232]. Dysregulation of these players can alter the normal functions of the cell and contribute to leukemogenesis. For example, in Ph-like ALL it has been observed IL7RA (IL7 receptor alpha) insertions and deletions, chimeric fusions of JAK2 due to deletions and translocations, rearrangements of the erythropoietin receptor (EPOR-R) and mutations in SH2B3 as alterations that keep JAK-STAT signalling activated [233]. Regarding RAS signaling, the pathway is abnormally activated by point mutations in NRAS and KRAS genes, as well as, loss-of-function mutations of negative modulator NF1, upstream regulator PTPN11 and kinase receptor FLT3 which activating mutations have been detected [181]. In pediatric patients, approximately around 30% have mutations in the RAS pathway [234]. Other overexpressed kinases due to fusions are the "ABL1 type" (ABL1, ABL2, CSF1R, and PDGFRB) which are also common in Ph-like ALL [233] and obviously BCR-ABL1 positive.

Epigenetic regulators

Aberrant acetylation and methylation contribute to changes of the transcriptome that can drive important leukemic consequences [195]. It is very common to find alterations in DNA and chromatin modifiers in different ALL subtypes, some of them enriched in relapse as mentioned below. For example, missense mutations in WHSC1 (NSD2) which is a histone methyltransferase are found in ETV6-RUNX1 [201] and CREBBP known as CREB-binding protein, a H3K18 and H3K27 acetylase is recurrently mutated among hypodiploid cases at diagnosis [198]. Also histone methyltransferase MLL family members and chromatin remodeling genes of the SWI/SNF complex such as ARID family are common in B-ALL [181]. Another recurrent altered gene is histone acetyltransferase EP300 [181].

Other recurrent genes

BTLA (B And T Lymphocyte Associated) and a type I membrane glycoprotein called CD200 are involved in lymphoid signaling and are recurrently deleted [220]. There are transcriptional cofactors such as TBL1XR1, BTG1 and NCOR1 that present deletions in B-ALL [234]. Hotspot mutations in ZEB2 have also been reported [235]. Another commonly deleted gene but less studied is ADD3 adducin gene [234,236].

1.3.2.2 T-ALL driver alterations

NOTCH1 pathway

NOTCH1 is a transmembrane type I protein of the NOTCH family. In children with ALL, it is found mutated in around 60% of the cases (e.g. 56.2% in Weng et., 2004 [237]) and in adults it seems to vary from publication to publication (53% Neumann et., 2014 [238] vs. 86% Kim et al., 2020). Anyway, it is one of the most prevalent mutated genes in T-ALL. This cellular receptor has an extracellular and intracellular domain facilitating transduction of external signals into transcriptional changes in the cell. In mammals there are 4 homolog NOTCH genes (NOTCH1-4) being NOTCH1 the one playing a major role in leukemogenesis [172]. Notch signalling is critical for prenatal hematopoiesis and postnatally it is expressed during thymocyte development and determines T-cell fate specificity [239]. Concretely, it is involved in the progression of cortical thymic developmental stages (DN1-3, see Figure 12), therefore, it is not surprising to find it highly mutated in lymphoblasts with early (Pre, Pro T-ALL) and late cortical immunophenotypes [240,241]. The oncogenic power of NOTCH1 pathway in T-ALL is its constitutively active signaling. The most common altered form of Notch1 is a truncated protein due to gain-of-function mutations and deletions in specific domains that produce the active intracellular part of the receptor (ICN1) which translocates to the nucleus

and constitutively triggers upregulation of PI3K/Akt/mTOR, c-myc, and NF-kb [178,240]. The most affected domains, HD and PEST, in NOTCH1 are those which, once altered, create the active Notch1 form [237]. The first is actually divided into two HD^N and HD^C, these are the heterodimerization domains that hold the extracellular and intracellular subunits linked, therefore, mutations in it spares Notch1 from being cleaved by proteases to release the active ICN1 form. The second domain, is called proline (P) glutamate (E) serine (S) threonine (T) rich (PEST), which contains a degron for the proteasome-dependent degradation, is responsible of the stability of the active ICN1 and, thus, mutations impairing the degron recognition contribute to the overactivation of the pathway. Related to that, another contributing factor in maintaining NOTCH1 active is the presence of loss-of-function mutations in FBXW7 which encodes for a subunit of the E3 ubiquitin protein ligase complex that, among others, regulates NOTCH1 stability [241]. Mutations in FBXW7 can be found between 15 to 25% of the T-ALL patients [241,242]. Another case of constitutively activation of NOTCH1 is the chromosomal translocation t(7;9)(q34;q34.3) which generates a truncated form of NOTCH1 and can be found in less than 1% of the patients [178]. The prognostic impact of NOTCH1/FBXW7 mutations seems to be favorable with early response to treatment in pediatric and adult patients but seems to lack association with a good long-term outcome and, moreover, differences in therapy protocols suggest a dependency of NOTCH1 prognostic value to the intensification of the treatment [240].

Proto-oncogenes expressed by rearrangements

As mentioned before, both T and B-ALL are characterized by rearrangements and gene expression profiles determining certain subtypes. Due to these rearrangements, transcription factor oncogenes are aberrantly expressed under the control of strong enhancers of the TCR loci [215].

These genes are: LIM-only domain (LMO) genes, HOX genes such as TLX1(HOX11), TLX3(HOX11L2), HOXA cluster genes, class II bHLH transcription factors (TAL1, TAL2, LYL1, BHLHB1), NKX2.(1,2,5) and also MYC and MYB [178]. In addition, some of them can also be overexpressed by translocations with other non-TCR partners or by other lesions such as deletions.

TAL1 overexpression creates an upregulation of a positive feedback loop with GATA3, RUNX1 (see below) and MYB. The latest is associated with developmental arrest since it has been observed that interference with MYB activity affects differentiation [243]. LMO1 and LMO2 are frequently co-expressed with TAL1 and LYL1 suggesting a cooperative role in T-cell leukemogenesis [178].

HOX genes also play a role in T-ALL. For example, evidence suggests that TLX1 contributes to the blocking of T-cell differentiation by dysregulation of mitotic checkpoint machinery and thus promoting aneuploidy events caused by missegregation of chromosomes [244]. Regarding TLX3, it seems that TLX3 and TLX1 have a lot of overlapping set of target genes and their transcriptional signature resembles [178]. Both are mainly found dysregulated in early cortical stages and co-occur with mutations in tumor suppressors such as protein tyrosine phosphatase non-receptor type 2 (PTPN2), Wilms tumor 1 (WT1) and PHD finger protein 6 (PHF6) (see Figure 13, more information below; [215]).

The transcription factor MYC contributes to cell growth and proliferation downstream NOTCH1 signaling. Besides, its protein stability also depends on FBXW7. Therefore, alterations in the NOTCH1 pathway transducts in upregulation of MYC and acts as a driver of leukaemia-initiating activity [178].

Tumor suppression and cell cycle regulation

Somatic mutations in tumor suppressors such as ETV6, GATA3 and RUNX1 are abundant in the immature ETP immunophenotype and adult patients (between 8 to 14 % frequency [214]). ETV6 is essential for the development of HSC, GATA3 is an important regulator of development of T-cell progenitors and, as previously mentioned in 1.3.2.1, RUNX1 is crucial to hematopoiesis [215]. Other altered tumor suppressors that can be found in 10-15% of T-ALL are BCL11B (Kruppel-like C2H2-type zinc finger transcription factor), LEF1 (lymphoid enhancer factor/T cell factor) and WT1 (Wilms Tumor 1 transcription factor) [178].

In T-ALL, as well as in B-ALL, dysregulation of cell cycle (a hallmark of cancer) can be achieved by altering CDKN2A and/or CDKN2B. However, in the T-cell type, around 70% of the cases present deletions in these loci [178]. Together with NOTCH1, these are the most altered genes in T-ALL with an incidence of more than 50% of the cases having at least one of NOTCH1 or CDKN2A/B loci affected [214]. Other deleted genes altering the cell cycle are RB1 and CDKN1B (12% frequency) and translocations affecting CCND2 which is another cell cycle regulator which have also been observed in a few cases (~1%) [215].

JAK-STAT pathway and RAS signaling

As indicated above (1.3.2.1), Janus kinases are activated either by IL7R-activating mutations that transduce in phosphorylation of JAK1 and JAK3 or activating mutations in these two (frequency of each one of these genes ranges from 5 to 12 % T-ALL [214]). In both cases, there is activation of STAT5 which transcriptionally regulates proliferation and survival especially in ETP T-ALL. Other alterations affecting this pathway but less

frequent are ETV6-JAK2 fusion, loss-of-function mutations in DNM2 or SH2B3 and deletions in PTPN2 [214].

Similarly as in B-ALL, also some T-ALL patients have NRAS and KRAS activating mutations (e.g. K-Ras^{G12D}) especially in cases of early immature arrest. In addition, also the other frequent genes affecting this pathway such as FLT3, NF1 and PTPN11 (see 1.3.2.1) have been observed in T-ALL [178].

Epigenetic regulators

In T-ALL, there is a notorious epigenetic regulator called PHF6 which is a plant homeodomain (PHD)-containing factor that is frequently mutated and deleted in adult male patients (38 %) [178]. Among its functions, this gene encodes for a protein that interacts with nucleosome remodeling deacetylase (NuRD) complex and therefore, helps regulate nucleosome positioning and transcription. A recent study revealed that Phf6 is associated with HSC homeostasis and has oncogenetic power to leukemia initiation which suggests that alterations in PHF6 are an early event in leukemogenesis [245]. Another epigenetic regulator highly mutated in adults compared to pediatric patients, specifically in ETP T-ALL subtype [246], is DNMT3A. This gene encodes for DNA methyltransferase 3A which has been identified in AML and the preceding stages: myelodysplastic syndrome (MDS) and elderly individuals with clonal hematopoiesis (CH) [247]. Another group of epigenetic modifiers mutated are members of the polycomb complex such as EZH2, SUZ12 and EED which normally is involved in transcriptional repression and are also abundant among ETP T-ALL cases [195]. Other epigenetic regulations altered are lysine demethylase 6A (KDM6A, also called UTX) and histone acetylation modifiers (CREBBP, EP300, HDAC7, HDAC5, NCOA3) [248].

Other recurrent genes

Around 10-15% of T-ALL present loss-of-function mutations or deletions in PTEN [178]. The inactivation of this tumor suppressor activates Akt of the PI3K-AKT-mTOR pathway. This pathway can also be activated by mutations AKT1, PI3KCA, PI3KR1, and IL7R and also by cross-activation from other key signaling pathways such as JAK-STAT or NOTCH1 [248].

Other genes mutated are ribosomal protein genes such as RPL5, RPL10, and RPL22 which frequency sums up to approximately 20% of the T-ALLs. In addition, there have been detected inactivating mutations in CNTO3 (3.8% almost exclusively in adult patients) which encodes for a subunit of the CCR4-NOT complex that regulates mRNA degradation [249].

1.3.2.3 Somatic mutation rate and signatures

In Figure 6 one can observe that overall, leukemias have a lower number of mutations compared to other cancers. Respect to ALL in adults, data from samples collected at diagnosis, shows one of the lowest mutation burdens [94]. Looking at the number of mutations in T-ALL compared to B-ALL the mutation rate is very similar in pediatric patients [234]. However, when comparing adults to pediatric patients with ALL, older patients tend to accumulate more mutations than the youngests. As mentioned before, all organisms accumulate mutations through time so, as time goes by, mutational processes attributable to aging tend to increase their contributions in the total mutational burden. In fact, the major mutational signatures detected in ALL are clock-like signatures (signature 1 and 5) both in adult patients [94,99], as well as, in pediatric patients [234]. There are some cases of B-ALL that also show activity of APOBEC[14] mutational

[14] APOBEC here refers to the group of APOBEC cytidine deaminases the activity of which generates mutations with specific patterns: Signature 2 and 13 of

processes but not in T-ALL. However, APOBEC expression in a few T-ALL cases has been detected so its mutational activity cannot be discarded [250].

1.3.2.4 Germline mutations and predisposition

Briefly, setting aside somatic mutagenesis, TP53 is the gene with more germline detected mutations with predisposition character [251]. Not only in ALL but also across pediatric cancers. However, in the most recent pediatric pan-cancer study no pathogenic germline mutations were detected in T-ALL cases (n=19; [252]). However in both cell lineage types, exist some predisposing mutations such as in IKZF1 (as mentioned earlier [253]). Other genes that have been observed with predisposing leukemia mutations are PAX5, RUNX1 and ARID5B [223,251,253].

1.3.3 Treatment Resistance and Relapse

As outlined above, the percentages patients undergoing relapse are 15-20% of pediatrics and 40-75% of adults [182]. Although pediatric ALL has one of the highest cure rates in cancer (around 90% [202]), relapse of this disease remains the major cancer-related death cause in children [254]. In adults, the outcome is dismal with an overall 5-year survival of around 30-40% [179]. Different studies have evidenced the existence of minor subclones called “relapse-fated” existing at time of diagnosis [255–257]. The characterization of subclones contributing to relapse is one of the current main research focuses with the objective to find markers of recurrence and time its appearance and progression for an early detection [182,254].

COSMIC. These enzymes, most likely APOBEC3A in the majority of cases, are responsible for local hypermutation mutagenesis that creates these signatures [96].

1.3.3.1 Clonal evolution and relapse in ALL

Over the last years, different lines of evidence derived from the study of clonal evolution suggest that leukemias present more clonal complexity that it was initially thought. In the leukemic scenario, during tumor initiation, selection happens as normal HSC and early progenitors compete for resources whereas, during the progression of the cancer, selection acts upon the different leukemic clones [258].

Pre-leukemic development

As noted earlier, different studies support an *in utero* origin of pre-leukemic clones carrying aneuploidies or fusion gene aberrations in pediatric B-ALL. In this review of Dr. Mel Greaves, he summarizes the evidence from his own work and others regarding tumor initiation [168]. First, studies of monozygotic twins with concordant B-ALL revealed that there are several genomic lesions considered founders of B-ALL that are shared in pre-leukemic clones such as ETV6-RUNX1, Hyperdiploidy, BCR-ABL1 and MLL-AF4 and are acquired in utero and transmitted from one twin to the other by blood transfusion. Another experiment with pediatric discordant twins further confirmed the shared existence of the lesion in pre-leukemic cells of the healthy twin. Other studies mentioned in the review backtracked early B-ALL events in neonatal blood spots in which for the majority of the B-ALL checked patients they confirmed that the ETV6-RUNX1 and MLL-AF4 fusions were already at birth. All these studies also suggested that, in the case of ETV6-RUNX1+ or hyperdiploidy, the event is necessary but not sufficient to drive leukemogenesis whereas MLL fusions found in infant patients is sufficient by itself. Discordant healthy twins and some further studies of cord blood samples in the population also suggested that: (1) secondary events are necessary to drive B-ALL, (2) *in utero* pre-leukemic origin is more common than the incidence of B-ALL itself. Therefore, the

leukemogenic transition of the pre-leukemic clone to the disease seems low which may indicate either a frequent loss of the pre-leukemic population or difficulties to acquire secondary leukemic events (strong bottleneck). Related to that, Greaves bets for the "delayed infection hypothesis" as a plausible explanation of acquisition of post-natal secondary lesion by viral causation which could explain the higher incidence of childhood B-ALL in modern societies. In addition to B-ALL, the group of Greaves also pointed out a prenatal origin of T-ALL with NOTCH1 being detected in neonatal blood spots of one pediatric patient [259].

Unlike AML, where there are measurable pre-leukemic stages such as CH and MDS in adult patients, our knowledge of pre-leukemic cells progressing towards ALL initiation in adults is scarced. According to Greaves, the fact that ETV6-RUNX1 or hyperdiploidy, which are the subtypes with more evidences of *in utero* origin, are less prevalent in adults suggests a low persistence of the pre-leukemic clone with aging and points towards adults having a different cancer respect to children [168]. Apart from that, examples of pre-leukemic stages have been observed in familial ALL in which germline mutations are the first event to settle tumor initiation. Briefly as an example, in a study of 5-generation kindred with 10 individuals suffering from B-ALL and other hematological diseases (DLBCL, aplastic anemia, and/or thrombocytopenia) they discovered a common germline deletion of ETV6 as the most likely predisposition event [260].

Order of acquisition and relapse-enriched alterations

A recent study [261] performed single-cell targeted sequencing in 4 pediatric patients of T-ALL in pair samples at diagnosis and remission. Analysis of the CD34+CD38- multipotent compartment with a graph-based algorithm revealed the most probable order acquisition for each patient.

According to their results, loss of CDKN2A and known oncogenic fusion genes are intermediate events whereas NOTCH1 activating mutations tend to happen late. Early events detected were of unknown significance except for STAT5B mutation detected in one of the patients. This contrasts with the idea of NOTCH1 alterations being an early prenatal event as explained above. The notion that NOTCH1 mutation gain can appear as both, an early or late event, in T-ALL has already been suggested before as mutations in this gene sometimes appear as secondary events [262].

In a similar study using single-cell sequencing with B-ALL samples, CRLF2 rearrangements were mostly early but sometimes can be late events in leukemogenesis too [263]. In another study using single-cell sequencing combined with bulk sequencing of 6 B-ALL patients, they determined the temporal ordering of events and reported that ETV6-RUNX1 translocation and structural variation due to RAG-mediated activity are early events, followed by clone-specific APOBEC punctuated mutagenesis and showed that acquisition of oncogenic SNVs such as proliferative KRAS point mutations are late events and not sufficient to boost a clonal dominance in the developed primary leukemia [264]. They also showed that VDJ recombination can occur at different progression moments of leukemogenesis since it can also be ongoing in more evolved clones.

Other alterations are considered late events as they appear in most blasts at relapse. Recurrence of the disease happens after a stringent bottleneck generated by the treatment which (if fortunate) leads to a disease remission. Reasonably, within the mutations carried by the relapse clone there might be genomic event/s driving drug resistance. Therefore, a lot of research projects are uncovering the relapse genomics. The focus is made on relapse-enriched genes which are those with mutations retained in diagnosis until

relapse (therefore shared between samples of the same patient) or those genes with acquired mutations specific or private to relapse.

One of the first ones to look at enriched-relapse alterations was Mullighan and colleagues back in 2008 who checked CNA in 61 ALL children comparing primary to relapse samples of each [255]. Results revealed a list of common deletions of genes in relapse samples: CDKN2A/B, ETV6, IKZF1, NR3C1 and TCF3. For some patients, some of these deletions were not shared with the primary sample but acquired at relapse. NR3C1 which encodes for the glucocorticoid receptor postulated as a treatment-resistant driver gene since this type of steroids are administered during treatment of ALL. Other genes involved in the glucocorticoid signaling have also been detected at relapse samples such as: BTG1/BTG2 and TBL1XR1. The first operates as a co-activator of the glucocorticoid receptor and the second is involved in regulation of the receptor responsive elements [265]. IKZF1 deletions have also been recurrently found in relapse samples of B-ALL both in pediatric and adult [266,267]. Another relapse enriched altered gene is TP53 in which copy number loss and mutations have been associated with nonresponse to chemotherapy [230].

In a mutational landscape study by Dr. Adolfo Ferrando's lab, NR3C1 is not only deleted but relapse-specific mutations in both T-ALL and B-ALL have also been detected [257]. In addition, in a recent study the paralog NR3C2 has also been reported as relapse enriched [112] together with other genes further explained in the upcoming paragraphs. Going back to the mutational landscape, one of the highlighted results in Ferrando's paper is the high frequency of activating mutations in members of the RAS-MAPK in the relapse samples which seem to have a dual role regarding resistance and sensitivity to different chemotherapies.

Apart from that one of the main discoveries regarding resistant relapses by Ferrando's team and others is the role of activating mutations in NT5C2 in ALL [268,269]. In a back to back publication, both reported that this gene encodes for a 5′-nucleotidase enzyme that confers resistance to purine analogs and that it was found with relapse-specific mutations in 20% T-ALL cases and 3-10% of B-ALL. Moreover, other relapse-specific alterations have been detected in genes involved in purine metabolism such as PRPS1 and PRPS2 which are also related to resistance [112,270]. Concretely, PRPS1 gene encodes for an enzyme regulator of the "the novo" purine synthesis. The authors reasoned that the resistance to thiopurines happens when mutants of PRPS1 protein cannot be inhibited by reduced negative feedback loop so they enhanced "the novo" purine synthesis that competes with the metabolization of thiopurine drugs and thus generating tolerance to them.

Other groups of common relapse genes are epigenetic regulators, metabolic genes and mismatch-repair pathway members. Among the genes in the first group the most notorious ones are CREBBP (acetyltransferase), its paralog EP300, NCOR1 (nuclear corepressor complex), WHSC1 (methyltransferase), EZH2 (methyltransferase), SETD2 (methyltransferase), CTCF (zinc finger) and KDM6A (demethylase) [265,271]. Deletions and sequence mutations of CREBBP are believed to interfere with glucocorticoid responsive genes [272]. Apart from that, in a recent study of relapse B-ALL in adults [267], they specifically highlighted the enrichment of novel alterations in metabolic genes in the recurrence of this disease. Briefly, they detected relapse-specific mutations in FPGS (Folylpolyglutamate Synthase) which catalyzes polyglutamylation of methotrexate (a step necessary in the processing of this drug) and ABGL1 (ATP/GTP Binding Protein Like 1) which has a a glutamate decarboxylase function also involved in glutamylation processing. FPGS has also been

detected in pediatric relapse cases [112]. Related to metabolism, in a recent study where relapse clones at diagnosis have been isolated and characterized, they detected a gene expression signature of mitochondrial metabolism as a hallmark of the relapse subclones [182]. Finally, DNA mismatch repair genes such as MSH6, MSH2 and PMS2 are also frequently altered in relapse samples [265]. As an example, not only MSH6 deletions are detected in relapse samples but it has been seen that knockdown of MSH6 gene resulted in higher levels of thiopurines in cells that become unable to initiate apoptotic cascade thus, conferring insensitivity to this drugs [273].

Although the alterations driving primary ALL in T and B-cell lineages are different, since both receive similar multiagent treatment (see 1.3.3.2 below) the relapse-enriched genes suspicious of being resistant mechanisms of the treatment are common among them.

Relapse patterns and leukemia progression

There are 3 models in which the treatment can accelerate clonal evolution in ALL. Landau et al., 2014 called them:

- differential sensitivity model
- mass extinction and competitive release model
- chemotherapy-induced mutagenesis model

The first model explains how therapy selects a minor clone containing a mechanism of resistance to it which then grows and establishes a relapse population. In contrast, the second model refers to cases in which there is a heavy cytoreduction which is insufficient to eliminate all leukemic blasts but that settles the possibility of a change in the clonal landscape allowing a fitter minor clone to expand. If all the remaining clones are equally fitter

then the higher the frequency in the population, the more chances to become the major clone again and resemble the diagnostic composition of cells.

In many backtracking studies of driver alterations, the majority of ALL recurrences come from a minor subclone at diagnosis previous to the treatment. For example, in Mullighan et al., 2008, he estimated that 52% of the relapse cases came from a pre-existing minor clone [255]. Probably, these cases belong to the *differential sensitivity model* whereas the 34% of the relapses that were reported, evolved from diagnosis clone which in this case, better fits with the *mass extinction and competitive release*. In a study with 20 pediatric cases with primary-relapse samples, they reported that 75% of the relapsed B-ALL arised from minor subclones at diagnosis with 45% of them harboring NT5C2 mutations [256]. In a similar study with T-ALL, they analyzed a total of 13 cases in which all relapses came from subclone at diagnosis. However, 6 out 13 patients had a relapse with mutations already detectable at primary leukemia whereas in the rest, the major primary clone was lost in relapse pointing towards an ancestral pre-existing clone. Again, mutations in NT5C2 were observed in 5 of the patients. A very recent study used research techniques such as deep digital mutation tracking and xenografting to classify 92 cases of childhood ALL and concluded that 50% of the times the relapse-fated clone arises from a minor clone at diagnosis, 27% comes from the primary major clone and 18% has a multiclonal origin. Regarding relapse in adults, similar numbers were obtained when comparing childhood vs adult B-ALL cases (46% and 58% respectively) which had a relapsed-leukemia coming from minor clones [267].

Given that, it has been shown to what extent chemotherapies can leave a mutational footprint or signature due to their DNA damaging effect [111], it might be that the chemo-mutagenesis confers therapy-induced resistance

to a clone that then expands and generates a relapsed leukemia. In fact, a recent study of pediatric ALL, reported a new signature derived from thiopurines from which they could estimate the probability of this purine analog's treatment causing relapse-specific driver mutations and found that, for some cases, resistant mutations in genes PRPS1, NT5C2 and TP53 were most likely chemotherapy-induced [112]. This is a clear example of a *chemotherapy-induced mutagenesis model.* In addition, in another recent study [254], mutations in NT5C2 were not detected by ddPCR in primary leukemias but were present exclusively in relapse which further support the idea that mutations at this gene are acquired during treatment or even due to treatment. On top of that, another recent study from the same authors was able to isolate relapse initiating clones from diagnosis samples with limiting dilution xenografting experiments. Results confirmed the existence of relapse clones that in fact, were showing already intrinsic tolerance capability to some drugs at diagnostic samples. In addition, two patients had NT5C2 at relapse, for one patient, mutations in this gene were not detected in the relapse-fated clones coming from its PDX from primary samples whereas in the other patient the mutation was at a very low frequency at primary PDX [182].

Related to a previously mentioned study describing a thiopurine signature [112], the authors also tried to characterize early from late relapses. Concretely, they categorized relapses in three groups according to the elapsed time between diagnosis and relapse: *very early* (less than 9 months), *early* (between 9 and 36 months) and *late* (more than 36 months). Results show that, based on their estimates of population growth, pre-existing resistant subclones at diagnosis fit with observed timing of *very early* relapse whereas *early* relapse adjusts better with a relapse arising from a persistent subclone that acquires resistant alteration during treatment allowing proliferation before therapy ends. Therefore, *later* relapse may

come from a survivor subclone that restarts proliferation after the therapy period is over. In fact, early relapses were the ones presenting more relapse-specific mutations in known genes associated with resistance than the other relapse types (65% over 17% *very early* or 32% *late*). They are not the only ones to characterize early/late relapses. Another recent study [183], came up with a score to represent "clonal dynamics" based on VAF shifts of different mutations at different samples (primary-relapse/s). They reported more clonal dynamics in *early* relapses than *late* and associated this to more plasticity of the tumor favoring quick emergence of fitter clones and a change in predominant populations post-treatment whereas late relapses were considered to arise from quasi-inert persistent clones.

Above all of the relapse population characterization, studies of phylogenetic trees built with primary-relapse samples revealed a general branching pattern in the evolution of ALL. In a study of 55 pediatric patients with ALL [257], the shape of the trees based on mutations detected with WXS, showed enough private primary mutations to consider a branch instead of a linear evolution process. Another study checking for CNA with multiplexing fluorescence *in situ* hybridization demonstrated a complexity in the clonal architecture of ALL and branching evolutionary trajectories [274].

1.3.3.2 Standard treatment

Diagnosis and risk factors

Usually, identification of lymphoblasts by morphology and cytochemistry[15] and assessment of peripheral blood and bone marrow infiltration is

[15] Lymphoblasts lack myeloperoxidase so they stain as Sudan black negative. It helps to distinguish AML from ALL [275].

performed (>20%). Immunophenotyping determines cell lineage and precursor commitment [275]. Cytogenetics identifies main chromosomal changes and aneuploidies and helps stratify patients according to most likely outcome (adverse/poor, intermediate or favorable/good) as described in 1.3.1. In fact, it is believed that part of the overall worst outcome of adults compared to children is because older patients tend to present ALL with adverse cytogenetics such as hypodiploidy, Ph positive type or ETP T-ALL [276]. During the last years, and with the progress of NGS many research groups have developed panels of driver genes to perform targeted sequencing of the diagnostic sample and fine-tune risk assessment. For instance, in a study where they analyzed a panel of genes with main hotspot exons of TP53, JAK2, PAX5, LEF1, CRLF2 and IL7R, it was demonstrated that mutations in TP53 and JAK2 are associated with poor prognosis since they obtained lower overall survival (OS), lower event-free survival (EFS) and higher relapse rate (RR) in an heterogeneous cohort of 340 B-ALL patients with children and adults [277]. Another example is a study of alterations in IKZF1 in two independent cohorts which resulted in very poor outcomes for the patients carrying them [220]. There are also controversial examples such as deletions in CDKN2A and mutations in NOTCH1 in which their assessment as genetic markers have been tested a great number of times without concordance in the results. Pediatric patients under the ALL-97 (n = 55) protocol and adults in LALA-94 (n = 87) and GRAALL-2003 (n = 54) clinical trials showed good outcome of individuals harboring NOTCH1 and FBXW7 mutations whereas in UKALLXII and ECOG protocols presented no significant association [240]. However, although there is some disagreement in long-term outcome between studies, in general, mutations in the NOTCH1 pathway are associated with good early response to treatment. Similarly, in the majority of the studies deletions in CDKN2A/B are associated with poor prognosis but there are also some that found no prognostic value of it [278]. A recent metastudy seems to support

that CDKN2A/B deletions are indeed related to a bad outcome and serve as an independent prognostic marker in adults and children [279].

Risk is systematically assessed according to clinical features: age of the patient and leukocytes or white blood cells count (WBC). The National Cancer Institute has determined as Standard Risk (SR) children between 1-9 years old and peripheral-blood leukocyte count at diagnosis <50000/μL whereas high risk (HR) children are those with 10-15 years old and leukocyte count ≥50000/μL [280]. Other determined categories are adolescents with ages 16-20 years and young adults with ages of 21-39 years which are usually referred together as AYAs. Therefore, in some cases adults are just considered those older than 40. These categories are widely used by the St. Jude Hospital research projects that are cited along this work. Later, risk is re-evaluated according to the first response to treatment. Concretely, it is performed a quantification of the residual disease, called minimal residual disease (MRD) by microscopic morphological assessment, which helps monitoring ALL under therapy and, so far, it is the strongest predictive feature of this disease [281]. Patients whose MRD stays high and never achieve complete remission are called refractory.

Overview of treatment

Following the generalized summary of this recent review [276], front-line treatment of ALL has 4 major components which are blocks with specific drugs and dosages adjusted for a particular period of time:

- Induction
- Consolidation
- Intensification
- Maintenance

The induction block aims to make the most cytoreduction possible to reach a complete remission (meaning absence of detectable disease) and restore normal hematopoiesis. It usually lasts 5 weeks approximately. The chemotherapy given is a multi-drug cocktail of glucocorticoids (immunosuppressor), vincristine (mitotic inhibitor-cytotoxic alkaloid), L-asparaginase (cytotoxic enzyme) and anthracycline (antitumor antibiotic). The most common glucocorticoid is prednisone but some protocols have used dexamethasone too which seems to present higher toxicity. Also the most used anthracycline is daunorubicin. Patients with BCR-ABL1 translocations which had very bad prognosis have significantly improved their outcomes as tyrosine kinase inhibitors such as dasatinib have been incorporated as part of the therapy. Usually, an MRD measure is taken in the middle and at the end of induction. A current measure to determine whether the patient should follow high-risk or standard-risk procedure is to determine an MRD > or < 0.01% respectively. Therefore, intensification of the therapy is adjusted according to the risk to reduce toxicity and long-term effects on those with good prognosis. At this point, allogeneic hematopoietic cell transplantation (allo-HSCT) is also considered depending on comorbidities and overall status of the patient [282].

The consolidation block consists of administering chemotherapy in frequent pulses every 2-3 weeks. The main objective of this block is to get rid of the remaining leukemic cells. Induction seems to be more similar between protocols compared to consolidation which tends to vary more. The usual drugs administered in this phase are: cytarabine (antimetabolite-pyrimidine analog), high-dose methotrexate (antimetabolite-folic acid analog), vincristine, asparaginase, mercaptopurine (antimetabolite-purine analog), and glucocorticoids.

The intensification block consists in a reinduction phase so drugs supplied are very similar to those in induction. It is also common to use cyclophosphamide (alkylating agent). Similarly to consolidation, the aim of this phase of treatment is to ensure the eradication of any left leukemic cells. High-risk patients that have achieved remission are usually the ones that receive this therapeutic block.

The maintenance block is the longest period of treatment (approximately up to 2 years from induction). Mercaptopurine is the main drug administered in this treatment phase which is daily given. It is normally combined with weekly doses of methotrexate. During this period there might be short reinductions too in which mercaptopurine and methotrexate are interrupted for the delivery of induction drugs.

Apart from these, along the different blocks, patients also receive intrathecal chemotherapy (methotrexate, cytarabine, and hydrocortisone) to avoid CNS relapse. Furthermore, as mentioned above, allo-HSCT is also part of the therapy for high-risk relapsers and poor responders [275].

Some of the success of cure rates in ALL, especially in children, are due to some of the following improvements [275,283]:

- The development of specific drugs to trigger response in blood cancers such as folic acid antagonists, corticosteroids, and purine analogs such as (6-mercaptopurine and thioguanine)
- Establishment of multiagent drug schedules to overcome resistance and toxicity
- Prevention and/or treatment of CNS blast infiltration from the very beginning of the treatment first by prophylactic cranial irradiation and then changed to intrathecal chemotherapy.

- Testing of new drugs as well as intensification measures by well design clinical trials (e.g incorporation of asparaginase in treatment protocols or incorporation of reinduction after consolidation)
- Stratification of patients due to risk factors and ALL subtypes and adjustment of dosages and regimes according to that. Notoriously, MRD monitoring to precise patient stratification
- Allo-HSCT as consolidation for high risk-patients and poor responders
- Incorporation of tyrosine kinase inhibitors to treat Ph+ patients
- Improvements in AYA patients with therapy that resembles pediatric therapeutic regimens.

Recently, new treatments have been developed to improve ALL cure rates, especially targeting relapsed leukemias. For example, rituximab and inotuzumab ozogamicin are monoclonal antibodies against B-cell lineage leukemia markers such as CD20 and CD22 respectively that have shown very promising results on relapse and refractory adult patients with reasonable toxicity. In fact, inotuzumab ozogamicin has already been approved by FDA and EMA[16] to treat adult patients with relapsed or refractory leukaemia. Similarly, there are also new antibodies developed to target CD19 such as Blinatumomab which has already been approved for the treatment of adult Ph negative patients. CD19 can also be targeted using immunotherapy with anti-CD19 chimeric antigen receptor (CAR) T cells. Results with patients of a wide range of ages showed promising results but anti-CD19 CAR T treatment has severe effects and usually is only recommended after different alternative lines of treatment have been given or after allo-HSCT.

[16] FDA: US Food and Drug Administration
EMA: European Medicines Agency

In the case of T-ALL, there has not been as much improvement as in B-ALL. A notorious advance is the usage of nelarabine (antimetabolite-purine analog) to treat relapsed and refractory T-ALL patients of all ages. Some inhibitors of key players in T-ALL such as JAK inhibitors, BCL-2 inhibitors are being tested to improve EFS rates. CAR T for T-ALL is also under development but issues regarding the similarities between leukemic T-cells and genetically engineered CAR T-cells must be overcomed [219].

The poorer outcome in adults compared to pediatric patients may be attributable to different factors. As mentioned before, adults tend to have more incidence in adverse subtype groups than children. They suffer from higher toxicity than children, for instance, severe hepatotoxicity due to asparaginase, so as consequence, this drug is not administered at the intensity of children or it is even dropout from treatment protocols. Older patients also tend to have more comorbidities associated with treatment. Another suggested explanation is that, since ALL is rare in adults, there is a lack of specialized centers so there is lesser awareness of stratification and management of toxicity and most adults are treated outside clinical trials [284].

Described treatment resistance

As indicated before, NT5C2 activating mutations are one of the main alterations that drive relapse as they confer resistance to 6-mercaptopurine [268]. This gene encodes an enzyme called cytosolic 5′-nucleotidase II which is responsible for the dephosphorylation of purine nucleotides. The dephosphorylated nucleotides can then be exported out of the cell, therefore reducing their intracellular levels. Gain-of-function mutations of NTC52 increase the export of dephosphorylated purine analogs like 6-mercaptopurine or thioguanine conferring resistance [285]. Evidence

supports that NT5C2 mutations are disadvantageous to the leukemic cells as they cause an excess of nucleotide exportation but they provide selective advantage when mercaptopurine is administered [286]. Mutations observed are classified on those locking the protein in its constitutively active form, those blocking the switch-off mechanism and those that truncate the break for allosteric activation [285]. In addition to NT5C2 and as mentioned above (see 1.3.3.1), PRPS1-mutated clones can also generate tolerance to purine analogs since the mutants reduced the feedback inhibition loop of the protein which ultimately results in inhibition of the drug metabolization into its active damaging form [270].

There are studies associating polymorphisms and mutations of the glucocorticoid receptor encoded in NR3C1 with glucocorticoid resistance but there was a need for functional studies, such as the ones performed with NT5C2, to better understand it [287]. A couple of years ago, a study demonstrated how the glucocorticoid receptor associated with CTCF interact at lymphocyte-specific open chromatin domains (LSOs) to regulate chromatin accessibility critical for the glucocorticoid-induced apoptosis. They reported how glucocorticoid resistant cells had an increased methylation of DNA at the enhancer preventing formation of DNA looping and, therefore, impeding the binding of transcriptional machinery necessary to trigger apoptosis [288].

Furthermore, in B-ALL, relapse samples recurrently present mutations in CREBBP which have been shown to impair regulation of glucocorticoid-receptor-responsive genes [272]. Instead of a particular genomic mutation conferring resistance, some studies have focused on the overall alteration of pathways such as JAK-STAT and PI3K-AKT-mTOR and resistance to glucocorticoids [287]. For example, a few years ago, a study with PDX showed that in T-ALL with active JAK-STAT pathway, removing IL7 or

inhibiting JAK-STAT signaling stimulated by IL7 sensitizes cells to glucocorticoids and might help to overcome resistance [289]. Another mechanism of resistance to chemotherapy has been described with p53 and ABCB1. As indicated above (1.1.3.4), the gene ATP-binding cassette sub-family B member 1 (ABCB1) encodes for P-gp protein which is a ATP-dependent membrane efflux pump that is able to export several drugs such as vincristine, anthracyclines, and glucocorticoids [290]. It has been shown that p53 transcriptionally regulates ABCB1 and that mutations and deletions in TP53 lead to increased expression of P-gp [287]. A recent study reported that cells having TP53 mutants with expressed truncated forms of p53 showed insensitivity to doxorubicin whereas when WT expression was rescued cells were re-sensitized in B-ALL [291].

As described above, loss-of-function of mutations in genes that encode for members of PRC2 complex (EZH2, EED, or SUZ12) are enriched in relapse and are also abundant in adverse subtype ETP-TALL. Depletion of PRC2 is associated with resistance to chemotherapy-induced apoptosis in human T-ALL cell lines [292].

Finally, there are other factors contributing to chemoresistance that can be cell-extrinsic. Some studies summarized elsewhere [293] have shown that the bone marrow microenvironment by modulating cell-cell interactions and by the production of soluble factors in the niche induce survival signaling which can contribute to chemoresistance. For example, several lines of evidence related the abnormal high expression of surface integrin VLA-4 in leukemic cells to chemoresistance, particularly in tests with cytarabine. Specifically, the binding of VLA-4 with VCAM-1 (surface protein in vascular endothelium cells) mediates activation of pro-survival signaling in ALL cells which can shield them against treatment induced apoptosis. Another example is, the chemokine receptor CXCR4 which has

also been involved in chemoresistance. High expression of this receptor is associated with higher relapse rates and inferior survival and its silencing has been related with restore of chemosensitivity and induction of apoptosis in ALL.

2. OBJECTIVES

General goals

- Study the clonal evolution from primary to relapse of T-ALL in adult patients
- Identify genomic candidates of treatment resistance in adult T-ALL
- Analyze similarities and differences between T-ALL and B-ALL and between the pediatric and adult diseases
- Contribute to the generation of a compendium of mutational cancer genes across tumor types

Specific goals

- Characterize the emergence of the relapse clone
    - Estimate the time before clinical presentation when the primary and relapse clones diverged
    - Infer the size of the relapse subpopulation at the time of diagnosis of the primary
    - Identify recurrent relapse-enriched alterations suspicious of conferring resistance
- Comparison of drivers between different types of ALL
- Define which are the mutational processes operating in leukemogenesis and, also, whether the relapse samples show signs of chemotherapy signatures
- Collect, curate and annotate datasets of tumor somatic mutations across cancer types

3. RESULTS

3.1 Chapter 1

The evolution of adult T-ALL patients

The results section contains the accepted manuscript of my main project during the PhD. Like most studies in our field, the publication shows only part of the large amount of work done during the project. For this reason, in the following introductive part, I decided to describe a detailed overview of all the progression of it, to provide a more realistic and complete picture of the work carried out. Furthermore, this beginning of the results section, also aims to define the contribution of all the people involved in the project since there are other authors of our lab and external collaborators in it.

When does the project start?

The research explained in the manuscript of the upcoming section is a collaborative project with the lab of Dr. Anna Bigas[17] and the group of Dr. Josep Maria Ribera[18]. During my first PhD year, I joined the project at the initial planning phase in January 2017 (see the diagram of Table 5). Given the high incidence of childhood B-ALL, most of the genomic knowledge generated in the past years such as the implicated genes and pathways and the evolution under treatment of ALL is clearly biased towards the pediatric disease. Consequently, little is known about the evolution of adult T-ALLs under treatment. The motivation of the project was to find mechanisms of therapy resistance in T-ALL, check their detection at diagnosis and include them in the regular diagnosis test of the Hospital Germans Trias i Pujol for

[17] Stem cells and cancer Group-Institut Hospital del Mar d'Investigacions Mèdiques (IMIM)
[18] Acute lymphoblastic leukemia Group-Josep Carreras Leukaemia Research Institute (IJC)

relapse prevention. The project is funded by The Asociación Española Contra el Cáncer (AECC)[19]. The initial plan was divided into two phases: a first 3-year phase to sequence samples and to identify candidates and a second phase of the following two years to validate the candidates and develop xenograft models for further prediction analyses. The group of the IJC was already collecting data from adult T-ALL patients from different spanish hospitals and the group at IMIM was coordinating the project and bringing the experimental expertise to our ensemble. Our role was to lead the bioinformatic analysis which had the central weight in the investigation as evidenced in the objectives written above.

Project design and sequencing

We have frequently met with our collaborators along the past 4 years. The meetings, in particular during the first period, had as main objective to revise the clinical data of the patients and the sample availability to decide which ones could be included in the project. In order to study the evolution of leukemias and to search for alterations of therapy resistance, we selected primary and relapse bone marrow aspirates (preferably, otherwise peripheral blood) with reasonable sample purity of patients above 18 years old. After a few months, our collaborators were able to gather a collection of primary, remission and relapse samples (here referenced as trios) per individual from a small cohort of 9 adult patients with T-ALL. We decided to sequence the whole genome of the samples from this first batch with the expectation to sequence an in-house cohort between 20 to 30 T-ALL adult patients. However, the sequencing center was not up to us to decide since local regulation stated that the election of it must undergo public tender procedure. We received the sequences from the first batch of patients in June 2018.

[19] Translated: Spanish Association Against Cancer

Collection of ALL tumor sequences in the public domain

In the meantime, we decided to gather as much data as possible from ALL. We downloaded the somatic mutational catalogs from the published projects that were available at the moment. Logically, most of the data that we obtained was coming from pediatric cohorts. Usually, the data we collected were somatic mutations from the supplementary MAF files reported in publications. At that time, not only was I a first year PhD student but also I was finishing my master's degree so, the first cohort comparison and landscape of driver genes carried out served as my master thesis[20]. Soon, we realized that the diversity in mutation calling procedures among the cohorts could influence downstream analysis that we were willing to perform. For this reason, to continue the investigation, now as my main PhD project, we decided to download and re-analyze the raw data ourselves to accomplish as much homogeneity as possible between cohorts. Not all the published projects provided a link to a public repository from which to download the raw data. Finally, we ended up having mostly WGS data from pediatric projects of the St. Jude Pediatric Hospital and another pediatric project with WXS data from Columbia University [257] that had samples trios per patient (a table with the characteristics of the cohorts is provided in the manuscript; there were a total of 238 patients).

First steps of the analysis

Since it was the first time that our lab performed sequence alignments and calling of somatic variants for such a volume of data, I, with support from other lab members, had to come up with a pipeline to systematically analyze all the samples. After trying some aligners and exploring GATK possibilities, we decided to use Sarek pipeline [294] developed by

[20] https://www.upf.edu/web/bioinformatics/projects-2016-2017

SciLifeLab in Sweden in nextflow scripting language. At that time, the pipeline was still at the beginning of its development and it was called CAW. However, the first analysis was done with the alignments and somatic calls outputted by CAW of a few cohorts. The pipeline seemed appropriate since it was using a pretty standard way to make the BAM files while implementing the "Best Practices" of GATK and using Strelka and MuTect. After some benchmarking of other calling tools and "trial and error" we had the SNVs, InDels, CNV and SV of some of the patients (¼ of the total number of patients coming from public repositories). I made the first BAMs and test runs myself with the help of Jordi Deu Pons.

However, at some point of the analysis, Dr. Loris Mularoni helped run the rest of the alignments and mutation calls with a stable version of Sarek while I was working in the first batch and learning about single-cell RNA-seq.

Project	2017			2018			2019			2020		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Initial planning: Selection of first batch of T-ALL adult trios (n=9) to sequence and public tender release												
Planning a comparison project and download and collection of ALL WGS data from public repositories												
Testing for pipeline to align reads (CAW) and internal benchmark for variant callers												
Received first batch of adult T-ALL patient's sequences (n=9), start of the analysis												
Planning of pilot test of single-cell RNAseq												
Re-do alignments and calling with new version of pipeline Sarek												
Selection of second batch of T-ALL adult trios (n=6) to sequence												
Preliminary results of first T-ALL adult batch with initial comparison with the analysis of public cohorts												
Received second batch of adult T-ALL patient's sequences (n=6)												

Incorporation of the second batch data to the analysis												
Received sequence data from pilot scRNA-seq and learning how to analyze it (by the end of the year, droped it)												
Selection of third batch of T-ALL adult trios (n=4) to sequence. Planning of deep sequencing of mutations with dPCR												
Received sequence data from third batch												
Finishing analysis and preparation and submission of the manuscript												
Incorporation of data from third batch and dPCR to manuscript												
Thesis writing												

Table 5. Gantt diagram of the main steps of the project (1st, 2nd and 3rd refer to trimesters)

Related to that, a pilot study to try to quantify the heterogeneity of the primary tumor and the existence of a relapse form was performed using 1 patient (the only one at that time with cryopreserved cells). The inconclusive results and the lack of material suggested to abandon such idea.

Therefore, we saved that line of investigation for another time and started new angles to study clonal evolution from bulk sequencing. Finally, we have sequenced the whole genome of sample trios of 19 patients.

Sections of the manuscript and contributions

Figure 1 and 2 of the manuscript correspond to the first part of the project in which we decided to compare cohorts of ALL to our in-house cohort of T-ALL adults and make a landscape of the disease, as well as, search for therapy-resistant candidates. From figure 3 to 5 we decided to focus on the leukemic evolution of the in-house cohort. Since this is my main PhD project, I have actively participated in all parts of the project, including the initial meetings when we decided the samples to sequence, to the discussions of experimental validation of mutations. Regarding the computational analyses, Dr. Santi González and I have performed all of them. I did the analyses of the first part of the project from the somatic calling of alterations, the consecutive filtering steps, to the discovery of drivers running IntOGen and curating literature and to the fitting of mutational signatures. During this period I had constant feedback from Santi who helped guide some technical decision-making from one step to the next one. The second part of the analysis was built upon many discussions with Santi, my supervisors and I, with occasional input help from Dr. Ferran Muiños for more mathematical-related technical parts. Concretely, I coded a simple model to estimate the divergence time of the primary and relapse clones. However, more accurate modelling was needed

from where Santi took over this part and designed and implemented the mutation rate increment models that resulted in the estimates of the divergence time in figure 4.c. Furthermore, after running some tests with Clonex [295] for the simulations of cell population growth, Santi and I realized it required some adaptation of the code in order to simulate the growth of not only a primary tumor but also a relapse. Therefore, he adapted Clonex and ran the simulations that resulted in figure 5.c and then compared with the observed data in figure 5.d. Ferran implemented the model of tumor growth from Li et al., 2019 [112] so I could obtain the doubling time that allowed me to infer the population size of relapse at diagnosis that corresponds to figure 5.a. The dPCR experiments were conducted by Dr. Violeta Garcia-Hernández under the supervision of Dr. Anna Bigas and with the help of the Pathology Department in Hospital del Mar.

I have also participated in outlining and discussing the draft of the paper. The manuscript has been accepted for publication in Genome Biology for *Cancer Evolution and Metastasis* special issue.

Sentís I , Gonzalez S , Genescà E, Garcia-Hernández V , Muiños F , Gonzalez C, Lopez-Arribillaga E, Gonzalez J, Fernandez-Ibarrondo L, Mularoni L , Espinosa L , Bellosillo B , Ribera JM , Bigas A , Gonzalez-Perez A , Lopez-Bigas N. The evolution of relapse of adult T-cell acute lymphoblastic leukemia (Accepted, Genome Biology)

The evolution of relapse of adult T-cell acute lymphoblastic leukemia

Inés Sentís*[1], Santiago Gonzalez*[1,2], Eulalia Genescà[3], Violeta García-Hernández[4], Ferran Muiños[1], Celia Gonzalez[3], Erika López-Arribillaga[1], Jessica Gonzalez[4], Lierni Fernandez-Ibarrondo[5], Loris Mularoni[1,6], Lluís Espinosa[4], Beatriz Bellosillo[5], Josep Maria Ribera@[3], Anna Bigas@[4], Abel Gonzalez-Perez^[1,7], Nuria Lopez-Bigas@^[1,7,8]

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain
2. Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 10, 08028 Barcelona, Spain
3. Hematology Departments, ICO-Hospital Germans Trias i Pujol, Josep Carreras Research Institute, Universitat Autònoma de Barcelona, Badalona, Spain
4. Program in Cancer Research, Institut Hospital del Mar d'Investigacions Mèdiques, CIBERONC, Barcelona, Spain
5. Pathology Department, Hospital del Mar, IMIM, Barcelona, Spain
6. CMR[B] Center of Regenerative Medicine, Barcelona, Spain
7. Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain
8. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

* These authors contributed equally

^ Co-senior authors

@ Corresponding authors

Inés Sentís: ines.sentis@irbbarcelona.org
Santiago Gonzalez: santiago.gonzalez@irbbarcelona.org
Eulalia Genescà: egenesca@carrerasresearch.org
Violeta Garcia: violeta_gh@usal.es
Ferran Muiños: ferran.muinos@irbbarcelona.org
Celia Gonzalez: cgonzalez@carrerasresearch.org
Erika López-Arribillaga: erika.lopez@irbbarcelona.org
Jessica Gonzalez: jgonzalez3@imim.es
Lierni Fernandez-Ibarrondo: lfernandez@imim.es
Loris Mularoni: lmularoni@idibell.cat
Lluís Espinosa:lespinosa@imim.es
Beatriz Bellosillo: BBellosillo@parcdesalutmar.cat
Josep Maria Ribera: jribera@iconcologia.net
Anna Bigas: abigas@imim.es
Abel Gonzalez-Perez: abel.gonzalez@irbbarcelona.org
Nuria Lopez-Bigas: nuria.lopez@irbbarcelona.org

Abstract

Background: Adult T-cell acute lymphoblastic leukemia (T-ALL) is a rare disease that affects less than 10 individuals in one million. It has been less studied than its cognate pediatric malignancy, which is more prevalent. A higher percentage of the adult patients relapse, compared to children. It is thus essential to study the mechanisms of relapse of adult T-ALL cases.

Results: We profile whole-genome somatic mutations of 19 primary T-ALLs from adult patients and the corresponding relapse malignancies, and analyze their evolution upon treatment in comparison with 238 pediatric and young adult ALL cases. We compare the mutational processes and driver mutations active in primary and relapse adult T-ALLs with those of pediatric patients. A precise estimation of clock-like mutations in leukemic cells shows that the emergence of the relapse clone occurs several months before the diagnosis of the primary T-ALL. Specifically, through the doubling time of the leukemic population, we find that in at least 14 out of the 19 patients, the population of relapse leukemia present at the moment of diagnosis comprises more than one but fewer than 10^8 blasts. Using simulations, we show that in all patients the relapse appears to be driven by genetic mutations.

Conclusions: The early appearance of a population of leukemic cells with genetic mechanisms of resistance across adult T-ALL cases constitutes a challenge for treatment. Improving early detection of the malignancy is thus key to prevent its relapse.

Keywords: T-ALL, adult acute lymphoblastic leukemia, T-ALL evolution under therapy, evolution of leukemia relapse, ALL relapse

Background

Acute lymphoblastic leukemia (ALL) affects 3 children in 100,000 in the UK [1]. In the past 5 decades, intense research on this disease has succeeded in reducing the mortality of ALL-affected children by 82% [2]. Recently, with the development of cancer genomics, researchers have unraveled the most frequent somatic genetic alterations underlying its development [3–14], and molecular subtypes, as well as their clinical relevance [15–22]. Genetic alterations that elicit some relapse events have also been uncovered and the potential role of therapy in the development of such relapse cases has been explored [23–31].

ALL is less prevalent in adults (0.7 patients in 100,000 people [1]). Not only are there differences in incidence among age groups, but also relapses after treatment appear more frequently in adults (40-75% vs 15-20% among pediatric patients) [31]. Very few studies have been dedicated to understanding the genomic roots of the emergence of adult ALL, and in particular, of T-cell ALL (T-ALL) [32–36]. There is a larger gap in the study of the evolution of this malignancy under therapy and its relapse after treatment. Therefore, important questions regarding the genomic evolution of adult T-ALL remain unanswered. It is not entirely clear, for example, whether the same mutational processes are involved in the onset of pediatric and adult T-ALL cases, and if the chemotherapeutic drugs employed in the treatment leave a mutational footprint in relapse cells, as it has been shown for pediatric cases [36]. Furthermore, while some genetic mechanisms of resistance to treatment have been identified in pediatric ALL [26,27], it is not known whether these also contribute to resistance of the adult malignancy.

To explore the evolution of adult T-ALL under treatment and address these specific questions, we profiled the whole-genome somatic mutations of 19 T-ALLs from adult patients who relapsed after treatment (in-house cohort; Additional file 1: Table S1). Samples were taken at the time of diagnosis (primary) and at recurrence of the malignancy after treatment (relapse). We then analyzed the genomic evolution of these adult T-ALL cases in comparison with 238 pediatric and young adult ALL cases (71 with primary and relapse samples) available in the public domain (Table 1). Known or potential resistance mutations appear in 6 patients of the cohort. Nevertheless, our results show that in the 19 cases the relapse is driven by genetic mutations, and that resistant cells appear in the population of blasts several months before the diagnosis of the primary.

Results

The genomics of primary adult T-ALL

Previous studies on the genomic basis of pediatric ALL have identified somatic mutations across cohorts of patients suffering from this disease [5–8,10,12,13,28–30,37–40]. Therefore, we first aimed to compare the landscape of somatic alterations observed across primary adult T-ALL with that across eight other cohorts of T- and B-ALL patients of varying age, ranging from infancy to young adulthood, which we analyzed with a unified mutation calling approach (Table 1; Additional file 1: Table S1 and Table S2). Among cancer types, ALL presents a relatively low somatic mutation burden [41,42]. Nevertheless, the burden of somatic point mutations of adult ALL cases tends to be higher than that of cases of most of the

subtypes of the pediatric malignancy, as has been previously observed [43] (Fig. 1a).

Mutations in human somatic cells are contributed to by different molecular mechanisms involving the interaction of endogenous (for instance, spontaneous cytosine deamination or oxidative damage) and external DNA damaging agents (such as UV-light, tobacco carcinogens or chemotherapies) with the DNA repair machinery [41,44–46]. The study of these mutational processes in tumors reveals the lifetime exposures of patients to potential carcinogenic agents and consequently contributes to shedding light on the etiology of malignancies. Thus, we first asked whether the somatic mutations observed across nine cohorts of pediatric and adult ALL (Table 1) are contributed to by similar or different mutational processes. No clear differences are observed between the mutational profiles of B-ALL and T-ALL (Fig. 1b, top). However, the mutational profiles of pediatric and adult malignancies exhibit discernible, albeit slight differences (Fig. 1b, bottom). The same mutational processes appear to be active across pediatric and adult T-ALL and in pediatric B-ALL (Fig. 1c; Additional file 2: Fig. S1). In particular, mutational signature 5 (SBS5), which in blood has been demonstrated to behave in a clock-like manner [47], and has been associated with the process of hematopoietic cell divisions [48,49], appears as one of the main contributors of mutations in the evolution of both pediatric and adult ALL.

We next asked whether the driver alterations observed across primary adult T-ALL in the in-house cohort are different from those observed across pediatric B/T-ALL (Methods; Fig. 1d; Additional file 2: Fig. S2; Additional file 1: Table S3 and Table S4). Mutations in

some known ALL driver genes, such as NOTCH1 and FBXW7 (the E3-ligase charged with its recognition for ubiquitination [50]), are overrepresented among both pediatric and adult T-ALL with respect to B-ALLs (χ^2 p=1.05x10^{-16} and χ^2 p=8.37x10^{-9}, respectively). Similar overrepresentation of mutations in T-ALLs was found in JAK3 (χ^2 p=0.004). In contrast, RAS activating mutations do not appear to be differently represented in both ALL types (χ^2 p=0.05 and χ^2 p=0.634 for KRAS and NRAS).

Genomic alterations driving primary and relapse adult T-ALL

With the goal to study the evolution of adult T-ALL, the 19 patients in the in-house cohort were selected specifically because they relapsed several months after treatment (Fig. 2a; Additional file 2: Fig. S3; Additional file 1: Table S1). Seventeen of them received the same treatment protocol (ALL-HR-11 [NCT01540812]), while the remaining two were administered very similar protocols (LAL-07OLD and ALL-HR-2003 [NCT00853008]). To uncover the genomic similarities and differences between adult and pediatric T-ALL cases at relapse, we next compared the in-house cohort with 31 relapsed cases from the T-ALL Oshima and T-ALL SJ cohorts (Table 1; Additional file 1: Table S3 and Table S4). A list of potential driver events across the 19 patients in the cohort is presented in Additional file 1: Table S5 and Table S6.

Many NOTCH1 and FBXW7 mutations observed in the primary leukemias were also present in the relapse samples (Fig. 2b; Additional file 2: Fig. S4). Intriguingly, mutations affecting USP7, a known deubiquitinase of NOTCH1 were detected in 3 adult and 3 pediatric patients, raising the possibility of yet another form of alteration of the NOTCH pathway in leukemogenesis [51–53].

Overall, NOTCH1-affecting mutations in adults are distributed along the protein-coding sequence in a very similar manner than those observed in pediatric patients (Fig. 2c). Nine patients in the cohort present multiple mutations of NOTCH1 that affect different protein domains (mostly HD and PEST), in agreement with previous reports [54]. Interestingly, in 6 patients different NOTCH1/FBXW7 mutations were detected in the primary and relapse samples (Fig. 2d). These constitute examples of convergent evolution of mutations affecting the NOTCH1 pathway, also observed in eight pediatric patients in the cohorts analyzed. This suggests that NOTCH1 mutations tend to appear late [55] and recurrently (i.e., in several cells) during T-ALL development.

DNMT3A-affecting mutations, known to drive acute myeloid leukemias (AML), were observed in three adult patients in the in-house cohort and none of the pediatric T-ALLs. In fact, these three patients are classified as Early T-Cell Precursor (ETP), a T-ALL subtype that presents myeloid markers [33]. Similarly, PAT5 and PAT9, patients with mutations of ROBO2 --a gene associated with progression of myelodysplastic syndrome [56] to AML and recently reported as mutated in pediatric ALL [57]-- present the ETP phenotype. Clonal mutations of PHF6 are overrepresented (χ^2 p=0.001) in adult T-ALLs with respect to their pediatric counterparts, shared between primary and relapse samples. PHF6 is a zinc-finger transcription factor that suppresses ribosomal RNA (rRNA) transcription [32]. Loss-of-function mutations of this gene have been shown to decrease sensitivity to glucocorticoids [58], which are part of the standard first-line treatment of adult T-ALL patients. Interestingly, activating mutations of the NT5C2 gene, known to elicit resistance to mercaptopurine anti-ALL treatment in pediatric cases

[26,27] are also observed across 3 adult cases exposed to this drug (Fig. 2a), with PAT16 bearing two mutations of NT5C2 (R238G, R367Q, see Additional file 1: Table S5). In the relapse samples of two patients of the in-house cohort, we observed amplifications of ABCB1, an ATP-dependent membrane transporter known to mediate multidrug resistance in tumors [59,60] (Additional file 2: Fig. S5). Finally, SMARCA4 mutations and deletions were also detected across adult (2) and pediatric T-ALLs, but almost exclusively in relapse malignancies, suggesting a potential role in resistance to treatment.

In summary, in 6 of the 19 adult patients of the in-house cohort we were able to identify a candidate treatment-resistance mutation.

The evolution of relapse adult T-ALL measured through mutations

We next asked how much do the mutational processes active in primary T-ALLs also contribute to the overall burden of mutations of relapse adult T-ALLs. The incorporation of new mutational processes, like the exposure to chemotherapies used in their treatment, could leave a mutational footprint that may be detectable in the relapse clone, as recently demonstrated in metastases of different solid tumors, and in relapsed pediatric ALL cases [45,61].

The deconstruction of mutational signatures (representing mutational processes active during a person's life) of primary and relapse samples of each patient reveals very similar scenarios for primary-private, shared and relapse-private mutations (Fig. 3a). Signature 5 (SBS5), which represents a mutational process associated with hematopoietic cell division [45] contributes the vast

majority (~80%) of mutations in these three groups. We did not detect the mutational footprint of mercaptopurine or any other chemotherapy in the relapse samples (Additional file 2: Fig. S6). This does not preclude that chemotherapy-related mutations exist below the level of detection of the sequencing technology, for example if the evolutionary bottleneck caused by the treatment has not sufficiently reduced the T-ALL population.

Since signature 5 has been described as a clock-like process [47] and this type of mutations are the main contribution to the burden of clonal mutations of both primary and relapse T-ALLs, we used them to infer a molecular time of divergence between the primary and relapse populations (Fig. 3b, Additional file 2: Fig. S7). To this end, we counted the number of primary-private, shared and relapse-private signature 5 clonal mutations (Fig. 3b). In all cases the branch that corresponds to relapse-private mutations is longer than that representing primary-private mutations, because the relapse clone has continued accumulating mutations longer after its divergence from the primary (eliminated as a consequence of the treatment). As expected, fewer relapse-private mutations accumulate in the cases with shorter time elapsed between the diagnosis of the primary and the emergence of relapse.

Time elapsed between divergence of primary and relapse clones and primary diagnosis

The number of primary-private, shared and relapse-private signature 5 clonal mutations can also be used to estimate the precise time of the divergence of the primary and relapse clonal populations. To that end, we first needed to understand the rate of accumulation of signature 5 mutations during T-ALL development. The DNA of

normal hematopoietic cells has been shown to incorporate signature 5 mutations at a rate of roughly 12 per year (Fig. 4a; Additional file 2: Fig. S7; [48]). Regressing the number of signature 5 mutations across primary and relapse T-ALLs on the age of patients in the in-house cohort in comparison with healthy hematopoietic stem cells (HSCs) yields slightly higher mutation rates and an unanticipated high (~400) number of mutations at the start of life of hematopoietic cells (intercept of trendline in Fig. 4a). This deviation could be explained through an acceleration in the mutation rate that occurs upon malignization of hematopoietic cells [62].

To compute the moment of time before diagnosis when this acceleration started, as well as the value of the accelerated mutation rate, we assumed that the acceleration rate is the same for the primary and relapse malignancies of a patient. We then simulated a one-time increase of the mutation rate (constant rate model) during tumor evolution and, alternatively a steady increase (linear rate model) in the mutation rate for successive cell generations (Additional file 2: Fig. S8). For each patient, we assayed several trendlines of accelerated mutation rate (i.e., starting at different timepoints before diagnosis; dotted lines in Fig. 4b) approximating the observed number of signature 5 clonal mutations in the primary and relapse T-ALL clones. We computed the likelihood of each of these trends of acceleration following their accuracy to fit the observed number of mutations in the primary and relapse malignancies (Fig. 4b and Additional file 2: Fig. S8). For each trendline of accelerated mutation rate, the age of the patient at which the divergence of the two clones occurred can be computed from the number of shared mutations. The difference between this age and

the age at diagnosis then yields the time elapsed between this divergence and the diagnosis of the primary T-ALL.

Upon application of this approach to each patient in the in-house cohort, we obtained a number of estimates of the number of days elapsed between the divergence of both clones and the diagnosis of the primary T-ALL, each with varying likelihood (green circles, Fig. 4c). The estimates for each patient may be summarized as their weighted (by likelihood) averages (broken lines). The time estimated for each patient was subsequently refined using the distribution of all patients (see Methods). As a result, we obtained a robust prediction of the boundaries of the most likely time elapsed between the divergence of primary and relapse clones and the diagnosis of the primary malignancy. In the majority of cases shown in the figure (13 out of 15) less than a year passed between its emergence and the diagnosis (Additional file 1: Table S7).

The evolution of relapse of adult T-ALLs

Both the primary and resistant populations of T blasts across the adult T-ALL cohort are composed of a major clone and one or more subclones detectable through sequencing (see Additional file 3). In all the patients, including four that are refractory to treatment, the major clone in the primary and relapse leukemias differ, implying that in every case, the treatment obliterates the major clone in the primary malignancy.

To understand the effect of the therapy on the clonal architecture of adult T-ALLs, we first aimed to estimate the speed of growth of the population of T-ALL cells to determine the minimum size of the relapse population at the time of diagnosis. This growth speed may

be characterized through the doubling time of the population (the time needed by a population of cells to duplicate its number). This can be computed from the number of blasts estimated by the pathologist at remission and relapse, and the amount of time elapsed between both events [61] (Additional file 2: Fig. S9a; Methods). We computed a doubling time for the T-ALL leukemic population of 10.79 days (confidence intervals, 10.1-11.36), which is slightly longer than that recently estimated for pediatric B-ALL [61] (Additional file 2: Fig. S9b). We were then able to compute, with this doubling time, the minimum time necessary for the relapse population to achieve approximately $7x10^{11}$ cells that corresponds to a full grown leukemia [61,63]. This minimum time to expand from a single cell upon its divergence from the primary population informs us of the likelihood that the relapse clone has arisen before the diagnosis of the primary.

In three cases (PAT7, PAT11, PAT12), it is possible that the relapse clone appeared during treatment, given the estimated doubling time. In two more (PAT9 and PAT10), it is not completely clear whether there's enough time between the start of treatment and relapse to allow the emergence of a new clone. In all other cases, the relapse clone was most likely already present at the time of diagnosis and represented by more than one cell (Fig. 5a). Indeed, for fourteen patients in the cohort, the size of the relapse clone at the time of diagnosis of the primary malignancy probably comprises more than 100 of $7x10^{11}$ leukemia cells. (Note that this calculation is independent from the time elapsed between divergence of the primary and relapse clones and the diagnosis computed previously.) PAT2, PAT4, PAT5 and PAT17, with more than 0.01% minimal residual disease during treatment, show estimates of the relapse clone at the time of diagnosis which are, as expected, above 1 in

10,000 blasts. We then asked whether the relapse clone could be detected in the primary sample of ALL cases by a method with a lower limit of detection than Next Generation Sequencing technologies. Thus, we aimed to detect two non-synonymous SMARCA4 mutations (G1162S and T786I) that are private of the relapse sample of two patients in the corresponding primary samples of these patients (PAT8 and PAT14). With a limit of detection of around one in one thousand cells, a digital PCR was unable to detect this mutation in the primary sample of either patient (Fig. 5a and Additional file 2: Fig. S10a,b). The fraction of cells of the relapse clone estimated to be in the primary sample of these two patients is below this limit of detection ($1/10^5$ in PAT8 and $1/10^8$ in PAT14). These results thus provide further support to the estimation of the doubling time and the size of the relapse clone in the primary samples derived from it.

Although we were able to pinpoint known or putative resistance mutations in several cases, we asked whether other cases of relapse could be explained by a failure of the treatment to kill a subset of the leukemic cells independent of any genetic mechanism [28,57]. To answer this question, we modeled the emergence of the relapse clone following both a resistant and a non-resistant (not driven by a genetic mutation) scenario (Fig. 5b). First, a population of tumor cells with driver and passenger mutations was simulated. Then, to model the first scenario, a group of cells sharing one passenger subclonal mutation (the resistance mutation) were selected as survivors of the treatment, and were expanded again for 20, 40 and 60 generations (40 generations correspond roughly to the observed times elapsed between primary and relapse diagnoses for the cohort; Additional file 2: Fig. S11). To simulate the second scenario, a group of cells with

the same size as in the first case (but selected randomly and sharing no particular subclonal mutation) was selected and expanded for the same number of generations. We then compared the change in clonal composition --change of cancer cell fraction (CCF) of mutations in primary and relapse-- obtained for both simulated scenarios with the distribution of CCF in the primary samples of mutations fixed in the relapse samples for all patients, represented in Fig. 5c. For example, of all mutations fixed in the relapse ALL of PAT8 (dashed brown line), approximately 59% were present at CCF 0-0.1% in the primary. In other words, in the primary sample they appeared below the limit of detection of the sequencing, and thus correspond to the red star mutations in the toy diagrams in Fig. 5b. On the other hand, 30% of the PAT8 fixed mutations were detected in the primary ALL at CCF between 0.9 and 1, with the remaining mutations at intermediate CCF bins. All patients in the cohort yield similar bimodal distributions.

Only in the results of the simulation of the resistant scenario do we observe a distribution of CCF of the mutations in the primary sample that resembles that of the patients in the in-house cohort (Additional file 2: Fig. S10). By contrast, in the results of the simulations of the non-resistant scenario, no mutations undetectable in the primary leukemia (CCF in the 0-0.1 decile) become fixed in the relapse (Fig. 5d). This holds if the simulations are run between 20 and 60 generations, and even if a much higher (unrealistic) fitness is assigned to driver mutations. These results suggest that the non-resistant scenario of evolution under treatment is not feasible given the time elapsed between primary and relapse.

In summary, in 14 cases in the cohort the relapse population is most likely already present before the start of the treatment. Moreover, all relapse cases fit the model of genetic resistance --due to one genetic event common to all cells in this relapse population-- although we are only able to identify the responsible mutation in a few of them.

Discussion

Advancing our knowledge on how tumors respond to therapies and which of their features determine their relapse after treatment is key to improving clinical oncology practice. Here, we studied the genomic features and the clonal composition of nineteen adult T-ALL cases at diagnosis and at the time of relapse to understand their evolution and identify commonalities that may predict their likelihood to respond to current therapeutic approaches.

Our results suggest that for most adult T-ALL patients, the population of leukemia cells that dominate the relapse is already present at the moment of diagnosis, that is before the start of the treatment, and comprises more than one but fewer than 10^8 blasts. One evidence that supports this notion comes from the fact that, in most cases, the span of time between the diagnosis and the emergence of relapse is not enough (given the doubling time estimated from the cohort) to explain the repopulation of a full leukemic population starting from a single cell. This contrasts with the results reported recently for a pediatric cohort, in which some relapse cases could be explained by resistance mutations appearing during treatment [61]. This finding is relevant for the clinical practice, since early identification of such potential resistance populations in a patient's leukemia may support making clinical decisions regarding their treatment.

We were not able to detect the mutational footprint of chemotherapies employed in the treatment of patients of this cohort, such as mercaptopurine, which has already been characterized in pediatric T-ALL cases [61]. This does not preclude that these chemotherapies indeed cause mutations in leukemic cells that progress in the relapse. Since upon treatment chemotherapy mutations will be private to each blast, and likely many of them survive into the relapse, the variant allele frequency of these treatment mutations will never rise above the limit of detection of the sequencing. The detection in the relapse T-ALL population [61] of these treatment mutations would require that only one or few blasts survived the treatment, guaranteeing that sufficient numbers of cells in the relapse carried the same mutations to make them detectable through sequencing. The absence of treatment footprints in the relapse is therefore another evidence that the relapse population at the time of treatment already contains a large number of cells.

One intriguing result is the detection of multiple mutations affecting the NOTCH pathway in the same T-ALL case, which do not appear to be exceptions, but rather the rule. It is possible that mutations affecting different domains of NOTCH1 increase the fitness of leukemic cells more than a single mutation, and provide an advantage for relapse. Further studies comparing the pattern of NOTCH1 mutations in relapsing and non-relapsing T-ALLs are needed to clarify this.

Conclusions

All results show that, in the T-ALL patients of this cohort, the relapse is driven by genetic mutations that appear in the population of blasts

several months before diagnosis, giving rise to a resistant subclone of up to several million cells at the beginning of treatment. Upon treatment thus, this subclone comes to dominate the T-ALL population at relapse.

References

1. Acute lymphoblastic leukaemia (ALL) incidence statistics | Cancer Research UK [Internet]. [cited 2020 Mar 16]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-all/incidence?_ga=2.138922035.1884636715.1584377747-1833693179.1584377747#heading-Four

2. Acute lymphoblastic leukaemia (ALL) mortality statistics | Cancer Research UK [Internet]. [cited 2020 Mar 16]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-all/mortality#heading-Two

3. Bhojwani D, Pei D, Sandlund JT, Jeha S, Ribeiro RC, Rubnitz JE, et al. ETV6-RUNX1-positive childhood acute lymphoblastic leukemia: Improved outcome with contemporary therapy. Leukemia. Nature Publishing Group; 2012;26:265–270.

4. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LAA, Miller CB, et al. Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. 2009;

5. Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. Nature Research; 2012;481:157–63.

6. Roberts KG, Morin RD, Zhang J, Hirst M, Zhao Y, Su X, et al. Genetic Alterations Activating Kinase and Cytokine Receptor

Signaling in High-Risk Acute Lymphoblastic Leukemia. Cancer Cell. 2012;22:153–166.
7. Lilljebjörn H, Rissler M, Lassen C, Heldrup J, Behrendtz M, Mitelman F, et al. Whole-exome sequencing of pediatric acute lymphoblastic leukemia. Leukemia. 2012;26:1602–7.
8. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:242–52.
9. Shah S, Schrader KA, Waanders E, Timms AE, Vijai J, Miething C, et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:1226–1231.
10. Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang Y-L, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med. 2014;371:1005–15.
11. Lindqvist CM, Nordlund J, Ekman D, Johansson A, Moghadam BT, Raine A, et al. The mutational landscape in pediatric acute lymphoblastic leukemia deciphered by whole genome sequencing. Hum Mutat. 2015;36:118–128.
12. Zhang J, Mccastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet. 2016;48.
13. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature [Internet]. Nature Publishing Group; 2018; Available from: http://www.nature.com/doifinder/10.1038/nature25795
14. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet. 2019;51:296–307.

15. Mullighan CG, Downing JR. Global Genomic Characterization of Acute Lymphoblastic. Semin Hematol. 2009;46:3–15.
16. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. The Lancet. Elsevier Ltd; 2013;381:1943–1955.
17. Hunger SP, Mullighan CG. Redefining ALL classification : toward detecting high-risk ALL and implementing precision medicine. Blood. 2015;125:3977–3988.
18. Pui CH, Pei D, Coustan-Smith E, Jeha S, Cheng C, Bowman WP, et al. Clinical utility of sequential minimal residual disease measurements in the context of risk-based therapy in childhood acute lymphoblastic leukaemia: A prospective study. Lancet Oncol. 2015;16:465–474.
19. Belver L, Ferrando A. The genetics and mechanisms of T cell acute lymphoblastic leukaemia. Nat Rev Cancer. Nature Publishing Group; 2016;16:494–507.
20. Inaba H, Azzato EM, Mullighan CG. Integration of next-generation sequencing to treat acute lymphoblastic leukemia with targetable lesions: The St. Jude Children's Research Hospital approach. Front Pediatr. 2017;
21. Iacobucci I, Mullighan CG. Genetic basis of acute lymphoblastic leukemia. J Clin Oncol. 2017;35:975–983.
22. Genescà E, Morgades M, Montesinos P, Barba P, Gil C, Guàrdia R, et al. Unique clinico-biological, genetic and prognostic features of adult early T cell precursor acute lymphoblastic leukemia. Haematologica. 2019;haematol.2019.225078.
23. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, et al. Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. Science. 2008;322:1377–1380.
24. Yang J, Bhojwani D, Yang W. Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in

childhood acute lymphoblastic leukemia. \ldots. 2008;112:4178–4183.
25. Mullighan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. NIH Public Access; 2011;471:235–9.
26. Meyer JA, Wang J, Hogan LE, Yang JJ, Dandekar S, Patel JP, et al. Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:290–294.
27. Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat Med. Nature Publishing Group; 2013;19:368–71.
28. Kunz JB, Rausch T, Bandapalli OR, Eilers J, Pechanska P, Schuessele S, et al. Pediatric T-cell lymphoblastic leukemia evolves into relapse by clonal selection, acquisition of mutations and promoter hypomethylation. Haematologica. 2015;100:1442–1450.
29. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, Rusch M, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun. Nature Publishing Group; 2015;6:1–12.
30. Oshima K, Khiabanian H, da Silva-Almeida AC, Tzoneva G, Abate F, Ambesi-Impiombato A, et al. Mutational landscape, clonal evolution patterns, and role of RAS mutations in relapsed acute lymphoblastic leukemia. Proc Natl Acad Sci U S A. National Academy of Sciences; 2016;113:11306–11311.
31. M., Dobson S, García-Prat L, Vanner RJ, Wintersinger J, Waanders E, Gu Z, et al. Relapse fated latent diagnosis subclones

in acute B lineage leukaemia are drug tolerant and possess distinct metabolic programs. Cancer Discov. 2020;canres.0472.2019.
32. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2010;42:338–342.
33. Neumann M, Heesch S, Schlee C, Schwartz S, Gökbuget N, Hoelzer D, et al. Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. Blood. 2013;121:4749–4752.
34. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:186–90.
35. Neumann M, Vosberg S, Schlee C, Heesch S, Schwartz S, Gökbuget N, et al. Mutational spectrum of adult T-ALL. Oncotarget. 2015;6:2754–2766.
36. Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat Genet. 2017;49:1211–1218.
37. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med. 2015;373:2336–2346.
38. Paulsson K, Lilljebjörn H, Biloglav A, Olsson L, Rissler M, Castor A, et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2015;47:672–677.
39. Spinella J-F, Cassart P, Richer C, Saillour V, Ouimet M, Langlois S, et al. Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations. Oncotarget. 2016;7:65485–65503.

40. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2014;46:116–25.
41. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. Nature Research; 2013;500:415–421.
42. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555:321–327.
43. Liu Y-F, Wang B-Y, Zhang W-N, Huang J-Y, Li B-S, Zhang M, et al. Genomic Profiling of Adult and Pediatric B-cell Acute Lymphoblastic Leukemia. EBioMedicine. 2016;8:173–183.
44. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
45. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51:1732–40.
46. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177:101–14.
47. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. Nature Publishing Group; 2015;47:1402–1407.
48. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. ElsevierCompany.; 2018;25:2308–2316.e4.

49. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun [Internet]. Springer US; 2019;10. Available from: http://dx.doi.org/10.1038/s41467-019-11037-8

50. Mészáros B, Kumar M, Gibson TJ, Uyar B, Dosztányi Z. Degrons in cancer. Sci Signal. 2017;10:eaak9982.

51. Richter-Pechańska P, Kunz JB, Hof J, Zimmermann M, Rausch T, Bandapalli OR, et al. Identification of a genetically defined ultra-high-risk group in relapsed pediatric T-lymphoblastic leukemia. Blood Cancer J. 2017;7.

52. Shan H, Li X, Xiao X, Dai Y, Huang J, Song J, et al. USP7 deubiquitinates and stabilizes NOTCH1 in T-cell acute lymphoblastic leukemia. Signal Transduct Target Ther. 2018;3:29.

53. Q J, Ca M, Km A, Y Z, Bt G-D, Kk W, et al. USP7 Cooperates with NOTCH1 to Drive the Oncogenic Transcriptional Program in T-Cell Leukemia. Clin Cancer Res. 2018;25:222–39.

54. Saito Y, Koya J, Araki M, Kogure Y, Shingaki S, Tabata M, et al. Landscape and function of multiple mutations within individual oncogenes. Nature [Internet]. 2020 [cited 2020 May 26]; Available from: http://www.nature.com/articles/s41586-020-2175-2

55. Mansour MR, Duke V, Foroni L, Patel B, Allen CG, Ancliff PJ, et al. NOTCH1 mutations are secondary events in some patients with T-cell acute lymphoblastic leukemia. Clin Cancer Res. 2007;13:6964–6969.

56. Xu F, Wu LY, Chang CK, He Q, Zhang Z, Liu L, et al. Whole-exome and targeted sequencing identify ROBO1 and ROBO2 mutations as progression-related drivers in myelodysplastic syndromes. Nat Commun. 2015;6.

57. Waanders E, Gu Z, Dobson SM, Antić Ž, Crawford JC, Ma X, et al. Mutational Landscape and Patterns of Clonal Evolution in Relapsed Pediatric Acute Lymphoblastic Leukemia. Blood Cancer Discov. 2020;
58. Xiang J, Wang G, Xia T, Chen Z. The depletion of PHF6 decreases the drug sensitivity of T-cell acute lymphoblastic leukemia to prednisolone. Biomed Pharmacother. Elsevier; 2019;109:2210–2217.
59. Kosztyu P, Bukvova R, Dolezel P, Mlejnek P. Resistance to daunorubicin, imatinib, or nilotinib depends on expression levels of ABCB1 and ABCG2 in human leukemia cells. Chem Biol Interact. Elsevier Ireland Ltd; 2014;219:203–210.
60. Ankathil R. ABCB1 genetic variants in leukemias: current insights into treatment outcomes. Pharmacogenomics Pers Med. Dove Press; 2017;Volume 10:169–181.
61. Li B, Brady SW, Ma X, Shen S, Zhang Y, Li Y, et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood. 2020;135:41–55.
62. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. Nature Publishing Group; 2020;578:122–8.
63. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Ann Hum Biol. 2013;40:463–71.
64. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Research. 2020;9:63.

65. Shen R, Seshan VE. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44.
66. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:333–339.
67. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. Genome Biology; 2016;17:1–11.
68. COSMIC. https://cancer.sanger.ac.uk/cosmic/signatures/SBS/.
69. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. Nature Publishing Group; 2020;1–18.
70. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696.
71. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science. 2020;367:1449–1454.
72. Sentís I , Gonzalez S , Genescà E, Garcia-Hernández V , Muiños F , Gonzalez C, Lopez-Arribillaga E, Gonzalez J, Fernandez-Ibarrondo L, Mularoni L , Espinosa L , Bellosillo B Ribera JM , Bigas A , Gonzalez-Perez A , Lopez-Bigas N. The evolution of adult T-cell acute lymphoblastic leukemia. European Genome-phenome Archive. https://ega-archive.org/search-results.php?query=EGAS00001004750 EGAS00001004750

73. Sentís I , Gonzalez S , Genescà E, Garcia-Hernández V , Muiños F , Gonzalez C, Lopez-Arribillaga E, Gonzalez J, Fernandez-Ibarrondo L, Mularoni L , Espinosa L , Bellosillo B Ribera JM , Bigas A , Gonzalez-Perez A , Lopez-Bigas N. Code of the analysis performed in the T-ALL relapse evolution in adult patients project.https://github.com/bbglab/evolution_TALL_adults /DOI: 10.5281/zenodo.4120326 (2020).

Figures

Fig. 1

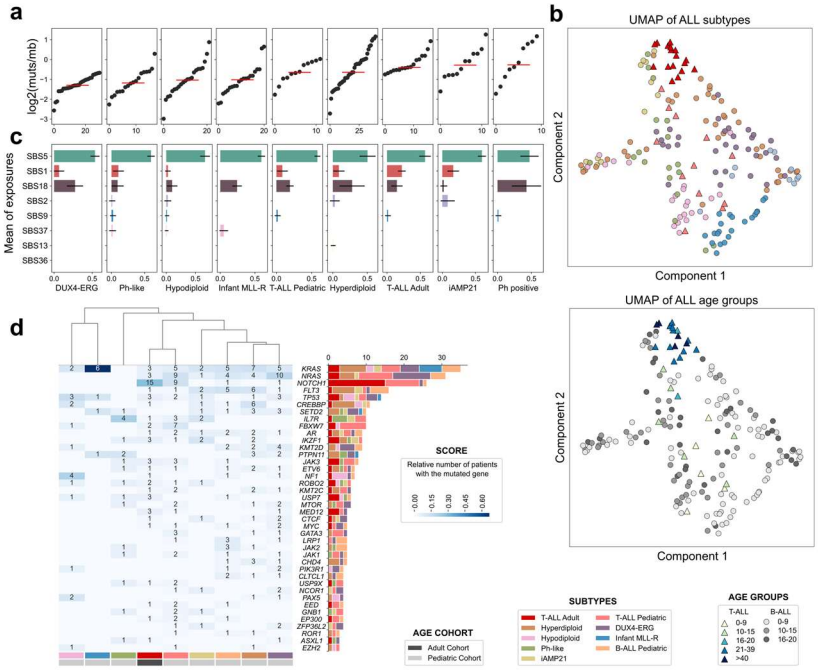


Fig. 1. Comparison of primary adult and pediatric ALL cases.
a) Clonal mutation burden (per megabase) of primary T-ALLs of nine cohorts. Red line shows the median mutation burden of the cohorts. Tumors are represented as dots, sorted along the x-axis according to their mutation burden.
b) Mutational profiles of primary ALLs in the nine cohorts in a Uniform manifold approximation and projection (UMAP) dimensionality reduction graph (see Methods). The UMAP was run on a matrix of the counts of all possible tri-nucleotide changes (96) across ALL patients of all cohorts. Each dot represents a patient, colored according to their cohort (top panel) or their age (bottom panel).
c) Mutational processes active across primary ALL cohorts, represented by their mean (and standard deviation) contribution of the mutation burden of each cohort. SBS1, SBS5, SBS2, SBS9, SBS37, SBS13, SBS36, respectively, single nucleotide substitutions signatures 1,5,2,9,37,13,36.
d) Rate of mutations of selected frequently mutated cancer genes across primary T-ALL cohorts. Cohorts are clustered according to the similarity in their profile of cancer genes mutation frequency (see Methods). The total number of patients in each cohort with mutations of each cancer gene are represented by bars at the right side of the graph. Here are represented genes with mutations in at least two patients (for the full list see Additional file 2: Fig. S2 and Additional file 1: Tables S3 and S4)

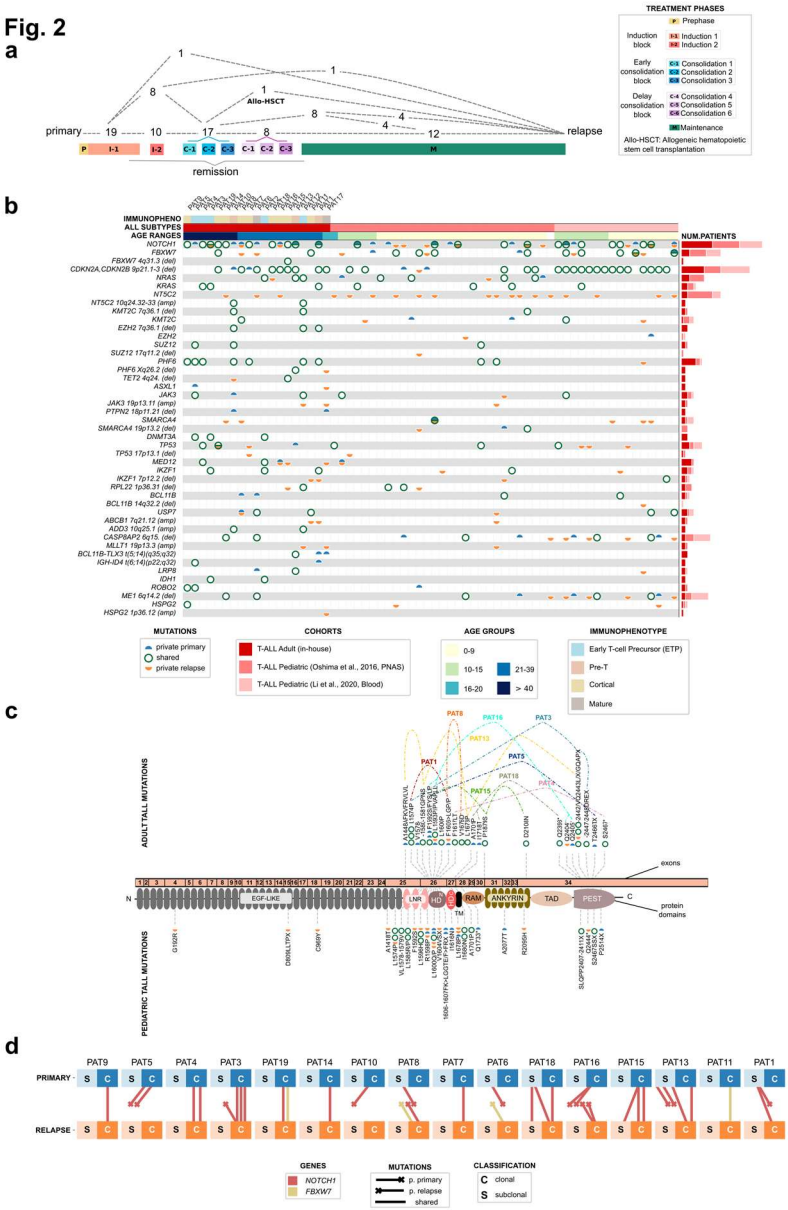


Fig. 2. Comparison of different age groups in T-cell acute lymphoblastic leukemia.

a) Schematic representation of the clinical course of all patients in the in-house T-ALL cohort. Colored boxes (following the legend) at the bottom depict common stages in this clinical course. The broken lines represent specific trajectories followed by groups of patients, with the numbers in each trajectory.

b) Summary of driver mutations (single nucleotide variants, InDels, copy number variants and translocations) identified in the primary and/or relapse T-ALLs of adult and pediatric patients. The original cohorts and ages of the samples included in the

table are indicated above it. The sample where the mutation is identified (primary, relapse, or both) is indicated by color semicircles and circumference at each cell of the table. The total number of patients affected by mutations of each gene are indicated as bars at the right-side of the graph. The table contains the genes that have alterations in at least two patients of the adult cohort (for full table see Additional file 2: Fig. S4 and Additional file 1: Table S5)
c) Protein affecting mutations identified in NOTCH1 gene within adult (above graph) and pediatric (below graph) T-ALLs. Multiple mutations in one patient are represented as dashed colored lines that connect the mutated positions.
d) Clonality change in multi-mutated NOTCH1 pathway genes. Blue and orange squares depict, respectively, primary and relapse T-ALL samples of each patient. Lines connecting them represent shared (connecting lines) or private (lines ending in a cross) NOTCH1 or FBXW7 mutations. In seven out of 19 patients only one mutation in this pathway is identified, while in the other 9 multiple mutations are detected. We did not detect any mutation affecting this pathway in only 3 of the 19 patients.

Fig. 3

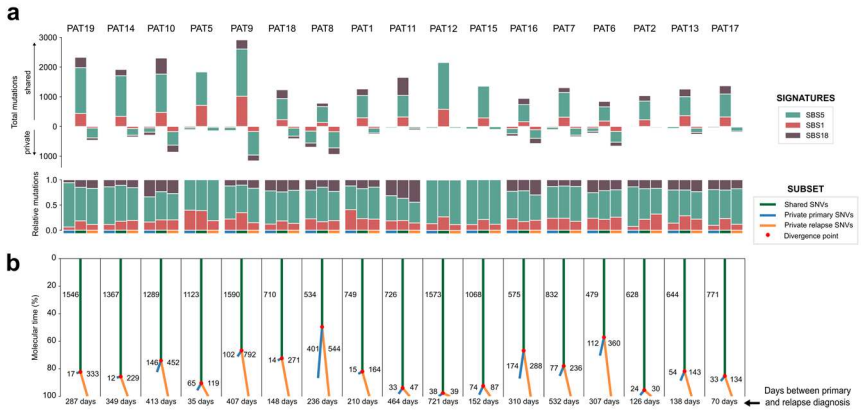


Fig. 3. Shared and private mutations in major primary and relapse T-ALL clones.
a) Contribution of different mutational processes to the mutation burden of each T-ALL case in the adult cohort. The contribution to primary-private, relapse-private and shared clonal mutations are indicated separately in absolute (top panel) and relative (bottom panel) terms.
b) Molecular evolution of adult T-ALL cases represented in a tree-form showing the number of shared clonal mutations (green trunk), clonal private-primary (blue branch) and clonal private-relapse (orange branch) mutations. Only signature 5 mutations are considered to build the tree (for further explanation see Additional file 2: Fig. S7). The relative length of the trunk and branches is proportional to the number of mutations in the respective group. Patients are sorted by decreasing order of age.

Fig. 4

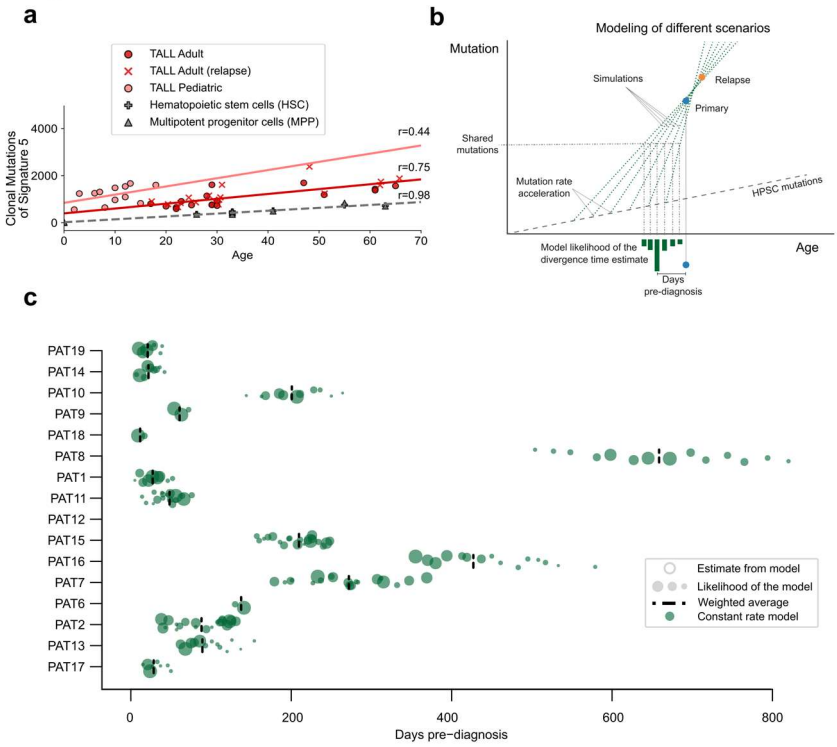


Fig. 4. Time of divergence between major primary and relapse T-ALL clones.
a) Relationship between the mutation rate of ALL samples and the age of patients. Red line shows the regression line estimated from the data points which are the number of mutations attributed to signature 5 (red dots are primary sample and red crosses represent the relapse) of the in-house adult T-ALL cohort. In pink the regression line estimate for the pediatric primary samples (here represented as pink dots). The grey cross and triangle correspond to the signature 5 somatic mutations from healthy tissue (MPP and HSC cells) of Osorio et al., 2018 [48]. Pearson correlation coefficient (r) is indicated above each of the previously mentioned regression lines.
b) Schematic representation of the different mutation rate increment models to decipher the divergence time of the leukemic (primary and relapse) cells.
c) Divergence time of the primary and relapse clone represented as days before diagnosis. The dots are the estimates from the models used and the size of the dots represents their likelihood (see Additional file 2: Fig. S8). The dashed line is the weighted mean of the likely model estimates (see Additional file 1: Table S7).

Fig. 5

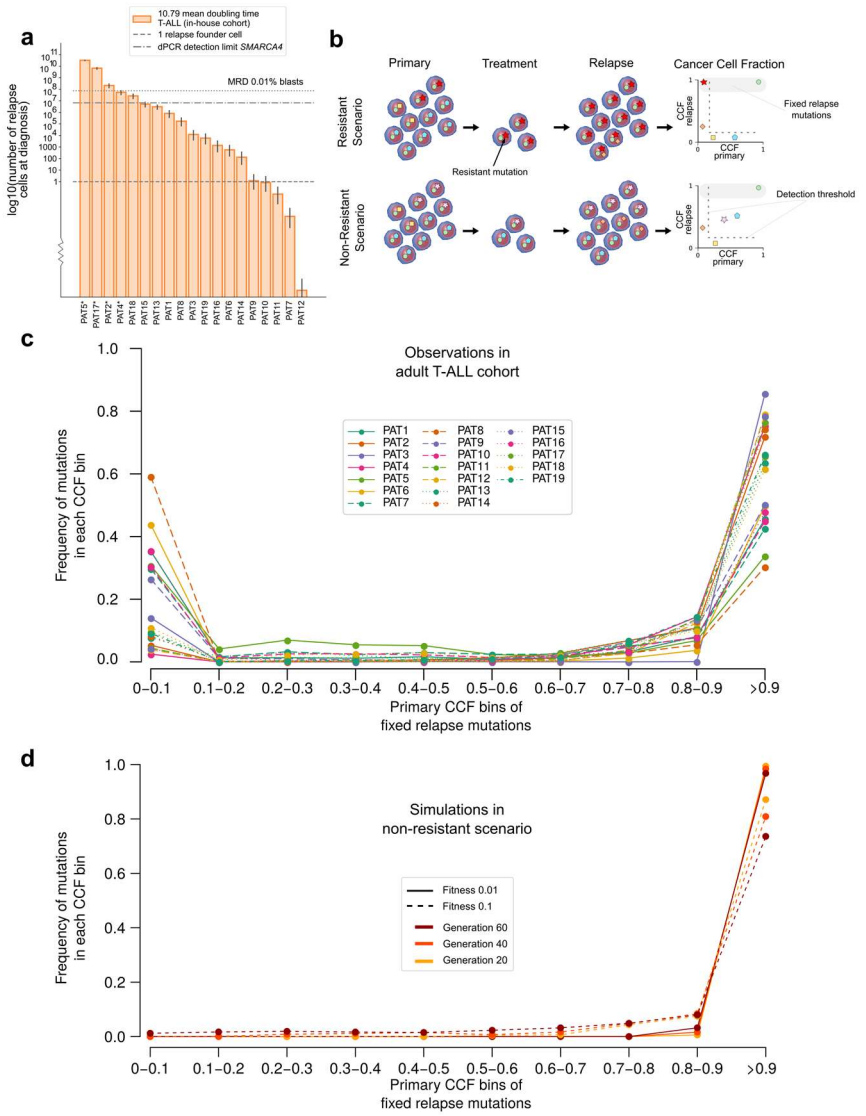


Fig. 5. Evolution of relapse lymphoblast population.
a) Estimated size (number of cells) of the relapse population at the time of diagnosis according to the computed doubling time. Error bars represent the estimates of cell populations from the first and third quartile of the doubling time estimates which are 10.1 and 11.36 respectively (see Additional file 2: Fig. S9). Horizontal dotted lines represent sizes corresponding to one cell and 108 cells (0.01% of the population: the threshold of clinical relapse). Patients with asterisk are the ones with estimates above the pathologist threshold of 0.01. The resolution limit of the dPCR is also represented (~1:10000).
b) Schematic representation of the two considered scenarios of relapse of T-ALL patients after treatment. Mutations in T-ALL cells are represented as different

geometric figures. In the first scenario (resistant), one mutation in the primary T-ALL below the limit of detection of the sequencing and the digital PCR (red star) provides resistance to the treatment. All cells with this mutation survive the bottleneck posed by the treatment, and thus this mutation and all other common to the resistant cells (hitchhikers) are fixated in the relapse population at CCF 1. In the second scenario (non-resistant), a group of cells with an ensemble of mutations survive the treatment.
c) Distribution (frequency) of CCF values of mutations in primary T-ALLs in the in-house cohort that are identified in their relapse counterparts as fixed (>0.9 relapse CCF). Mutations are grouped by CCF bins. Each line represents one patient, for example, the dash brown line corresponds to PAT8, discussed in the text.
d) Distribution (frequency) of CCF values of mutations in synthetic primary T-ALL populations in evolutionary simulations following the non-resistant scenario. The dots represent mutations binned at different CCF values with the frequency that each bin represents with respect to all mutations in each synthetic relapse population. The average results of six simulation settings with different values of fitness of driver mutations and number of cell generations are presented.

Table 1. Summary of ALL cohorts analyzed

ALL subtype cohort name	Subtype cohort information	Reference^	Sequencing	Type	Num. patients
DUX4-ERG	Rearrangement and overexpression of DUX4 and transcriptional deregulation or deletion of the transcription factor gene ERG	St. Jude	WGS	B-ALL	30
Infant MLL-R	Infant patients with a fusion of the N-terminus of the MLL gene with the C-terminus of a partner gene	St. Jude	WGS	B-ALL	21
Ph positive	Patients with the "Philadelphia" chromosome present a translocation of chromosomes 9 and 22. This translocation creates the BCR-ABL fusion	St. Jude	WGS	B-ALL	11
Ph-like	Cell gene expression profile of the lymphoblasts of Ph-like ALL is similar to that of Ph positive ALL; however, they do not present BCR-ABL1 rearrangement	St. Jude	WGS	B-ALL	18
Hyperdiploid	Hyperdiploid patients are characterized by multiple chromosomal gains	St. Jude	WGS	B-ALL	40
Hypodiploid	Hypodiploid patients are characterized by chromosomal losses	St. Jude	WGS	B-ALL	22
iAMP21	Patients with intrachromosomal amplification of chromosome 21	St. Jude	WGS	B-ALL	12
T-ALL Zhang	Patients with T-cell ALL from Zhang et., 2012 *Nat Gen*	St. Jude	WGS	T-ALL	13
T-ALL Oshima	Patients with T-cell ALL from Oshima et al., 2016 *PNAS*	Columbia University	WXS	T-ALL	31*
B-ALL Oshima	Patients with B-cell ALL from Oshima et al., 2016 *PNAS* (B-cell lineage subtype unspecified)	Columbia University	WXS	B-ALL	24*
T-ALL Li[#]	Patients with T-cell ALL from Li et al., 2020 *Blood*	St. Jude	WGS	T-ALL	16*
T-ALL in-house	In-house cohort	In-house	WGS	T-ALL	19*

*Cohorts with primary and relapsed paired samples
[#]Mutations called by the authors of the original analysis; in all other cohorts a uniform mutation calling pipeline was applied

^References: St. Jude cohorts were defined according to their ALL subtype in different publications (see Methods) except for the T-ALL pediatric cohort from Li et al., 2020 [61].
WGS: Whole-genome sequencing
WXS: Whole-exome sequencing

Methods

In-house cohort selection and samples collection

Samples from adults (>= 18 years old) with T-cell acute lymphoblastic leukemia were collected during 15 years under therapy protocols (LAL-07OLD, ALL-HR-03, LAL-AR-2011) as part of the PETHEMA (Programa Español de Tratamientos en Hematología) trials (with the exception of patient 16). Patients have signed the corresponding consents of the protocols. Cohort clinical data is specified in Additional file 2: Fig. S3 and Additional file 1: Table S1. There are three collected samples per patient: one taken at diagnosis (primary), a second one when the percentage of lymphoblasts is reduced during treatment (remission) and a final sample when the leukemia reappears after some months (relapse).

Whole genome sequencing

The short-insert paired-end libraries for the whole genome sequencing were prepared with KAPA HyperPrep kit (Roche Kapa Biosystems) with some modifications. In short, in function of available material 0.1 to 1.0 microgram of genomic DNA was sheared on a Covaris™ LE220-Plus (Covaris). The fragmented DNA was further size-selected for the fragment size of 220-550bp with Agencourt AMPure XP beads (Agencourt, Beckman Coulter). The size selected genomic DNA fragments were end-repaired, adenylated and ligated to Illumina platform compatible adaptors with Unique Dual matched indexes or Unique Dual indexes with unique molecular identifiers (Integrated DNA Technologies).

The libraries were quality controlled on an Agilent 2100 Bioanalyzer with the DNA 7500 assay for size and the concentration was estimated using quantitative PCR with the KAPA Library Quantification Kit Illumina® Platforms (Roche Kapa Biosystems). To obtain sufficient amount of libraries for sequencing it was necessary for the low input libraries (0,1 - 0,2 ug) to amplify the ligation product with 5 PCR cycles using 2x KAPA-HiFi HS Ready Mix and 10X KAPA primer mix (Roche Kapa Biosystems). The libraries were sequenced on HiSeq 4000 or NovaSeq 6000 (Illumina) with a paired-end read length of 2x151bp. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (HiSeq 4000 RTA 2.7.7 or NovaSeq 6000 RTA 3.3.3).

Analysis of ALL cohorts in the public domain

We downloaded public whole-genome and whole-exome sequencing data from EGA and dbGap. We included samples from St. Jude Children's Research Hospital associated with EGAD00001001052 and EGAD00001001432 EGA accession codes. We have used the samples from which we could recover clinical information given with the associated publications [5,8,10,37,38]. We downloaded the DNA sequencing data of Oshima et al., 2016 [30] from dbGap under the accession code phs001072.v1.p1. The information of the cohorts with the clinical information that we could gather for each sample is summarized in Additional file 1: Table S2.

For some of the samples we could not find information regarding the sex so in those cases we inferred it from the normal sample BAM of

each patient. For that, we applied the following reasoning: (1) we determined that the patient is a female if the average coverage of chromosome X is greater than the minimum of average coverages of the autosomal chromosomes and (2) we also assumed that the patient is a female if the mean coverage of chromosome Y is 10 times smaller than the average coverage of the autosomal chromosomes of the sample.

All the samples in Additional file 1: Table S2 have been analyzed with the same pipeline (for detailed information see the following section: Alignment and variant calling). However, in order to compare the T-ALL Adult cohort with other T-ALL cohorts with pre- and post-treatment samples we added the mutations reported in the supplementary materials in Li et al., 2020 [61] only in Fig. 2.a and 2.b.

Alignment and variant calling

Alignment, SNV, small InDels: We performed the alignment and calling of mutations (SNVs and small InDels) using Sarek pipeline v2.2.1 [64]. This workflow performs the alignment from raw FASTQ applying the steps referred to as “best practices” according to GATK. We converted the downloaded BAMs from public repositories to FASTQ with biobambam v2.0.72 and used them as input for the pipeline. We used the Strelka caller implemented in Sarek to generate mutation calls. Only the T-ALL adult cohort was aligned with GEM-mapper v3.6 by the CNAG but the calls were done with Strelka. The mutation calls were performed using primary and relapse as tumor samples and the remission as “normal” sample. Variants have been annotated with VEP v.92 run locally with the

canonical flag and using gnomAD r2.0.1 to get population frequencies of the potential polymorphisms.

CNV: We have used FACETS v0.5.6 [65] to call copy number changes in WGS and WES samples. Following FACETS documentation, we first created its input with snp-pileup which imputed common SNPs and made the reference and alternative read counts at nucleotide resolution. We have run snp-pileup with the recommended parameters except for the --min-read-counts that was set to 10,0. We run FACETS for WES as mentioned in the documentation but setting preProcSample function parameters to cval = 15, ndepth = 5, snp.nbhd = 500 and procSample function parameters to cval = 80, min.nhet = 20. Similarly, we run FACETS for the WGS data as preProcSample(snp.nbhd = 5000, ndepth = 5, cval = 75) and procSample(cval = 800, min.nhet = 25).

SV: We ran Delly v0.7.9 [66] to detect duplications, inversions and translocations. First we ran the *call* function and then the *filter* function of Delly for each one of the alterations mentioned. The map-quality parameter of the call function was set to 20 and we also passed a file provided in the github of Delly with regions to exclude through the --exclude argument. The filter function was run with the following parameters: --filter somatic --minsize 0 (expect for duplications which was set to 100) --qual-tra 0.75 --altaf 0.1.

Filtering steps

SNVs and InDels: From the VCF output from Strelka we filtered the calls labeled as PASS and DP from the FILTER column. For the patients with trio samples we recovered the shared mutations between primary and relapse that are not PASS or DP but are present in the original VCF. This was not possible for patients with

paired samples (primary and remission). In addition, we checked for miss-called DNVs (dinucleotide variants) by inspecting consecutive SNV positions with Samtools v1.4.1 and changed the reference and alternative if needed. Once the variants were annotated with VEP, we took the variants in the canonical transcript. In case of more than one consequence type predicted for the same variant we took the most damaging (more impact) one according to VEP. We also filtered out mutations with population frequency greater than 0.01 according to the gnomADg_AF column added. Finally, we discarded low coverage variants as the ones with a total depth of 5 reads. Further details regarding filters applied to called SNVs are provided in Additional file 3.

CNV: We discarded the variants that were called with low reliability. Those are the segments reported with NAs in the cellular fraction and minor allele copy number columns of FACETS output which, to our knowledge, indicate that the region does not have sufficient numbers of heterozygous SNPs to guide good estimates (Additional file 2: Fig. S5).

SV: We converted the VCFs into bedpe format with bcftobedpe function from svtools v0.4.0 and kept the variants with the flag PASS in the FILTER column. We manually check recurrent SV that have not been described before in the literature by performing BLAT of the breakend points (BND) and their flanking regions in the UCSC and discarded those that were Alu regions or mappable to many parts of the genome.

Purity and clonality estimations

We inferred the purity of the samples from the variant allele frequency (VAF) distribution of the mutations as follows. Since the overall ploidy of the samples was mostly around 2 (diploid) we computed density plots of the VAF multiplied by the CNV of each mutation as a rough proxy of the CCF and determined the purity as the maximum point. We recomputed the CCF with the inferred purity and fitted a beta binomial distribution (betabinom function from scipy v1.4.1 python package). For each mutation, we derived a probability from it and categorized them as clonal or subclonal according to a threshold of 0.01 (above or below it respectively). Exceptionally for PAT16, upon inspection of the CCF distributions in primary and relapse samples, we detected a more complex clonal structure, and thus used a threshold of 0.05 for a clearer categorization of the clonality of the mutations.

Signatures analysis

Several runs of deconstructSigs v.1.8.0 [67] were carried out depending on the context of the analysis. Firsty, following the guidelines proposed by Maura et al., 2019 [49], we have included all hematological meaningful described signatures for the fitting of primary samples (see Additional file 2: Fig. S1). From those, we selected the signatures that we believed had a substantial activity in the primary leukemias in at least one patient of the cohort analyzed and re-run deconstructSigs with them (see Fig 1.c). Secondly, we re-fitted the T-ALL adult samples with only those signatures that presented activity (SBS1, SBS5, SBS18) to better estimate their contribution in Fig.3.a. Lastly, we have fitted known-treatment signatures for the primary and relapse samples to see whether there is any contribution of those in the mutational profile of the relapse. In this case, we have included Signature 32 (SBS32) which the

proposed etiology in COSMIC [68] suggests prior treatment with azathioprine. The adult T-ALL patients have not been treated directly with this compound but it is known that azathioprine is metabolized to 6-mercaptopurine which is used in the maintenance phase of received therapy (see Additional file 2: Fig. S3 and Additional file 2: Fig S6). Apart from SBS32, we have also included two treatment signatures recently extracted in Li et al., 2020 [61] as SBSA_new and SBSB_new. They assigned the usage of thiopurines to SBSB_new signature so that is why we have decided to include it. There is not much said about SBSA_new but since pediatric and adult ALL patients receive similar treatment we decided to give it a try in the fitting analisis. In all cases we set the signature cutoff parameter of deconstructSigs to 0.1.

Clustering of driver genes of ALL subtypes

The distances computed to build the dendrogram on Fig. 1d were based on Jensen-Shannon divergence measures between the distributions of the number of patients per mutated gene of each cohort. We only took into account genes with mutations in at least three patients.

Dimensionality reduction

We used a Uniform Manifold Approximation and Projection (UMAP) implemented in the python package umap-learn v0.3.10 to simplify the mutational profiles (96 dimensions that represent each trinucleotide channel) into two dimensions with the size of the local neighborhood (n_neighbors) to 20 and minimal distance (min_dist) of 0.2.

Identification of ALL driver variants

Driver Gene Discovery: We have run the IntOGen pipeline [69] for SNVs and small InDels (https://www.intogen.org/search) locally for each of the defined cohorts (see above). For each one of the outputs we have proceeded as follows. First, we have discarded all genes in Tier 3 and 4 that are not in the Cancer Gene Census (CGC) [70]. Second, we have discarded all genes in all tiers that have been defined as potential artifacts (see this list of genes in https://bitbucket.org/intogen/intogen-plus/src/master/extra/data/artifacts.json). Third, we have manually inspected the remaining genes and defined a list of potential false positives (FP). From this list of suspicious genes, we have discarded those not present in the CancerMine. With the rest of the FP candidates that were present in the CancerMine, we have decided their level of credibility as driver genes of leukemia according to the publications reported. Apart from that, we have also manually searched in PubMed for any other missed relation by CancerMine of the gene and hematopoietic neoplasms (see Additional file 1: Table S3)

Literature lists of cancer genes of ALL: We have defined 3 lists of known driver genes in ALL:

- Genes with SNVs/InDels mutations
- Genes affected by CNV
- Genes affected by SV that are know to drive ALL

The genes and their sources to build these lists are listed in Additional file 1: Tables S4.a,b,c respectively.

Annotations of alterations: For SNVs and InDels we have defined as potential driver all the mutations with a predicted protein affecting consequence type (in the canonical transcript) according to VEP

(transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, start_lost, transcript_amplification, inframe_insertion, inframe_deletion, missense_variant, protein_altering_variant, splice_region_variant, incomplete_terminal_codon_variant, start_retained_variant, stop_retained_variant) in a cancer gene from the list defined as the combination of the results from the Driver Gene Discovery and the curated literature list of SNVs and InDels. Results from that are summarized in Fig. 1d, Additional file 2: Fig. S2 and Additional file 1: Table S5.

For CNV and SV we have flagged the alterations we have found as "known driver" (contained in the curated literature lists respectively) or with "alteration in gene of interest" if it affects any cancer gene related to leukemia of all the lists. In the case of CNV affecting genes of interest, we consider as candidate drivers those oncogenes that are fully amplified and tumor suppressors affected by any deletion. Results are reported with the annotated "classic" Giemsa cytobands by mapping where the BND genomic coordinates fall within them (see Additional file 1: Table S6 a and b.

We have also annotated the genes affected grouping them by some meaningful information such as their protein family, biological process or pathway (see Additional file 2: Fig. S2, Additional file 2: Fig S4 and Additional file 1: Table S4). We created those groups with information from the sources in Additional file 1: Table S4.

Estimations of divergence time

Considering the differences between the mutational burden of T-ALL samples compared with the expected number of mutations of healthy

hematopoietic cells seems clear that some acceleration on the mutation rate has occurred (Fig. 4a). Additionally, the regression between age and signature 5 of healthy cells and T-ALL show close slope (12.21 ∓ 1.24 vs 20.61∓ 6.58, see Fig. 4a and Additional file 2: Fig. S7) but a much higher intercept (22.35 ∓ 45.53 vs 397.4 ∓ 251.81, see Fig. 4a and Additional file 2: Fig. S7). We hypothesize these similarities on slope and differences on intersect can be explained by a late stage acceleration during tumorigenesis that affects in a similar way the different T-ALL samples.

Based on this hypothesis of tumorigenesis acceleration of signature 5 we have built 2 different models which represent the upper and lower boundary of the estimations: (I) the change of mutation rate is a one-time, discontinuous event, shared between primary and relapse; (II) the change on the mutation rate grows linearly during all lifetime of the tumor. In both scenarios, the mutation rate can only increase and both primary and relapse clones are under the same mutational process. In terms of divergence time, the constant model is the most conservative showing the earliest times of divergence between clones, while the linear model is the one generating larger divergences times. The rest of the models based on N acceleration steps will generate estimates within the previous described.

We established 120 different time-points t_n evenly spaced along the 10-year period immediately preceding diagnosis: we refer to them as “acceleration times”, since they are bound to represent the time-points when the mutation rate first deviates from neutral, clock-like behavior. For each acceleration time we first computed a function assigning a plausible mutation rate for each time point, consistently with either the constant or linear model. To this end, we fitted the

mutation curve to go through the average number of mutations of primary and relapse N(t*) at the middle time-point t* between these two events. More specifically, the following conditions must hold:

$$Constant: N(t^*) = N(t_n) + \mu \cdot (t^* - t_n)$$
$$Linear: N(t^*) = N(t_n) \cdot (1 + r)^{t^* - t_n}$$

where the values of μ and r have to be determined, depending on the model used. Now we did 100 stochastic simulations of the mutation curve by randomly sampling 0 or 1 mutations from a beta binomial distribution with a 1-day granularity, only in cases the mutation rate per day exceeds one a smaller granularity has been used. Thus, mean parameter μ(t) may change with time (linear model) while correlation parameter ρ=0.0002, estimated with the dispersion observed on healthy hematopoietic stem cells described on Osorio et al. 2018 [48], remains constant. Therefore the number of mutations simulated at time t is defined recursively as:

$$N(t_m) \sim N(t_{m-1}) + BetaBinom(\mu(t_m), \rho, 1)$$

where (t_m) is either μ (constant model) or log(1+r)·N(t_{m-1}) (linear model). As the 100 stochastic curves generated for each hypothesis (determined by the acceleration time and mutation rate model) cut the time levels at primary and relapse, they cast a distribution of the possible number of mutations about the observed that yields a likelihood that the hypothesis explains well the observed number of mutations at primary and relapse. Thus each combination of acceleration time and mutation rate model has an associated prior likelihood. We calculated the Bayes posterior distribution using the combinations of parameters with a higher success (likelihood) on the cohort which is then used to select the most plausible models underlying the observation, then provide a plausible set of divergence times weighted by the likelihood. In order to avoid the

deviation of the divergence time estimation due to a long tail of low likelihood simulations, only the more likely scenarios have been selected (10% percentile).

Doubling time and lymphoblast population estimates

The doubling time of the T-cell lymphoblast population was estimated following a similar approach as in Li et al., 2020 [61]. We assumed that blast cell growth is consistent with a logistic model, i.e., the population fraction represented by the T-lymphoblast population as a function of time t fits a logistic function of the form:

$$\sigma(t, a) = (1 + e^{-a})^{-1}$$

where a is the parameter of the logistic model and t is assumed to be given in standard time units such that the T-lymphoblast subpopulation reaches 50% of the total population at time t =0.

Assuming the parameter a is known, the doubling time is given by the following expression:

$$T_D = log(2) / a$$

Therefore the doubling time estimate resorts to fitting a logistic model to our data, i.e., provide an estimate for the parameter a.

Our approach intends to provide an estimate of a that corrects for the likely inconsistencies between time annotations provided in the patients' data. We make the general assumption that some error Δt_i has been introduced for each patient P_i when associating a standard time to the T-lymphoblast population measurements -- mainly due to the difficulty to estimate the initial time for paired data points with a low initial T-lymphoblast population fraction. A standard goodness-of-fit criterion for logistic models is given by the cross-entropy loss:

$$C(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n} y_i \, log \, \hat{y}_i + (1 - y_i) \, log \, (1 - \hat{y}_i)$$

where y and $\hat{y}$ are the observed (resp. predicted) data samples.

Our approach intends to simultaneously estimate the errors Δt_i and the parameter a by minimizing the following cross-entropy loss:

$$L(a, \Delta t_1, \ldots, \Delta t_n) = -(\sum_{i=1}^{n} C(y_{i,0};\, t_{i,0};\, \Delta t_i) + C(y_{i,1};\, t_{i,1};\, \Delta t_i))$$

where $C(y;\, t;\, \Delta t) = y \, log \, \sigma(t - \Delta t, a) + (1 - y) \, log \, (1 - \sigma(t - \Delta t, a))$

where for each patient P_i the values $y_{i,0}$ and $y_{i,1}$ are the initial (resp. final) population fractions and the values $t_{i,0}$ and $t_{i,1}$ are the initial (resp. final) times.

Minimization of the cross-entropy L was implemented in Python with the function “minimize” of the scipy.optimize module. For a more robust minimization, we ran it several times with different randomly generated initial values (see Additional file 2: Fig. S9).

Upon estimation of the doubling time T_D, we proceed to compute the number of cells N_d at the time of diagnosis as a function of the time Δt elapsed between diagnosis and relapse:

$$N_d = N_B \cdot f \cdot 2^{-\Delta t \, / \, T_D}$$

where N_B is an estimate of the total number of bone marrow cells in adults (~$7.5 \cdot 10^{11}$ cells according to [61,63]) and f is the frequency of lymphoblasts of the biopsy.

Digital PCR analysis of *SMARCA4* mutations

The dPCR analysis was performed on a QuantStudio 3D dPCR System using the manufacturer's procedure and reagents (ThermoFisher Scientific). Data analysis and chip quality were assessed using the QuantStudio 3D Analysis Suite software online.

Simulations of relapse scenarios

In order to understand how likely our observations at primary and relapse can be obtained under a non-therapy selective scenario, we have performed several simulations using a Wright-Fisher model (https://github.com/gerstung-lab/clonex).

Firstly, we have established a set of parameters based on our observations of primary samples using a mutation rate of 10^{-8} and a total number of driver and passenger positions of 100 (0.01 fitness effect) and 150000 respectively on a population of 10^6 cells. As a result, after 5000 generations the population has fixed a number of driver mutations ranging from 3 to 8 (mean 5.2) and 122 to 753 (mean 505.8) passengers.

Secondly, from the primary population we randomly removed between $9*10^4$ and 10^6 cells to simulate a bottleneck effect. The resulting population has grown for 20, 40 and 60 generations which covers our estimations about the observed dataset (10% CI: 10.83-37.89 generations).

Finally, we have compared the VAF distribution at primary of those variants with a VAF at relapse higher than 90%, considered as fixed mutations, between the observed and simulated non-resistant scenario.

Due to the lack of fixation of low VAF variants in our simulations, two additional scenarios were performed under the previously described strategy: (I) A non-resistant simulation increasing the fitness up to 0.1 (considered as high fitness, [71]) to allow for faster fixation rates. (II) A resistant scenario where the bottleneck consists of the selection of all cells sharing a low population frequency passenger mutation, defined as resistant mutation.

Ethics approval and consent to participate

All patients were included in protocols (LAL-07OLD, ALL-HR-03, LAL-AR-2011) from the PETHEMA group, except PAT16. These protocols were approved by the Institutional Research Board (IRB) of the participating centers and patients provided informed consent before entering into the trials. The study was approved by the Comitè d'Ètica de la Investigació (Research Ethics Committee: PI-16-146) of the Hospital Germans Trias y Pujol (code approval AEC143). The study complies fully with the Helsinki declaration.

Availability of data and materials

The raw data of the genomic sequencing of the 45 samples (primary-remission-relapse) of the patients of the in-house cohort is deposited in the EGA repository (accession code EGAS00001004750; [72]). For the sake of reproducibility, the code of the analysis is available here: https://github.com/bbglab/evolution_TALL_adults (doi:10.5281/zenodo.4120326; [73]). Raw sequencing data of public

datasets produced by St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome Project (see Table 1) was obtained from the EGA repository (accession codes EGAD00001001052 and EGAD00001001432; some BAMS corresponding to published projects somewhere else [5,6,8,10,14,37]). Raw sequencing data of patients included in the study by Oshima et al., 2016 [30] (Table 1) was obtained from dbGap (phs001072.v1.p1). The somatic mutations identified in the patients included in the study by Li et al 2020 [58] were obtained from the Supplementary Data of the original paper.

Competing interests

The authors declare that they have no competing interests

Funding

The authors would like to thank the Asociación Española Contra el Cáncer (AECC) for financially supporting this project (GC16173697BIGA). N.L.-B. acknowledges funding from the European Research Council (consolidator grant 682398) and the ERDF/Spanish Ministry of Science, Innovation and Universities–Spanish State Research Agency/DamReMap Project (RTI2018-094095-B-I00). S.G work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 754510. I.S is supported by FPI fellowship from Spanish Ministry of Economy and Competitiveness (project reference SAF2015-66084-R). V.G-H. is supported by the AECC (project reference GC16173697BIGA-9). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness

(MINECO; Government of Spain) and is supported by CERCA (Generalitat de Catalunya).

Authors' contributions

A.B, JM.R. and N.L-B coordinated the project. I.S and S.G. carried out the analyses and prepared the figures. I.S. collected public sequencing data and re-analyzed them to call mutations systematically. I.S. also did mutation calling of the 19 ALL patient samples of the project and performed the analysis of driver and resistance mutations. S.G. conceived and carried out the analyses of mutation rate acceleration and the development of resistance models in different scenarios, presented in Fig. 4c and 5c and d. F.M. contributed in the design of the statistical model to compute the doubling time. I.S, S.G, A.G.-P. and N.L.-B. participated in the design of computational analyses and in the interpretation of the results. L. M. contributed to the mutation calling. JM.R. and E.G. collected the samples of the adult ALL patients and provided clinical information. I.S., E.G., E.L-A, Ll.E., A.G-P, A.B., JM.R., and N.L-B. participated in discussions of project design, patient data and sample selection. V.G-H., L. F-I and B.B contributed to digital PCR experiments and data analysis. J.G. and C.G. provided technical support to the project. A.G.-P. drafted the manuscript. I.S., S.G., A.G.-P. and N.L.-B. edited the manuscript. A.G.-P. and N.L.-B. supervised the analyses.

Acknowledgements

We would like to acknowledge the contribution of Jordi Deu-Pons and Iker Reyes to the mutation calling and general technical support of the project. We also want to mention Francisco Martínez-Jimenez

for his contribution to the analysis of drivers and Oriol Pich for his help on mutational signature analysis. We are grateful to the St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome Project (PCGP) for permitted access to pediatric data. Also, we would like to thank the data from Columbia University Medical Center Institutional published in Oshima et al., 2016 used in this study.

Supplementary Information

Additional file 1. Additional tables. This file contains the supplementary tables referenced in the main text. Table S1 that contains clinical information on the adult T-ALL cohort. Table S2 contains clinical information of the public pediatric cohorts. Table S3 contains the detected cancer genes by IntOGen. Table S4 contains the lists of ALL cancer genes of interest found in the literature separated in 3 subtables according to the type of alterations: SNVs and InDels (Table S4.a), CNV (Table S4.b), SV (Table S4.c). Table S5 contains the mutations (SNVs and InDels) that we consider as candidate drivers. Table S6 has the candidate driver CNV (Table S6.a) and SV (Table S6.b) of the cohorts analyzed. Table S7 has the time of divergence estimates between primary and relapse estimated as days pre-diagnosis of each patient.

Additional file 2. Additional figures. This file presents all supplementary figures referenced in the main text.

Additional file 3. Additional methods. Some of the filtering steps have been extended for clarification in this file.

Additional file 2

Fig. S1

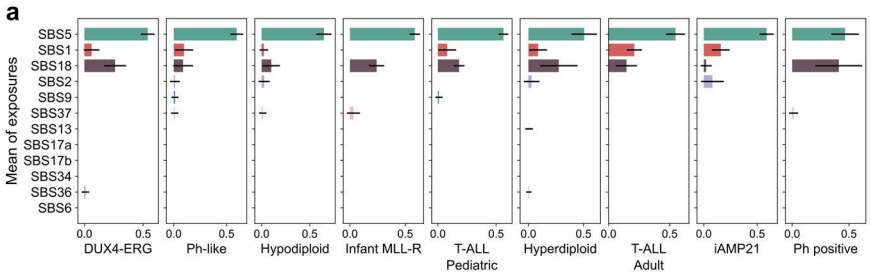


Fig. S1. Probing for active mutational processes across ALL cohorts. Mutational processes active across primary ALL cohorts, represented by their mean (and standard deviation) contribution of the mutation burden of each cohort. The list of signatures to fit was determined by their activity in any hematopoietic cancer according to COSMIC (see Supp. Methods). This linear fitting (Methods), was used to select the subset of mutational processes active in ALL tumors of the studied cohorts which are shown in the main Figures.

Fig. S2

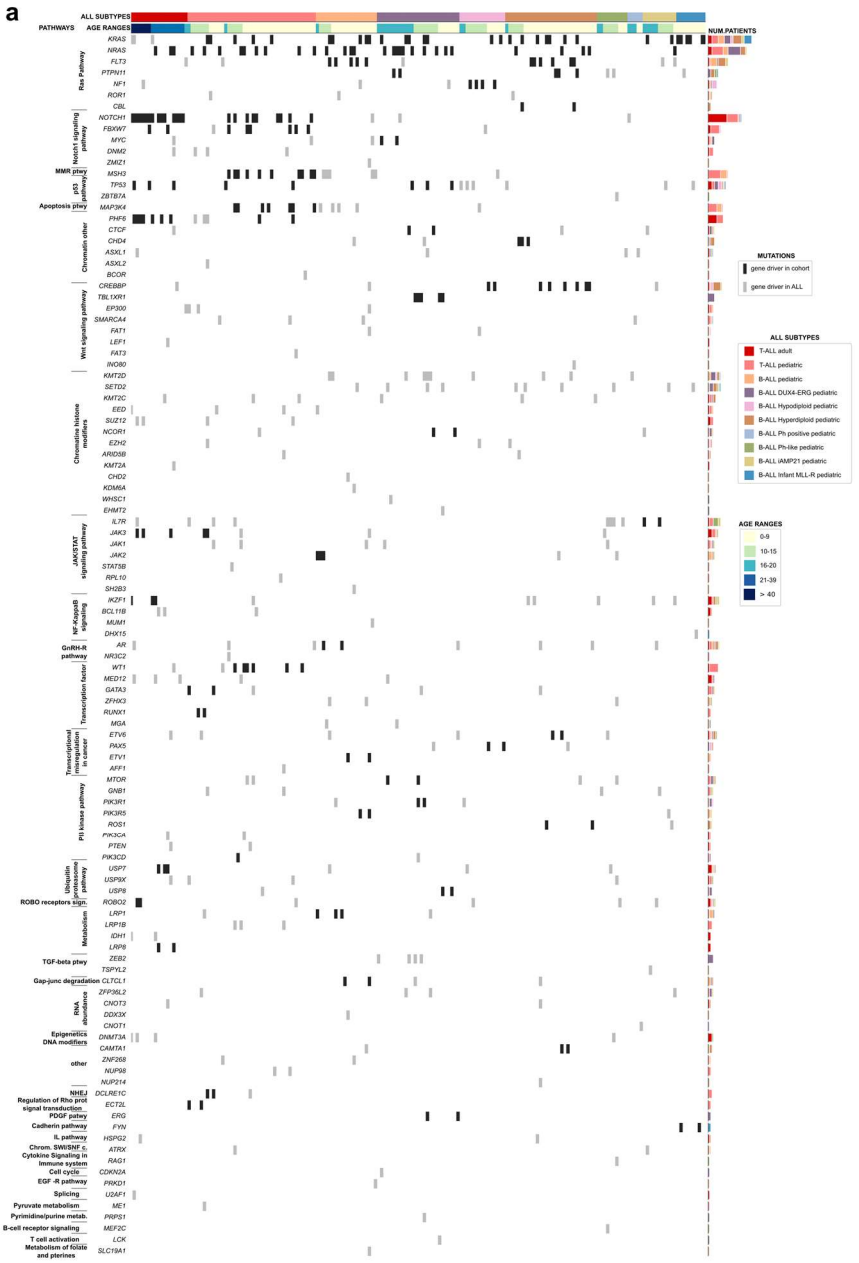


Fig. S2. Mutations in driver genes in primary ALLs.
Rows are driver genes in ALL (collected from the literature or identified across these cohorts using IntOGen; see Supp. Methods) grouped by protein family, biological process or pathway. Columns are ALL samples grouped by cohort, and sorted by age. Each full rectangular cell represents a protein-affecting mutation in a driver gene annotated from the literature (grey) or directly detected as driver in that cohort

through the IntOGen pipeline (black). The bars on the left represent the total number of patients in each ALL cohort with mutations of the gene.

Fig. S3

a

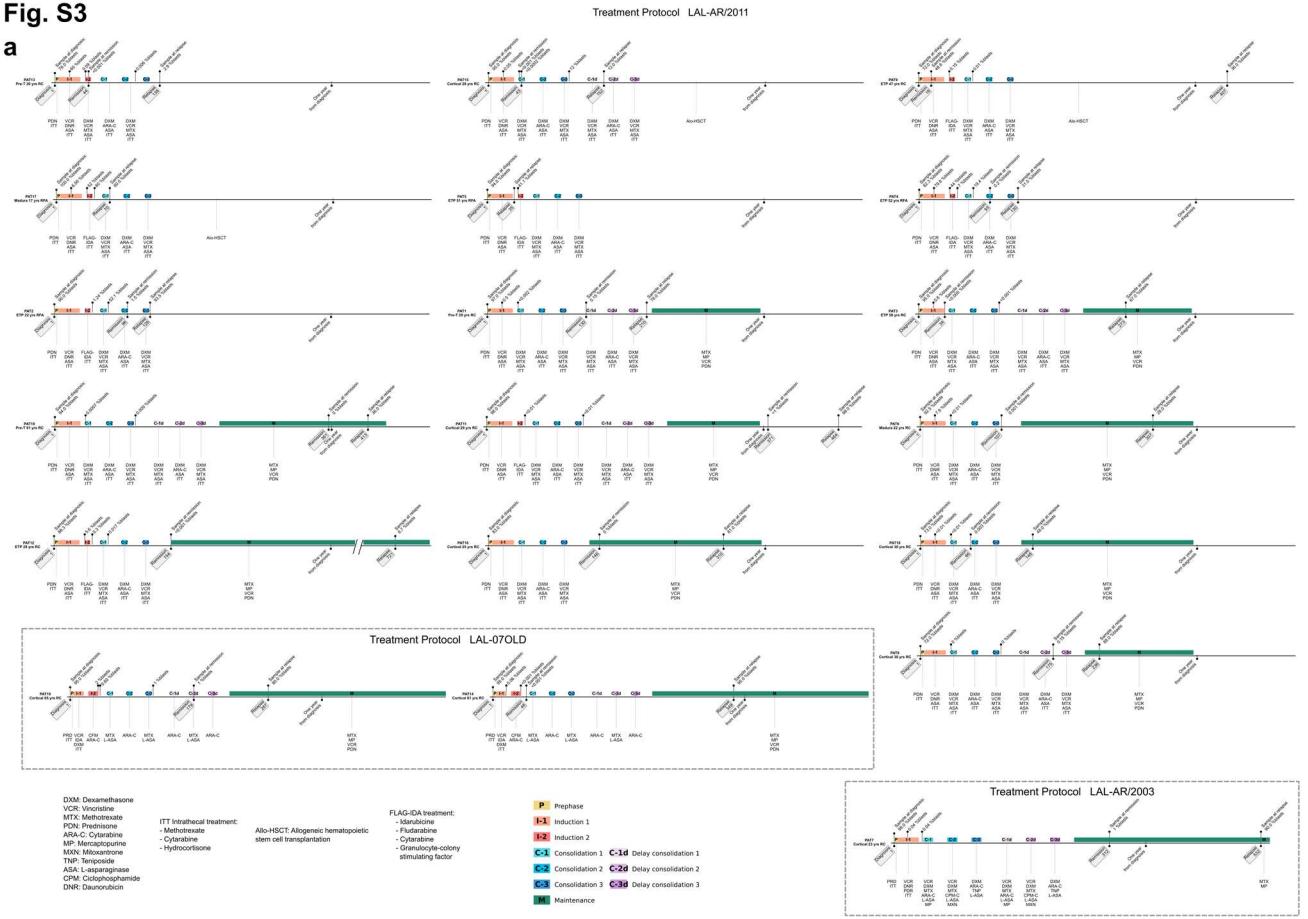

Treatment Protocol LAL-AR/2011
Treatment Protocol LAL-07OLD
Treatment Protocol LAL-AR/2003
DXM: Dexamethasone
VCR: Vincristine
MTX: Methotrexate
PDN: Prednisone
ARA-C: Cytarabine
MP: Mercaptopurine
MXN: Mitoxantrone
TNP: Teniposide
ASA: L-asparaginase
CPM: Cidophosphamide
DNR: Daunorubicin
ITT Intrathecal treatment:
- Methotrexate
- Cytarabine
- Hydrocortisone
Allo-HSCT: Allogeneic hematopoietic stem cell transplantation
FLAG-IDA treatment:
- Idarubicine
- Fludarabine
- Cytarabine
- Granulocyte-colony stimulating factor
P Prephase
I-1 Induction 1
I-2 Induction 2
C-1 Consolidation 1
C-1d Delay consolidation 1
C-2 Consolidation 2
C-2d Delay consolidation 2
C-3 Consolidation 3
C-3d Delay consolidation 3
M Maintenance

Fig S3. Clinical course of patients in the in-house T-ALL cohort

Each figure represents the clinical course of a patient. Day 1 represents the sample taken at diagnosis and the total extension of the line comprises one and a half years. Lymphoblast counts at regular checks and sample extractions performed at the hospital are represented above the line. Colored boxes represent treatment cycles according to clinical protocols (see Supp. Methods), and rectangular grey labels below correspond to sample extraction timepoints. The width of the treatment box is approximately scaled to the time that corresponds to protocols guidelines. Treatments received appear below the timelines. PAT16 most likely received treatment based on protocol LAL-AR/2011, however, this information has never been confirmed.

Fig. S4. Mutations in driver genes in primary and relapse T-ALL tumors of three cohorts. Rows are driver genes in T-ALL (collected from the literature or identified across these cohorts using IntOGen; see Supp. Methods) grouped by protein family, biological process or pathway. Columns are ALL samples grouped by cohort, and sorted by age. We added immunophenotypic information for the in-house T-ALL cohort. Primary-private and relapse-private mutations are represented

as blue and yellow semicircles, respectively. Shared mutations are represented as green circles. The total number of patients affected by mutations of each gene across the three cohorts are indicated by stacked bars at the right-side of the graph. Calls from the X chromosome in the pediatric cohorts are not included (only in adults).

Fig. S5
a

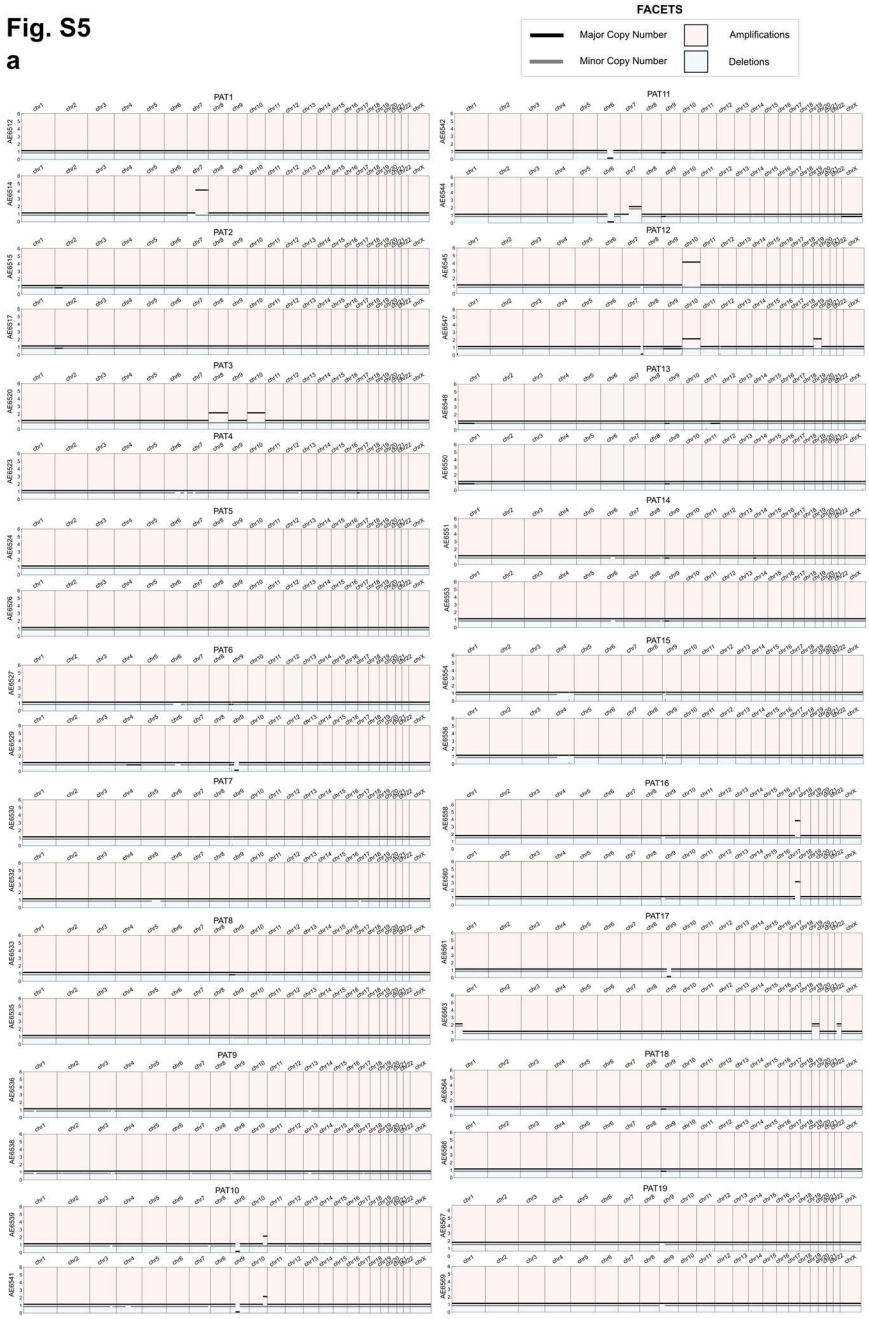


Fig. S5. Copy Number Variants detected in primary and relapse T-ALLs in the in-house cohort.

In each panel, that corresponds to one T-ALL sample in the cohort, chromosomes are represented in the x-axis, with their copy number in the y-axis. Copy number of the major allele is represented as a black line and that of the minor allele as a grey line. In diploid segments, both the black and grey lines appear close to 1 (total sum

of 2) but not overlapping for visual purposes. Red and blue shaded backgrounds represent amplifications and deletions, respectively. For all patients (indicated in the graph title), the top and bottom plots correspond to the primary and relapse samples, respectively.

Fig. S6

a

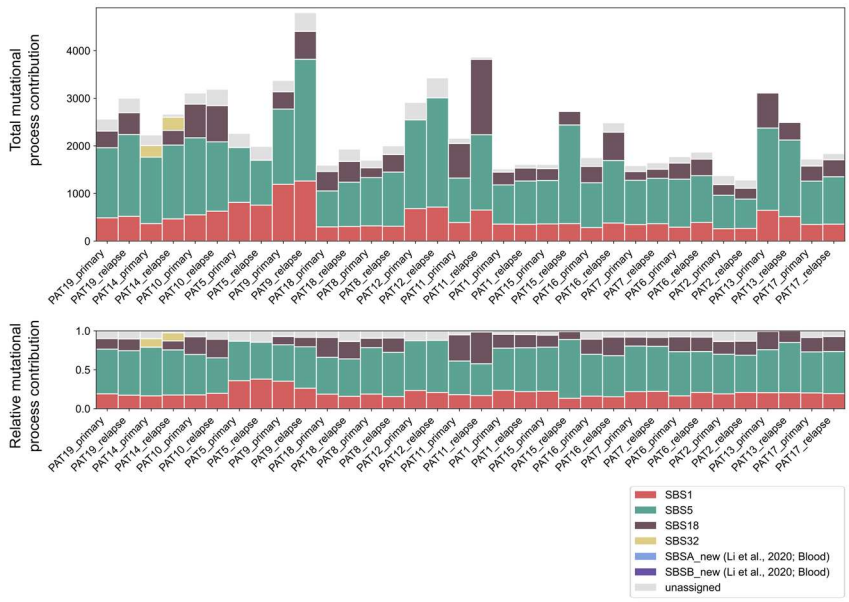


Fig. S6. No mutational footprints attributable to treatments are detected. Panels represent the absolute (top) and relative (bottom) contribution of mutational processes (signatures listed in legend) to the mutation burden of the primary and relapse malignancy of each patient. As indicated in the legend, colors of the bars indicate the signatures used in the fitting process. Note that, although included in the fitting, the mutational signatures (or footprints) of drugs used in ALL treatment (recently identified in pediatric relapse tumors; [57]) are not detected in relapse T-ALLs in the in-house cohort.

Fig. S7

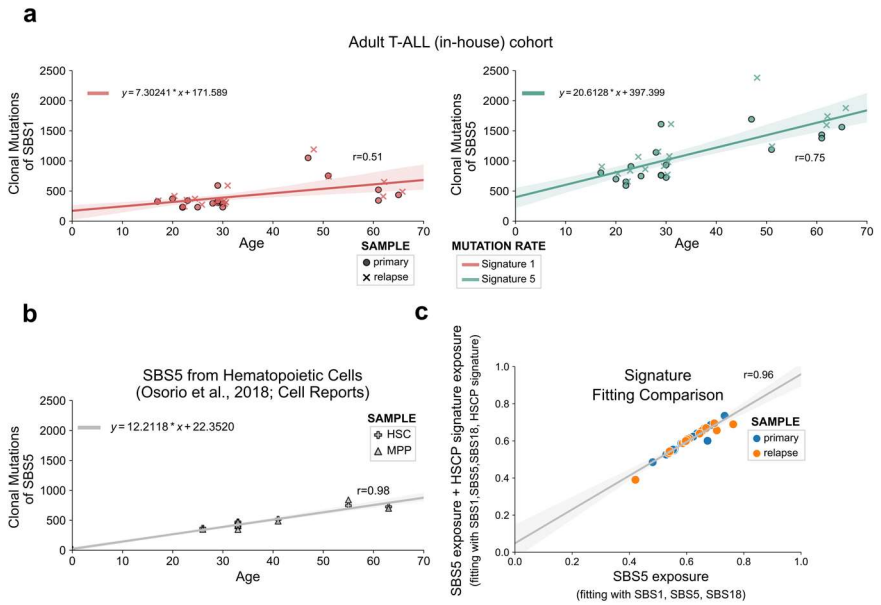
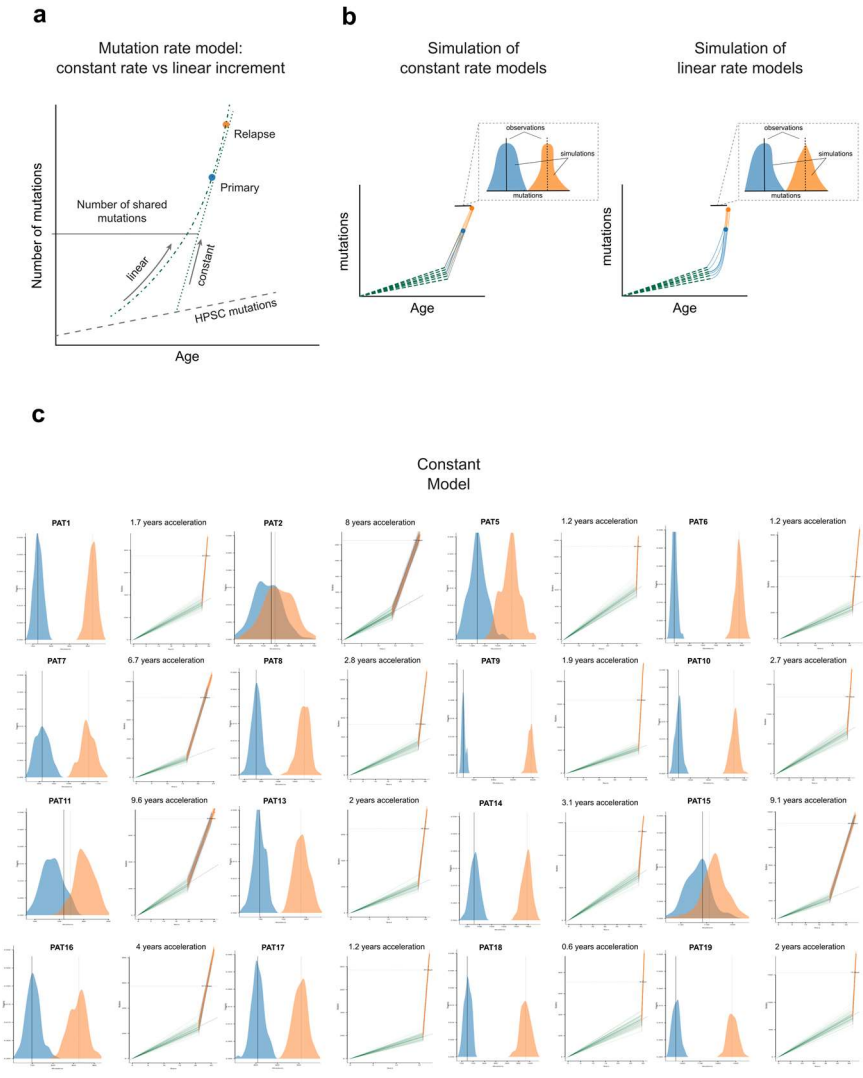


Fig. S7. The contribution of clock-like Signature 5 to the mutation burden of T-ALLs.

a) Known clock-like signatures 1 (left panel) and 5 (right panel) contribute clonal mutations at a steady rate across the lives of T-ALL patients in the in-house cohort. Although the number of clonal mutations contributed by each process significantly correlates with the patients' age (signature 1 $p=2.30x10^{-3}$, signature 5 $p=3.64x10^{-7}$), the correlation is stronger in the case of signature 5, which had been previously observed [46]. Moreover, signature 5 contributes more age-related mutations than signature 1. Dots are primary and cross relapse samples. Trendlines following the regression are added and their equations are indicated at the top right of each panel. b,c) Signature 5 also fits well the steady accumulation of mutations in healthy hematopoietic stem cells and multipotent progenitors (HSC and MPP; Osorio et al., 2018 [47]) with aging. Signature 5 contributed mutations fit very well (r=0.98) the mutational burden of clonally expanded HSC and MPP (b), implying that it is probably the main mutational process active in these cells and that signature 5 mutations accumulate steadily over time. When a de novo signature extraction (rather than the signature deconstruction presented so far in this article; see Supp. Methods) is carried out on the mutational profile of HSC and MPP a specific HSC population (HSCP) signature is extracted [47,48]. The activity (fraction of contributed mutations) of this signature across HSC and MPP cells correlates very well (r=0.96) with that obtained for the fitted signature 5. This supports the idea that the HSCP signature and signature 5 represent the same underlying mutational process active in HSC and MPP.

Fig. S8


a
Mutation rate model:
constant rate vs linear increment
Relapse
Primary
Number of shared
mutations
Number of mutations
linear
constant
HPSC mutations
Age
b
Simulation of
constant rate models
Simulation of
linear rate models
observations
simulations
mutations
mutations
Age
c
Constant
Model
PAT1
1.7 years acceleration
PAT2
8 years acceleration
PAT5
1.2 years acceleration
PAT6
1.2 years acceleration
PAT7
6.7 years acceleration
PAT8
2.8 years acceleration
PAT9
1.9 years acceleration
PAT10
2.7 years acceleration
PAT11
9.6 years acceleration
PAT13
2 years acceleration
PAT14
3.1 years acceleration
PAT15
9.1 years acceleration
PAT16
4 years acceleration
PAT17
1.2 years acceleration
PAT18
0.6 years acceleration
PAT19
2 years acceleration

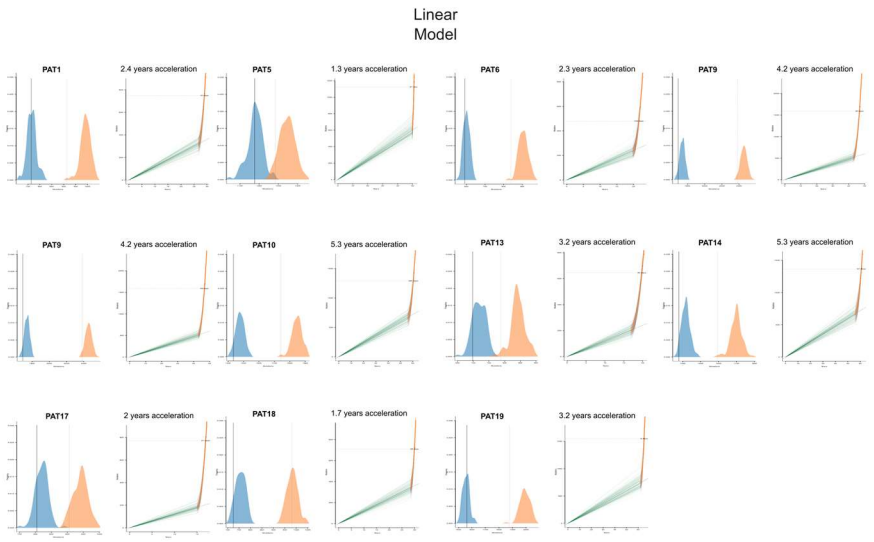


Fig. S8. Models of accelerated mutation rate in T-ALL.
a) Schematic representation of two extreme models of the accelerated mutation rate in T-ALL. The baseline mutation burden increase represents the steady accumulation of mutations of HSC and MPP shown in Supplementary Figure 7. The linear model consists in a steady acceleration of the mutation rate throughout all the evolutionary history of the T-ALL. On the other hand, in the constant model the acceleration occurs only once in the evolution of the T-ALL, which after this point maintains a steady increase of the mutation burden.
b) For both, the constant (left plot) and linear (right plot) models, a number of simulations of accelerated mutation rate are carried out, represented in these schematic graphs by dotted lines. The likelihood of each explaining the observed mutation burden of primary and relapse samples is then computed, as explained in the main manuscript.
c) Real examples of the more likely models given the observed data for each patient of both types of simulations (constant and linear). The years when the mutation rate accelerated and incremented are written above each pair of graphs per patient.

Fig. S9

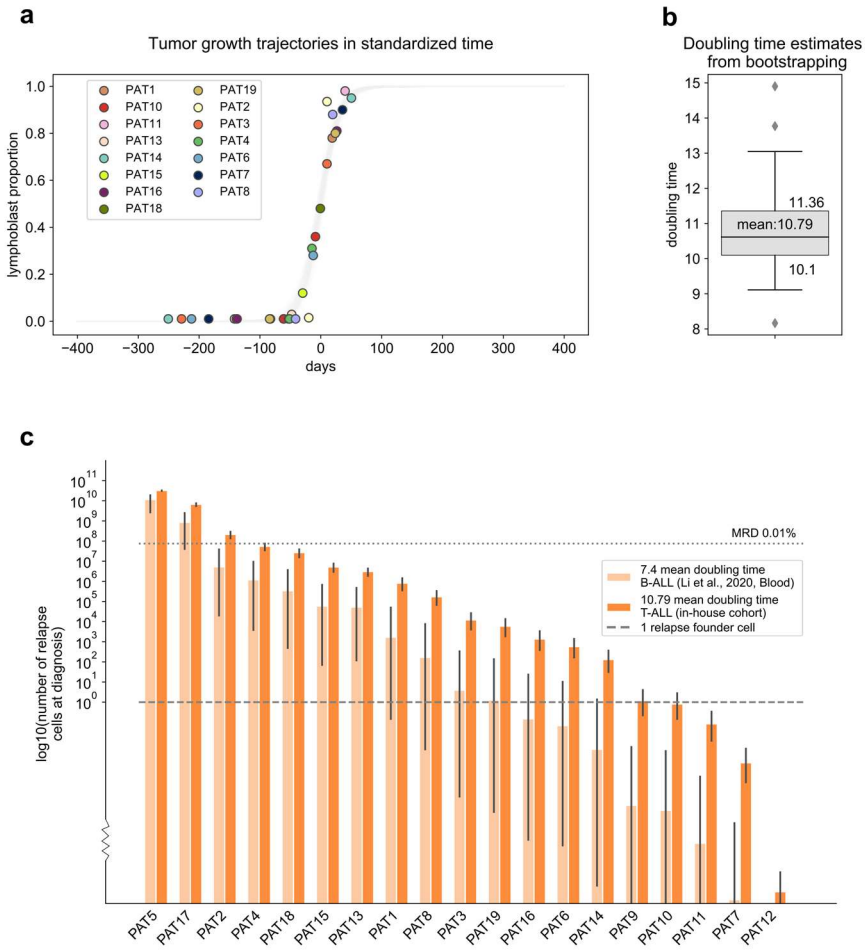


Fig. S9. Estimating the doubling time of T-ALL population from the pathologists' observations.

a) Bootstrapping adjustment of growth logistic curves to the counts of lymphoblasts in remission and relapse bone marrow samples carried out by the pathologist. The observations at these two points (dot pair) for each patient are represented with the same color. Paired-dates of the bone marrow sampling are re-scaled to a standardized time where 0.5 blast proportion (y-axis) falls at day 0 (x-axis) of the growth trajectory of the malignancy of each patient.

b) Boxplot with the doubling time estimates resulting from the bootstrapping. The line at the center of the boxplot is the mean. The first and third quartile of the distribution of bootstrapped doubling times are also represented.

c) Comparison of the number of relapse cells for all patients in the in-house T-ALL cohort computed at time of diagnosis using two different estimates of the doubling time of the lymphoblastic population. Light orange bars represent the size of the T-ALL relapse population computed using the doubling time estimate obtained recently for pediatric B-ALLs (Li et al., 2020) Dark orange bars represent the size of the T-ALL relapse population computed using the doubling time estimate obtained

in (b). Error bars are the estimates corresponding to the doubling time in the first and third quartiles of the distribution.

Fig. S10

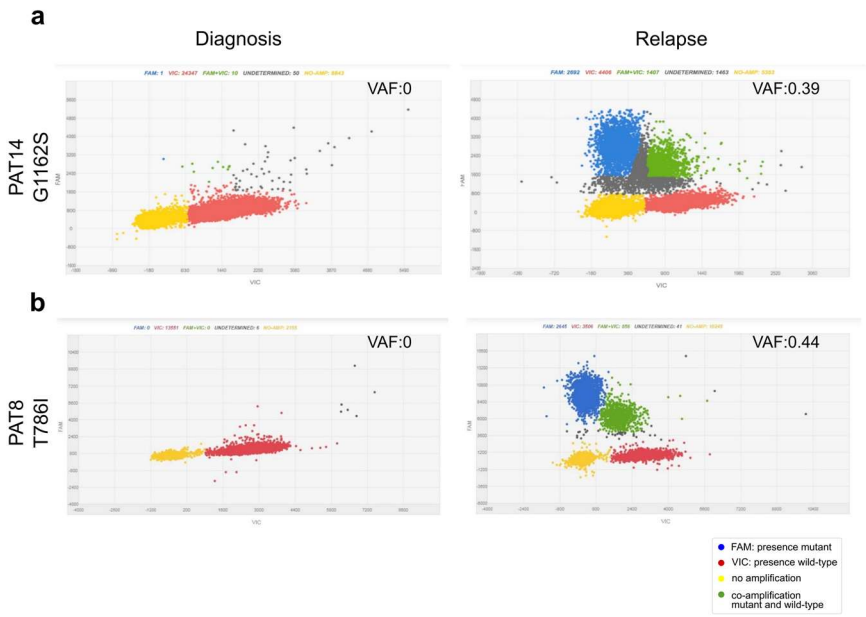


Fig. S10. Results of digital PCR on mutant SMARCA4 in two primary samples. Detection of mutations a) G1162S and b) T786I in the relapse-enriched SMARCA4 gene in primary samples of PAT14 and PAT8 respectively was negative in both (VAF= 0). The resolution of the dPCR in PAT14 was 0.089% whereas in PAT8 it was 0.11%. The VAF detected of the mutants in relapse derived from the dPCR is similar to the one detected by NGS which are both close to the expected 0.5 for an heterozygous variant with no normal contamination (0.39 vs 0.403 and 0.44 vs 0.346). Scatter plots showing the distribution of the data points based on the dyes used (VIC and FAM). Blue dots (FAM) represent presence of mutant SMARCA4 and red dots (VIC) represent wild-type SMARCA4; yellow refers to no amplification and green to co-amplified wild-type and mutant species.

Fig. S11

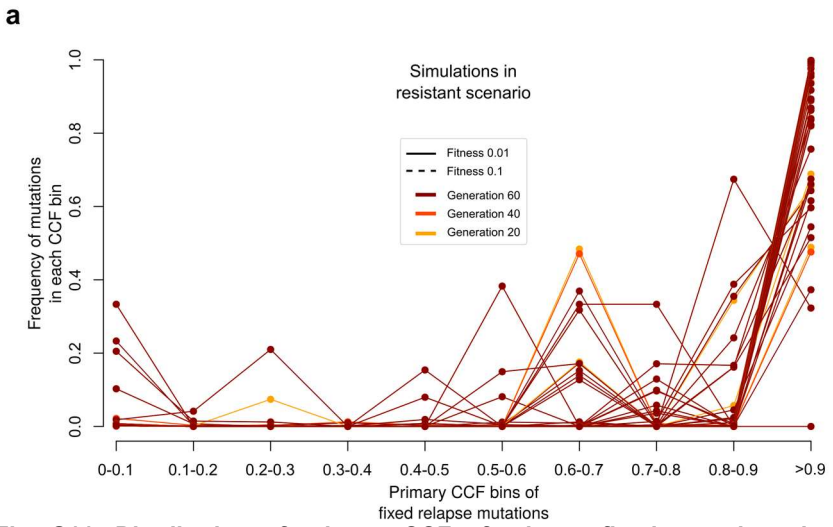


Fig. S11. Distribution of primary CCF of relapse fixed mutations in the simulated resistant scenario.

Distribution (frequency) of CCF values of mutations in synthetic primary T-ALL populations in evolutionary simulations following the resistant scenario defined in the toy example of Figure 5b. The dots represent mutations binned at different CCF values with the frequency that each bin represents with respect to all mutations in each synthetic relapse population. All the results of six simulation settings with different values of fitness of driver mutations and number of cell generations are presented.

Additional File 3

Preprocessing and filters of somatic mutation calls

As explained in the Methods section we have processed the mutations (SNVs and InDels) from the original VCF output of Strelka to the final MAF file of calls analyzed. The first thing we did was to filter out any mutation non-labelled as PASS or DP in the FILTER columns of their corresponding VCF. We noticed that for a few patients the number of mutations in the relapse sample was lower than the primary, contrary to what we would expect taking into account that the relapse cells had more time to accumulate mutations compared to the primary. Therefore, we decided to check whether there were mutations labeled as PASS or DP in the primary that were present in the relapse original VCF that we missed at filtering. We realized that this was the case so we decided to also do the reverse exercise and add the missed calls to the filtered set of mutations of each sample. We have called these shared mutations FISHED (see below Additional file 3: Fig. S1 a).

Another critical point was that we observed substantial differences between samples of the same patient regarding tumor burden as well as within the entire cohort. We suspected that there could be some polymorphisms within the somatic calls of the samples. We used gnomAD to annotate the variants with population frequencies and decide to filter out those with a frequency above 0.01 (see Additional file 3: Fig S1b).

The clonal classification, that is separating clonal from subclonal mutations, is explained in detail in the Methods section. In the Additional file 3: Fig S1 c we are showing the Cancer Cell Fraction

(CCF, see equation below) of each mutation in the primary and relapse samples colored or shaped according to their clonal classification in the primary and relapse respectively. In almost all patients, the shared clonal mutations are a well-defined blue dotted cloud of points with its centroid approximately at CCF 1 of both axis (samples).

$$CCF = \frac{VAF * (p * cn + 2 * (1 - p)}{p}$$

being p the purity of the sample and cn the copy number of the region where the mutation falls. VAF means variant allele frequency and is calculated as follows:

$$VAF = \frac{ar}{tr}$$

where ar refers to reads mapping to the alternative allele and tr the total number of reads mapping to that particular position.

Finally, we were notified by the sequencing center (CNAG) that PAT3 and PAT4 primary samples seemed to have the DNA damaged. In this figure, PAT3 shows evidence of that by its large number of mutations and CCF values. We have not included those samples and as a consequence, these two patients were out in most of our analysis except for reporting protein affecting mutations in known ALL cancer genes of interest.

Fig. S1

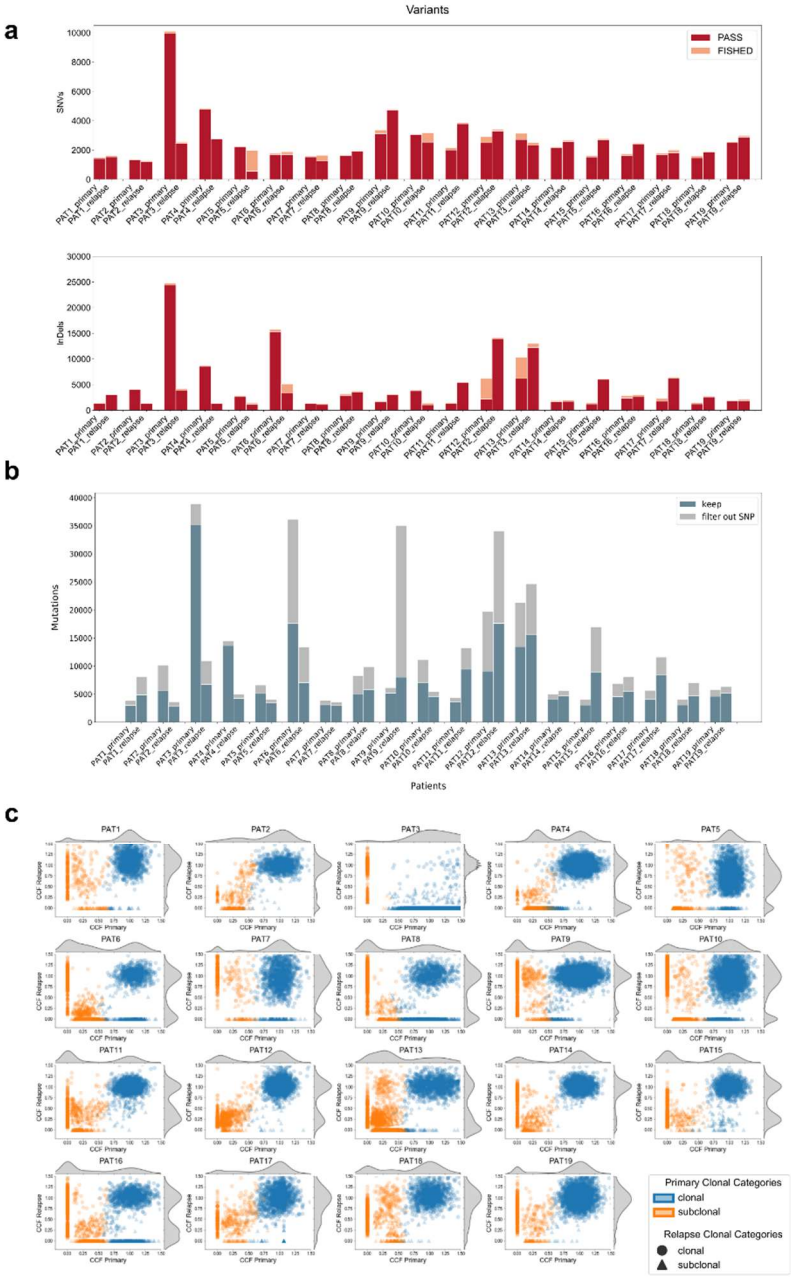


Fig. S1. Filter steps of mutations.
a) Barplot of primary and relapse sample of each patient showing the number of SNVs (up panel) and InDels (low panel) that have the PASS label in FILTER column of the VCF (red) and the rescued shared variants called FISHED (orange). b) Barplot of primary and relapse samples of each patient showing all mutations as the proportion of mutations that were filtered out due to their high frequency (> 0.01) in

the population as annotated by Gnomad and the proportion that are believed to be somatic. c) Scatterplots showing the CCF of each mutation in primary (x- axis) and relapse (y-axis). Color and shape of each data point (mutation) is indicated in the legend.

3.2 Chapter 2

Compendium of mutational cancer driver genes

One of the fundamental aims in cancer research is to discover the compendium of all cancer driver genes, which are those responsible for tumorigenesis and provide the scientific community with new targets for precision medicine. As explained in the above section, one of the objectives of the leukemia project was to find new drivers of ALL and also candidate genes of therapy resistance. With that intention, I joined another project of the lab, the generation of a Compendium of Mutational Cancer Driver Genes across cancer types (IntOGen project). The aims of the IntOGen project were two: i) to provide the research community with an automatic identification workflow of cancer genes and ii) generate a list of cancer driver genes across tumor types. To achieve the second objective the analysis of somatic mutations across a large number of cancer samples was required. At that time, I was already downloading a great amount of tumoral data from pediatric ALL cohorts with the intention to detect signals of positive selection for the identification of driver genes. I extended the search from ALL to somatic mutations in other cancer cohorts, with the aim to collect all datasets of tumor somatic mutations available in the public domain. Dr. Francisco Martínez-Jiménez and I downloaded, curated and annotated the catalogs of somatic mutations from cohorts of tumors in public repositories such as cBioPortal (https://www.cbioportal.org/), pediatric cBioPortal (https://pedcbioportal.org/login.jsp), ICGC [59], TARGET (https://ocg.cancer.gov/programs/target), St. Jude Cloud (https://www.stjude.cloud/) and additional cohorts directly obtained from literature studies summarized in Figure 2 of the paper. We also included the data from Pan Cancer studies such as TCGA [56-57], PCAWG [61] and the metastatic tumors from Hartwig Medical Foundation (https://www.hartwigmedicalfoundation.nl/en/). In fact, part of the novelty

of the framework is the amount of pediatric data and metastatic tumors included in the release, which is data frequently not well represented in this type of studies. All the data together comprises more than 28,076 tumors of 66 cancer types. I also provided help and feedback in the elaboration of a system for the cancer type annotations. The description of the new pipeline and the results of the analysis of all this great amount of data can be found published here [82] together with an historical view of the identification of cancer drivers genes. Also, all the results are uploaded in the IntOGen web (https://www.intogen.org/search).

The compendium of cancer driver genes which constitutes the main outcome of this work was published within the framework of a review article. Therefore, it should be considered as an analysis paper, since it reports an original research contribution.

Martínez-Jiménez F, Muiños F, Sentís I , Deu-Pons J , Reyes-Salazar I, Arnedo-Pac C , Mularoni L, Pich O , Bonet J , Kranas H , Gonzalez-Perez A, López-Bigas N. A Compendium of Mutational Cancer Driver Genes. Nature Reviews Cancer 20, 555–572 (2020). https://doi.org/10.1038/s41568-020-0290-x

A compendium of mutational cancer driver genes

Francisco Martínez-Jiménez[1], Ferran Muiños[1], Inês Sentís[1], Jordi Deu-Pons[1], Iker Reyes-Salazar[1], Claudia Arnedo-Pac[1], Loris Mularoni[1], Oriol Pich[1], Jose Bonet[1], Hanna Kranas[1], Abel Gonzalez-Perez[1,2]✉ and Nuria Lopez-Bigas[1,2,3]✉

Abstract | A fundamental goal in cancer research is to understand the mechanisms of cell transformation. This is key to developing more efficient cancer detection methods and therapeutic approaches. One milestone towards this objective is the identification of all the genes with mutations capable of driving tumours. Since the 1970s, the list of cancer genes has been growing steadily. Because cancer driver genes are under positive selection in tumorigenesis, their observed patterns of somatic mutations across tumours in a cohort deviate from those expected from neutral mutagenesis. These deviations, which constitute signals of positive selection, may be detected by carefully designed bioinformatics methods, which have become the state of the art in the identification of driver genes. A systematic approach combining several of these signals could lead to a compendium of mutational cancer genes. In this Review, we present the Integrative OncoGenomics (IntOGen) pipeline, an implementation of such an approach to obtain the compendium of mutational cancer drivers. Its application to somatic mutations of more than 28,000 tumours of 66 cancer types reveals 568 cancer genes and points towards their mechanisms of tumorigenesis. The application of this approach to the ever-growing datasets of somatic tumour mutations will support the continuous refinement of our knowledge of the genetic basis of cancer.

Cancer is a collection of diseases characterized by abnormal and uncontrolled cellular growth caused primarily by genetic mutations[1,2]. These mutations, called 'drivers' after their ability to drive tumorigenesis, confer on cells in a somatic tissue certain selective advantages with respect to neighbouring cells[1]. They occur in a set of genes (called 'cancer driver genes'), the mutant forms of which affect the homeostatic development of a set of key cellular functions. One of the main goals of cancer research, since the establishment of genetics, has been the discovery of these cancer driver genes across tumour types[3–6]. Their identification has led to the development of the paradigm of targeted anticancer therapies and, more generally, to the search for genomic biomarkers of prognosis and response to treatments[7].

The first part of this Review presents a historical perspective of the evolution of our knowledge of cancer genes from before the first whole-exome and whole-genome sequencing of tumours to the present day, and then provides an outlook for the future. It focuses on mutational driver genes, that is, those capable of driving tumorigenesis via single-nucleotide variants (SNVs) and short insertions or deletions (indels), which we collectively call 'point mutations'. However, it does not cover other types of somatic alterations that affect cancer genes and also contribute to tumorigenesis, such as amplifications or deletions, genomic rearrangements and epigenetic silencing. For comprehensive reviews on some of the other types of driver alterations not covered here, see, for example, REFS[8–10]. Also excluded are methods that identify driver genes on the basis of their proximity to significantly mutated genes in biochemical pathways or networks, which have also been reviewed elsewhere[11].

In the second part of this Review, we propose that the maturity of the methods for mutational driver identification and the wealth of tumour mutational datasets currently available in the public domain can advance the ultimate goal of uncovering the compendium of driver genes across all tumour types and also provide clues about their tumorigenic mechanisms. To demonstrate this proposition, we developed the Integrative OncoGenomics (IntOGen)[12,13] pipeline, aimed at the systematic identification of the compendium of mutational driver genes across tumour types. A snapshot of the compendium of driver genes described in this Review has been obtained through its application to 28,076 tumours grouped within 221 cohorts of 66 different tumour types. This snapshot of the compendium

[1]*Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain.*

[2]*Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain.*

[3]*Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.*

✉*e-mail: abel.gonzalez@irbbarcelona.org; nuria.lopez@irbbarcelona.org*

https://doi.org/10.1038/s41568-020-0290-x

Positional cloning
Technique for molecular cloning of all genetic material in a chromosomal locus with the aim of identifying genes.

Retrotransposition
Use of DNA retrotransposons to introduce pieces of foreign DNA into a genome with different research aims, such as transgenesis and insertional mutagenesis.

Sanger sequencing
Method of DNA sequencing developed by Sanger and colleagues in the 1970s which implements an in vitro DNA replication with selective incorporation of chain-terminating dideoxynucleotides.

Next-generation sequencing
(NGS). Also known as massively parallel sequencing, a group of high-throughput methods of DNA sequencing based on the concept of massively parallel processing.

of driver genes (and newer versions) and the automatic system to produce it are hosted on the IntOGen platform.

The genetic basis of cancer

The search for the causes of cancer is firmly intertwined with the development of genetics[14]. The first scientific notions on the causes of cancer derived from systematic record keeping in the eighteenth and nineteenth centuries, which linked the high incidence of specific types of tumours to exposures resulting from the practice of some professions[15,16]. The first known report on the heritability of cancer by Broca dates from the late 1800s, even before the genetic basis of inheritance developed by Mendel became widely recognized[17]. In the early 1900s, Peyton Rous was able to transmit tumours to healthy birds using cell-free extracts obtained from a diseased animal[18], thus suggesting that units smaller than cells were responsible for tumorigenesis. At approximately the same time, and before Morgan's work on chromosomes[19] as the seat of genes, Theodor Boveri proposed that cancer could arise as a result of incorrect chromosome combinations[20]. In addition, experiments with chemical carcinogens demonstrated that changes to the sequence of DNA promoted cellular transformation[21–24]. These and other findings brought the basis of cancer firmly within the realm of genetics.

The advancement of biochemistry and molecular genetics in the decades from 1940 to 1980 fostered the development of laboratory methods such as positional cloning, retrotranscription and Sanger sequencing. The application of these methods to cancer research led to the identification of the first cancer driver genes, named after the ability of their mutant forms to drive tumorigenesis. A small portion of the genomes of several birds that hybridized with part of the DNA of avian sarcoma virus was the first cancer gene to be identified and was thus named *SRC*[25]. The recognition of the existence of such viral DNA fragments, a variant of 'normal' genes present in the avian genomes, which had acquired transforming capability had already given rise to the term 'oncogene' in 1969 (REF. [26]). Oncogenes such as *HRAS* were then identified in human tumours[27,28], and the change of a single nucleotide in the gene sequence was demonstrated to be enough to provide the transforming capability[29,30]. With these discoveries, the genetic basis of tumorigenesis (including the aforementioned professional exposures) could finally be explained.

As the introduction of defective copies of the oncogene, despite the presence of normal alleles in the cell, was enough to produce transformation, it was concluded that oncogenes act in a dominant way[31]. However, the analysis of the incidence of retinoblastoma, a paediatric tumour, had shown that two hits, that is, genetic events inactivating both alleles of the gene (later named *RB1*, after the disease), are necessary for the development of the malignancy[32]. This apparent contradiction was solved by the mid-1980s with the acknowledgment of the existence of a second type of cancer gene, termed a 'tumour suppressor'[33]. Unlike in the case of oncogenes, transformation is caused by the inactivation of tumour suppressors, which in general requires loss of activity of both alleles of the gene. The discovery of tumour suppressors also provided an explanation for familial cancer cases[17]: an inherited mutation inactivating one of the alleles of a tumour suppressor increases the likelihood of developing a tumour as only the second hit is required.

Following this clear blueprint of two classes of cancer genes, between the 1980s and the early years of the first decade of the twenty-first century dozens of genomic loci encoding oncogenes, such as *MYC*, *RET*, platelet-derived growth factor receptor-α (*PDGFRA*), *MET*, *KIT*, FMS-like tyrosine kinase 3 (*FLT3*), epidermal growth factor receptor (*EGFR*) and *BRAF*[33–39], and tumour suppressors, such as *TP53*, transforming growth factor receptor-β2 (*TGFRB2*), *RB1*, *PTEN*, checkpoint kinase 2 (*CHEK2*), cyclin-dependent kinase inhibitor 2A (*CDKN2A*), *BRCA1*, *BRCA2* and adenomatous polyposis coli (*APC*)[40–52] were identified. Germline mutations in some of the latter were also shown to confer susceptibility to cancer development[40,48,50–51]. Further pioneering studies also established the importance of other types of alterations affecting these genes, such as amplifications, deletions, translocations or promoter hypermethylation, for cell transformation[33,50,54,55].

In 2004, a seminal article compiled a list of 291 cancer driver genes from the scientific literature[56], including genes altered through point mutations, translocations or copy-number changes. In an effort to conceptualize this heterogeneity, driver genes were recognized to affect primarily a handful of essential cellular functions, termed 'cancer hallmarks'[57] (reviewed and updated in 2011 (REF. [58])). According to this generalization, as a result of driver alterations, malignant cells become capable of (1) resisting apoptosis, (2) maintaining proliferative signalling (even in the absence of extracellular signals), (3) evading suppressors of cell growth, (4) initiating invasion and metastasis, (5) enabling replicative immortality, (6) inducing angiogenesis, (7) achieving deregulation of energy metabolism and (8) avoiding immune destruction. The development of these capabilities is supported by the promotion of tissue inflammation and the intrinsic genomic instability of tumours[58].

Somatic mutation patterns reveal drivers

In the early years of the first decade of this century, improvements introduced in DNA sequencing technologies and the rapid advance in the annotation of the human genome enabled projects aimed at revealing increasing shares of the landscape of somatic mutations in tumours. In 2005, a study sequencing 518 kinase-encoding genes found 76 non-silent mutations on average across 25 primary breast tumours and cell lines[59]. The following year, another group sequenced 13,023 genes of 11 breast tumours and 11 colorectal tumours and found 519 and 673 with mutations, respectively[60]. The development of next-generation sequencing (NGS) technologies in the middle of the first decade of this century[61] catalysed the beginning of cancer genomics. In 2008, two further analyses of 22 glioblastomas and 24 pancreatic tumours sequencing the entire exome found 1,007 and 685 mutated genes, respectively[62,63]. A similar landscape arose from the first whole-genome sequencing of tumours[64–67]. Nevertheless, the consensus

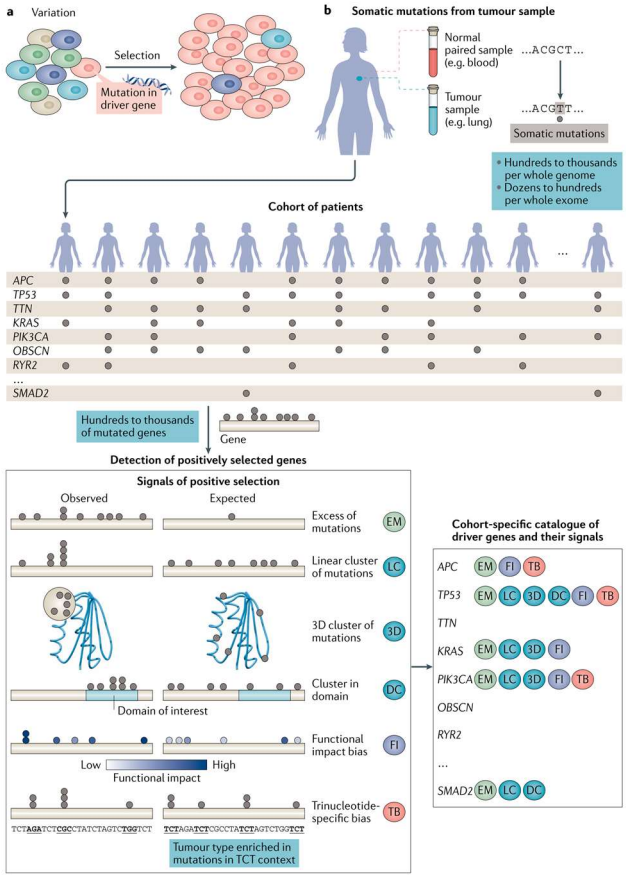


Fig. 1 | **Signals of positive selection identify driver genes. a** | Cells in somatic tissues accumulate mutations. Somatic mutations in certain genes provide the cell in which they occur with a selective advantage and are thus positively selected. Following a Darwinian process, over time, a clonal expansion occurs and the cells carrying mutations in these genes become dominant within the population. **b** | Deviations of the observed pattern of mutations of genes across samples of the same cancer type from the expected pattern reveal the genes under positive selection in tumorigenesis. Two biopsy samples are taken from a patient with cancer: one from the tumour and the other from a healthy tissue (for example, peripheral blood in solid malignancies). By comparison of the sequences of these two samples, the somatic point mutations in the tumour are identified. Between a handful and a few hundred somatic mutations are identified in the exome, a number that increases to tens of thousands if the whole genome is sequenced. As a result, between a few dozen and several thousand genes appear mutated in each tumour. The driver genes are those that exhibit one or more signals of positive selection across the tumours of a cohort.

Non-B-DNA structures
Local structures of chromosomal DNA that deviate (frequently in a transient manner) from the Watson–Crick double helix; they include stem–loop structures involving one or both DNA strands and G-quadruplexes.

viewpoint on tumorigenesis was that only a few mutational events affecting driver genes were expected to be the origin of malignization[31,68]. Therefore, the vast majority of these mutated genes would have no involvement at all in tumorigenesis; that is, their mutations are passengers rather than drivers. These studies first exposed the need for rigorous statistical tests that accounted for the heterogeneity of mutation rate and mutation types to identify the unexpected mutational patterns that reveal cancer genes[69–71].

These first studies paved the way for the launch of large tumour sequencing initiatives in several countries, such as The Cancer Genome Atlas (TCGA)[72], aimed at sequencing the exomes of hundreds of tumours of more than 24 frequent cancer types. As sequencing technologies continued to advance, more ambitious projects, many grouped under the umbrella of the International Cancer Genome Consortium (ICGC)[4], set their goal on sequencing the whole genome of thousands of tumour samples. With the recent conclusion of many of these initiatives, comprehensive pan-cancer analyses have laid out some of the most important findings of a little over a decade of cancer genomics research[73–76], including lists of identified driver genes[5,77]. The vast majority of these pioneering projects focused on the study of primary malignancies. It is only more recently that similar projects probing metastatic tumours have begun to reveal the landscape of driver alterations of advanced malignancies[78,79].

One of the main goals of all of these projects was the identification of the set of genes driving the malignancies, providing a road map for the systematic and comprehensive identification of mutational driver genes. The rationale behind it is that tumorigenesis follows a Darwinian evolution characterized by variation and selection[80,81]. Variation is provided by spontaneously arising somatic mutations that introduce genetic differences between somatic cells in a tissue. Positive selection then acts on cells carrying mutations that confer selective advantages over neighbouring cells, leading to clonal expansion of the mutants (FIG. 1a). (A variety of selective advantages, described above as the hallmarks of cancer, may be provided by mutations of different driver genes.)

As a result of this evolutionary process, when a cohort of tumours of the same cancer type is analysed, the deviation of patterns of mutations in some genes from their expectation under neutral mutagenesis may constitute signals that the mutations in those genes are under positive selection in tumorigenesis. For example, driver genes are mutated at abnormally high frequencies across the tumours of a cohort, and methods to detect this significant mutational recurrence were subsequently developed to analyse the mutational datasets produced by the aforementioned cancer genomics projects[68,82]. Other signals of positive selection in tumorigenesis (FIG. 1b), such as the abnormal clustering of mutations in certain regions of the proteins[83–88], a bias towards the accumulation of mutations with high functional impact[89] or a bias in the frequency of trinucleotide changes[90], have been used by driver identification methods[91,92]. Over time, many of these methods have been validated and tested on a number of cohorts of different cancer types and shown to be highly reliable. For thorough lists of methods see, for example, REFS[5,77,93,94].

The analysis of the first large mutational datasets revealed that different types of mutations appear with differing frequencies in tumours of different origin and that the rate of mutations across the human genome is highly heterogeneous (BOX 1). It quickly became apparent that driver detection methods are profoundly affected by the heterogeneity of the background mutation rate[95]. Building background models that accurately account for all of the factors that affect the mutation rate in the absence of selection has become a hallmark of most driver identification methods developed in recent years[90,95–101]. While several driver genes mutated at very high frequency may be spotted just by looking at their mutational pattern across tumours[68], the accurate modelling of the background mutation rate is key to avoiding the detection of false positive drivers and to identify those with lower mutation recurrence. The combination of the outputs of methods that use different signals of positive selection is the best approach for a comprehensive identification of driver genes, which may exhibit some but not all signals. Spurious discoveries by individual methods also have a higher chance of being filtered out by such a combination[5,13,94,102].

Box 1 | **The background mutation rate of genes**

The background mutation rate of a gene (that is, the rate and distribution of mutations) in a somatic cell is determined by its sequence, the identity of the cell and the mutational processes the cell or tissue as well as the person has been exposed to during their lifetime. A correct assessment of the background mutation rate of genes requires the ability to accurately model the variability introduced by all these factors. This is key to identifying which observed mutational patterns are actually unexpected and attributable to positive selection.

The mutational processes active in a tissue in an individual define a set of probabilities for each nucleotide in the gene to change taking into account its immediate sequence context[106,109,110,186,187]. These probabilities may be learned from the observed mutational profile of each tumour in a cohort or may be derived from the activity of a set of relevant mutational processes across the samples of a cohort[188].

The probability that a specific nucleotide change occurs in the gene is also influenced by the specific features adopted by the chromatin of the cell both at the large scale and at the small scales[95,189,190]. At the large scale, the time at which the gene is replicated relative to an origin[191], the level of compaction of the chromatin[192,193] at its locus and the level at which the gene is expressed[95] influence its mutation rate. The effect of these large-scale factors may be carefully modelled for each gene in each relevant tissue[95,97]. Alternatively, a background model within each gene may be built by permuting the mutations observed in the gene[98–100].

At the small scale, factors such as the occupancy by nucleosomes[194,195] and other proteins[196], the distribution of certain chromatin marks along the gene body[197,198] and the formation of local non-B-DNA structures[199–201] may alter the mutation rate locally at sequence stretches within the gene.

Systematic discovery of driver genes

The adoption of NGS by cancer research, fostered by pioneering initiatives such as the ones mentioned in the previous section, has generated a great amount of cancer genomics data available in the public domain. The tally of tumour samples sequenced at the whole-exome or whole-genome level that are currently available for systematic driver discovery is in the tens of thousands. This provides, in theory, the opportunity to identify the compendium of mutational driver genes ('compendium', for short); that is, the complete list of genes driving each malignancy upon mutations.

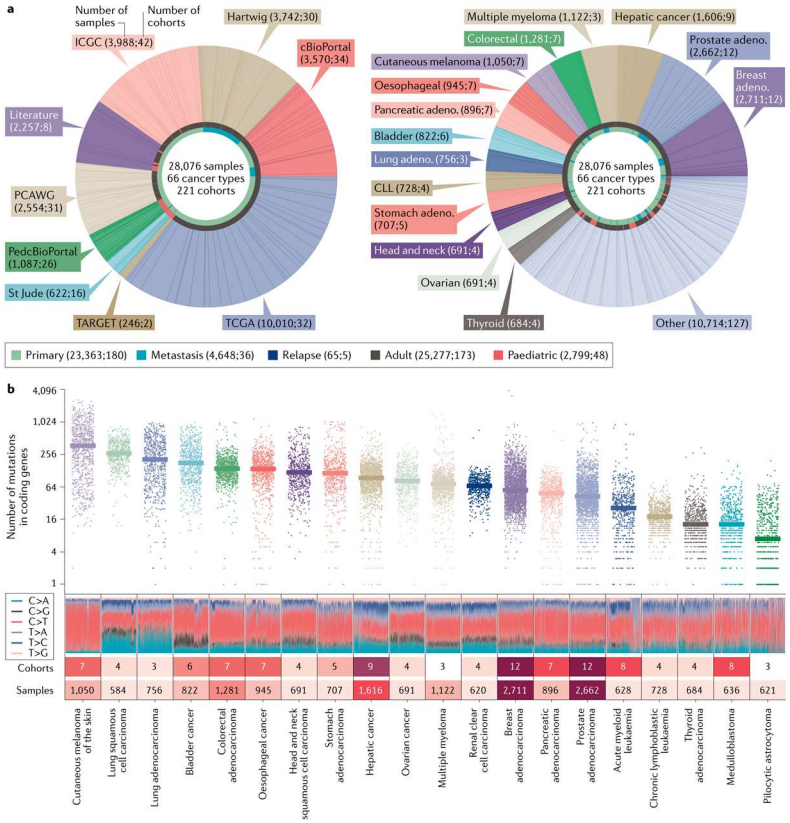


Fig. 2 | **Application of the IntOGen pipeline to datasets of tumour mutations. a** | Datasets of tumour mutations collected from the public domain for the construction of the current snapshot of the compendium of driver genes. Both doughnut plots represent all datasets classified by source (left) or cancer type (right). In both plots, the innermost ring signals the cohorts from primary or metastatic or relapse tumours, while the second ring highlights cohorts of adult or paediatric tumours. **b** | Mutation burden (top) and mutation type (bottom) of tumours from cancer types represented by at least two cohorts. The number of cohorts and samples contributing to the distribution of each cancer type are shown below the plot. Adeno., adenocarcinoma; CLL, chronic lymphocytic leukaemia; Hartwig, Hartwig Medical Foundation; ICGC, International Cancer Genome Consortium; PCAWG, Pan-Cancer Analysis of Whole Genomes; St Jude, St Jude Children's Research Hospital; TARGET, Therapeutically Applicable Research to Generate Effective Treatments; TCGA, The Cancer Genome Atlas.

Implementation of the system. To build a snapshot of this compendium, we have collected somatic SNVs and short indels across 221 cohorts (comprising between 10 and 973 samples) of 66 different cancer types totalling 28,076 samples (FIG. 2a; Supplementary Methods; Supplementary Table 1). We define a cohort as a set of tumour samples of the same cancer type analysed within a project with a uniform sequencing and

Box 2 | **Accessing the compendium of mutational driver genes**

The snapshot of the compendium of driver genes described in this Review as well as the automatic system used to produce it are hosted on the Integrative OncoGenomics (IntOGen) platform. Cancer researchers may explore the compendium, comprising the list of driver genes across tumour types and their mutational features, via the Web interface of the platform. All the information contained in it is also downloadable. Furthermore, the automatic system (the IntOGen pipeline) can be obtained by researchers from the platform for local installation and application to datasets of somatic mutations across cohorts of tumours. Details on the current implementation of the IntOGen pipeline can be found in Supplementary Methods. Building upon a practice that dates back to 2013, when the IntOGen platform for the analysis of cancer driver genes was first established[12,13], we will continue to collect tumour sequencing data as it becomes available in the public domain, and to produce more comprehensive snapshots of the compendium. For future versions of the pipeline and the compendium, regular updates may be found on the IntOGen website.

mutation calling pipeline. Most samples are contributed by large sequencing efforts, such as the ICGC[4,103] (3,988 samples), TCGA[72] (10,010 samples), Pan-Cancer Analysis of Whole Genomes (PCAWG)[76] (2,554 samples), Hartwig Medical Foundation[78] (3,742 samples) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET)[104] (246 samples). Importantly, the mutations across 60 other cohorts comprising 3,570 adult and 1,087 paediatric tumour samples sequenced by individual institutions were obtained via cBioPortal and PedcBioPortal[105], respectively. This highlights the importance of developing and maintaining centralized efforts to collect sequencing data produced within small projects. Finally, the mutations of 2,257 tumours sequenced as part of eight independent cohorts were obtained from the original studies. Most of the 221 cohorts (180) comprise primary tumours, while the remaining 41 are composed of metastatic or relapse samples (4,713 in total). Special effort has been made to include paediatric malignancies (2,799 samples grouped in 48 cohorts), which are traditionally under-represented in driver discovery efforts.

The number of coding mutations in tumours differs depending on the cancer type, and an important degree of variability across the samples of a given malignancy is also observed (FIG. 2b, top). For example, some breast adenocarcinomas bear mutations in several hundred genes, while other samples of the same malignancy exhibit only a dozen mutated genes. Part of this heterogeneity may be explained by differences in sequencing technology or depth, or in mutation calling methods. Nevertheless, most of the heterogeneity in mutation burden has a biological basis, owing to differences in the time or intensity of exposure to mutational processes, arising, for example, from the activity of ultraviolet light or faulty DNA repair[106–110]. While recalling all mutations across the cohorts would eliminate part of the variability of technical origin, this is not yet possible for such large numbers of samples owing to limitations in computational power. It is thus necessary, in the effort of systematic discovery of driver genes across cancer types, to analyse each cohort of tumours separately. Larger cohorts provide more statistical power to detect the signals of positive selection that characterize driver genes. Therefore, in this systematic discovery one expects that certain recurrently mutated driver genes will appear across many cohorts of the same malignancy, while others will be detected only in larger cohorts.

The construction of the compendium by use of these datasets of tumour mutations requires an efficient computational system that systematically runs state-of-the-art driver discovery methods. Our implementation of this system, which we refer to as the IntOGen pipeline[12,13] (BOX 2), consists of three basic steps, illustrated in FIG. 3 and explained at length in Supplementary Methods. A first preprocessing step guarantees that each method receives its input in the correct format and within operational parameters, for example, deduplicating samples taken from the same tumour, or removing those with an abnormal ratio of non-synonymous to synonymous mutations or with hypermutator phenotype. Seven recently published complementary methods of driver identification — dNdScv[97], OncodriveFML[98], CBaSE[101], OncodriveCLUSTL[99], a reimplementation of HotMAPS accounting for trinucleotide contexts of mutation types[87], smRegions[100] and Mutpanning[70] — are executed next. Then the lists of candidate drivers identified by each method are combined through a weighted vote in which the weight awarded to each method is based on its perceived credibility (Supplementary Fig. 1). The combination yields lists of driver genes per cohort that are more sensitive than those produced by individual methods without loss of specificity (Supplementary Fig. 2). In a final postprocessing step, spurious candidate driver genes that may appear owing to known confounders are automatically filtered out (Supplementary Methods). The IntOGen pipeline is designed to scale smoothly as the datasets of tumour mutations continue to grow into the hundreds of thousands, advancing our view of the compendium.

Each driver discovery method focuses on one or more features of the mutational pattern of genes across tumours. To identify the signals of positive selection, it assesses the deviation between the observed values and the expected values of the feature under the assumption of neutral mutagenesis (FIG. 5). These mutational features, collected by the IntOGen pipeline for all driver genes, provide key insights into the mechanisms of tumorigenesis for each of these cancer genes (see later), and are an integral part of the compendium (Supplementary Methods). They comprise (1) the clusters of mutations (both linear and 3D, which may arise owing to intraprotein or interprotein interactions), (2) domains in the protein that are preferentially affected by mutations and (3) the excess of mutations with different consequences.

Linear clusters are local accumulations of mutations along the sequence of a gene found across tumours, such as those formed by mutations at codons 12 and 13 of *KRAS* (FIG. 5). On the other hand, 3D clusters involve amino acid residues, which may be separated in the primary structure of the protein but are close in its tertiary structure (for example, mutations contributed by amino acids at positions 26, 39–42, 57 and 59–62 of RHOA). Preferentially affected domains bear a significant accumulation of mutations, such as the case of MH2 in SMAD4. The excess of mutations with different

Synonymous mutations
Single-nucleotide variants that cause a change of codon for a synonymous one.

Hypermutator phenotype
Tumours with abnormally high mutation burden in comparison with other samples of the same cohort (for example, more than three times the interquartile range above the median of the distribution), usually as a result of defective DNA repair mechanisms.

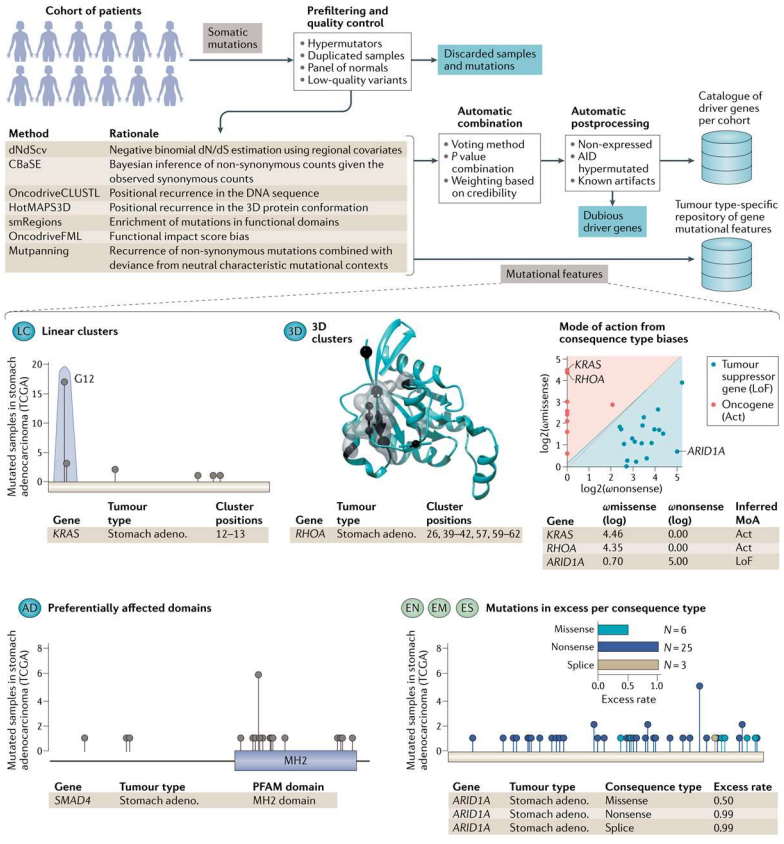
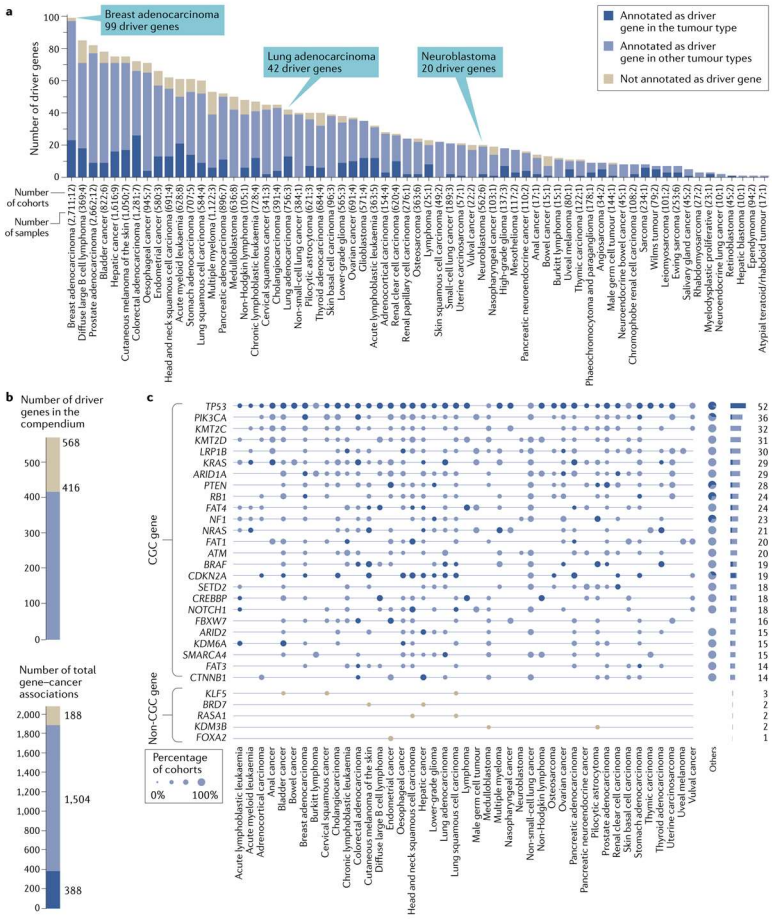


Fig. 3 | **Schematic representation of the Integrative OncoGenomics (IntOGen) pipeline.** The flow of data through the pipeline is illustrated starting from its application to a cohort of patients with stomach adenocarcinomas. The two outcomes of the pipeline — that is, the catalogue of driver genes in the cohort and the mutational features (linear and 3D clusters of mutations, mode of action (MoA), preferentially affected domains and excess of mutations with different consequences) — for each patient in the cohort are integrated to form the compendium of driver genes. Act, activating; adeno., adenocarcinoma; AID, activation-induced cytidine deaminase; LoF, loss of function; TCGA, The Cancer Genome Atlas.

Nonsense mutations
Single-nucleotide variants that cause the change of a stop codon for an amino acid-coding codon.

Missense mutations
Single-nucleotide variants that cause the change of an amino acid in a protein sequence.

consequences — 99% and 50% of nonsense mutations and missense mutations, respectively, for AT-rich interactive domain 1A (*ARID1A*) — informs about the mode of action (tumour suppressor or oncogene) of a driver gene. An excess of observed missense mutations in the absence of an excess of nonsense mutations indicates the activating mode of action of oncogenes. By contrast, tumour suppressor (or loss-of-function) genes tend to exhibit an excess of nonsense mutations. While the mode of action of some genes is very clear-cut, some



cases are harder to place within the binary oncogene–tumour suppressor model (close to the diagonal in the 'mode of action' scatterplot in FIG. 3). Furthermore, the mode of action of some genes may differ between tumour types.

A snapshot of the compendium

How much does the systematic compendium, or more appropriately, the current snapshot obtained from these 221 cohorts of tumours (BOX 2) add to the current knowledge of the genetic basis of tumorigenesis?

◀ Fig. 4 | **A snapshot of the compendium of mutational driver genes. a** | Number of cancer driver genes per tumour type in the compendium. The three-colour scale to denote genes annotated in the Cancer Gene Census (CGC) in the same tumour type as that identified in the compendium or in a different tumour type or to denote genes not annotated in the CGC is used throughout the figure. **b** | Total number of cancer driver genes in the compendium, indicating the overlap with the genes annotated in the CGC as drivers in any tumour type (top bar). Overlap between driver gene–tumour type associations in the compendium and those in the CGC in the same or a different tumour type (bottom bar). **c** | The range of tumour types in which 25 exemplary genes are identified as drivers by the compendium represented as dots in a matrix compared with the associations annotated in the CGC. The bottom of the plot presents the involvement in tumorigenesis of five previously unannotated drivers across tumour types. The size of the dots represents the percentage of all cohorts of the tumour type in which the gene is identified as a driver. The number of tumour types in which each gene appears as a driver in the compendium is represented in the bars to the right.

A systematic mining of the literature to establish a thorough and reliable catalogue of validated cancer genes is beyond the scope of our analysis. Thus, to address this question, we used the set of driver genes in the Cancer Gene Census[111] (CGC; version 87) as the 'ground truth' of the genes involved in the development of the 66 malignancies represented in the compendium. While the CGC is incomplete and may contain some false positives, it is, to our knowledge, the most comprehensive and accurate set of validated cancer genes annotated from the literature, and it thus serves this purpose. One part of the answer (FIG. 4a,b) then is that almost three quarters of the 568 mutational driver genes in the compendium are already annotated in the CGC (which also provides a strong validation of the compendium). However, because the compendium identifies the signals of positive selection unbiasedly across the cohorts of all cancer types, it has the possibility of more thoroughly mapping driver gene–tumour type associations. Indeed, more than 80% of all identified links between a driver gene and a malignancy are not annotated in the CGC (FIG. 4a,b). For example, while 21 known CGC drivers of breast adenocarcinomas are in the compendium, 75 genes annotated in the CGC but not previously recognized to drive this malignancy are shown to be under positive selection across one or more of the 12 breast cancer cohorts analysed (FIG. 4a). In other words, for many well-known driver genes, the compendium reveals that their role across cancer types is much more widespread than previously documented (FIG. 4c). For example, the pattern of somatic mutations in histone-lysine *N*-methyltransferase 2C (*KMT2C*) shows signals of positive selection across 31 tumour types. However, it is annotated in the CGC

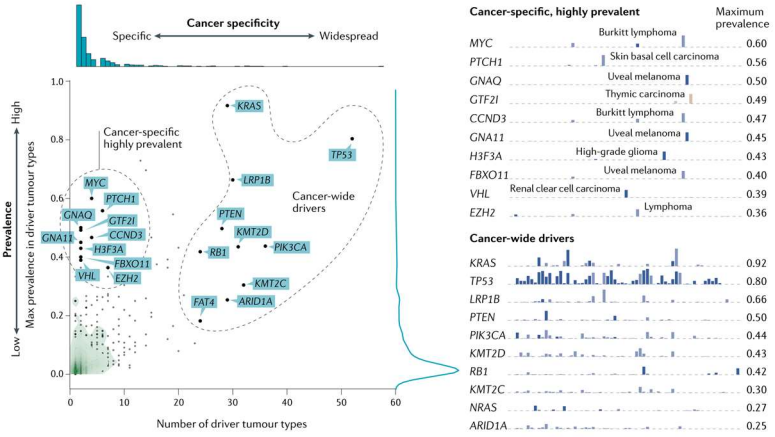
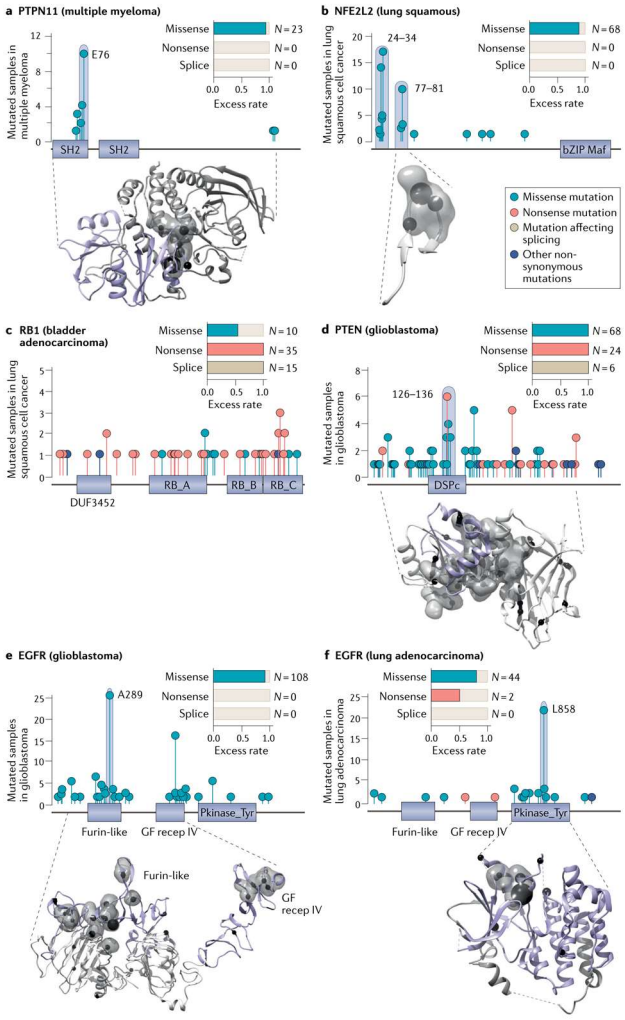


Fig. 5 | **Distribution of the prevalence of driver genes across cancer types in the compendium.** Each driver gene is represented as a single dot in the scatter plot. The horizontal axis represents the number of tumour types where a gene has been identified as a driver, and the vertical axis represents the maximum mutational frequency of the gene across the tumour types. The separate distributions of these two variables are represented through 1D histograms above and to the right of the graph. Two sets of drivers mutated at high frequency either across one or very few tumour types (cancer specific, highly prevalent) or across more than 20 cancer types (cancer-wide drivers) are circled and denoted by their abbreviations. While most cancer-wide drivers are bona fide well-recognized cancer genes, low-density lipoprotein receptor-related 1B (*LRP1B*) has long been assumed to be a potential spurious finding. This debate is not yet settled, as some studies have found its loss of function may be related to enhanced cell migration in several tissues[181–183]. The bar plots to the right of the graph present the mutational frequency across tumour types (corresponding to the *x* axis in the scatter plot) of selected cancer-specific, highly prevalent and cancer-wide drivers. The maximum mutational frequency of each of them appears beside the corresponding row. Bars are coloured following the legend in FIGURE 4.


a PTPN11 (multiple myeloma)
Missense N = 23
Nonsense N = 0
Splice N = 0
Excess rate
Mutated samples in multiple myeloma
E76
SH2
SH2
b NFE2L2 (lung squamous)
Missense N = 68
Nonsense N = 0
Splice N = 0
Excess rate
Mutated samples in lung squamous cell cancer
24–34
77–81
bZIP Maf
Missense mutation
Nonsense mutation
Mutation affecting splicing
Other non-synonymous mutations
c RB1 (bladder adenocarcinoma)
Missense N = 10
Nonsense N = 35
Splice N = 15
Excess rate
Mutated samples in lung squamous cell cancer
DUF3452
RB_A
RB_B
RB_C
d PTEN (glioblastoma)
Missense N = 68
Nonsense N = 24
Splice N = 6
Excess rate
Mutated samples in glioblastoma
126–136
DSPc
e EGFR (glioblastoma)
Missense N = 108
Nonsense N = 0
Splice N = 0
Excess rate
Mutated samples in glioblastoma
A289
Furin-like
GF recep IV
Pkinase_Tyr
Furin-like
GF recep IV
f EGFR (lung adenocarcinoma)
Missense N = 44
Nonsense N = 2
Splice N = 0
Excess rate
Mutated samples in lung adenocarcinoma
L858
Furin-like
GF recep IV
Pkinase_Tyr

◀ Fig. 6 | **Interpreting the mutational patterns of driver genes. a–f |** Six exemplary mutational patterns computed for five proteins across five cohorts, including multiple myelomas (obtained from a study published in 2018 (REF.[130])) and lung squamous cell carcinomas, bladder adenocarcinomas, glioblastomas and lung adenocarcinomas obtained from The Cancer Genome Atlas (TCGA). Clusters and their boundaries are defined by methods that assess the significant clustering of mutations. In all plots, *N* denotes the number of mutations of each type of consequence (that is, missense mutations, nonsense mutations or mutations affecting splicing) observed in the gene across the cohort. bZIP Maf, bZIP Maf transcription factor domain; DSPc, dual specificity phosphatase, catalytic domain; DUF3452; domain of unknown function 3452; EGFR, epidermal growth factor receptor; Furin-like, furin-like cysteine-rich region; GF recep IV, growth factor receptor domain IV; NFE2L2, nuclear factor erythroid 2-related factor 2; Pkinase_Tyr, protein tyrosine and serine/threonine kinase; PTPN11, protein tyrosine phosphatase non-receptor type 11; RB_A, retinoblastoma-associated protein A domain; RB_B, retinoblastoma-associated protein B domain; Rb_C, Rb C-terminal domain; SH2, Src homology 2 domain.

only as a driver of medulloblastomas. The unbiased discovery of cancer genes through the IntOGen pipeline is thus an essential complement to the annotation of experimentally validated drivers.

Not only does the systematic nature of the compendium add to our knowledge of the role of well-known cancer genes but it also points at 152 potential new driver genes (FIG. 4a,c); that is, genes that were not previously annotated in the CGC. As the CGC is most likely an incomplete proxy of the full catalogue of cancer genes, some of these potential new drivers may have been reported before in the literature. Indeed, we present and discuss below five of these unannotated genes that exhibit signals of positive selection in their mutational pattern across tumours and have been suggested by independent studies to be involved in tumorigenesis (FIG. 4c, bottom).

The pattern of mutations in RAS GTPase-activating 1 (*RASA1*) across lung and head and neck squamous cell carcinomas exhibits several signals of positive selection probed in the system. Its decreased expression or loss-of-function mutations have been recognized to increase RAS-mediated signalling in human bronchial epithelial[112] and melanoma[113] cell lines. It has also been linked to tumorigenic promoting functions in triple-negative breast cancer[114]. Because the protein encoding *RASA1*, like the protein encoding neurofibromin 1 (*NF1*), negatively regulates the RAS–MAPK pathway[115], both genes are thought to function as tumour suppressors, which is also suggested by their mutational patterns. Lysine-specific demethylase 3B (*KDM3B*), whose protein product specifically demethylates Lys9 of histone H3 to promote the transcriptional activation of target genes, exhibits significant excess of mutations and functional bias across two cohorts of pilocytic astrocytomas and medulloblastomas. However, neither nonsense nor missense mutations are clearly over-represented within this excess; thus, its mode of action is currently labelled as 'ambiguous' in the compendium. KDM3B has been shown to be involved in cell cycle regulation in hepatocellular carcinomas[116] and to function as an activator of the WNT signalling pathway in colorectal cancer stem cells[117]. Although these two studies suggest that *KDM3B* acts as an oncogene in tumorigenesis, a separate report proposes that some of its germline mutations cause susceptibility to Wilms tumours[118]. Thus, its exact mode of action in tumorigenesis remains to be determined. Several genes encoding forkhead box transcription factors are annotated in the CGC as drivers of several malignancies (for example, forkhead box A1 (*FOXA1*) of breast and prostate carcinomas and *FOXR1* of neuroblastomas). Nevertheless, *FOXA2*, with several signals of positive selection across uterine carcinomas, is not annotated in the CGC. *FOXA2* mutations frequently found in uterine carcinomas tend to affect the DNA-binding domain or cause truncation of the protein product[119], causing its failure to localize to the nucleus[120]. Some of these mutant forms are known to cause a decrease in expression of the *CDH1* gene (which encodes E-cadherin) and thus have been associated with epithelial-to-mesenchymal transition in the progression of certain tumours[121,122]. Krueppel-like factor 5 (*KLF5*), which encodes a transcription factor involved in the regulation of human development and identified as a cancer driver gene, altered through different mechanisms[123,124] exhibits signals of positive selection across cervical squamous, bladder and lung squamous cell carcinomas. We also identified bromodomain-containing 7 (*BRD7*), which has several paralogues already annotated in the CGC and has been postulated to act as a co-activator of the SMAD transcription factors[125] in driving the initiation of melanomas and liver carcinomas.

Some genes act as drivers across several cancer types, while others tend to be more specific. The compendium provides an opportunity to assess the specificity of driver genes across tumour types in a systematic manner (FIG. 5). Most genes (360) act as drivers in one or two tumour types, and only a small group of ten genes (cancer-wide drivers, bottom right panel) are able to drive more than 20 malignancies through mutations. Some very specific mutational drivers (upper left outliers in FIG. 5 and top right panel) are very frequently mutated in only one or two cancer types. For example, 60% of all Burkitt lymphomas bear driver mutations in *MYC*[126] and 47% bear driver mutations in cyclin D3 (*CCND3*)[127]. Half of the cases of uveal melanoma bear activating mutations in one of two hotspots of guanine nucleotide-binding protein G_q subunit-α (*GNAQ*), while almost all other cases bear mutations at one of two homologous hotspots of its paralogue *GNA11* (REF.[128]). Interestingly, general transcription factor II-I (*GTF2I*), mutations of which drive virtually half of all thymomas[129], is not yet annotated in the CGC.

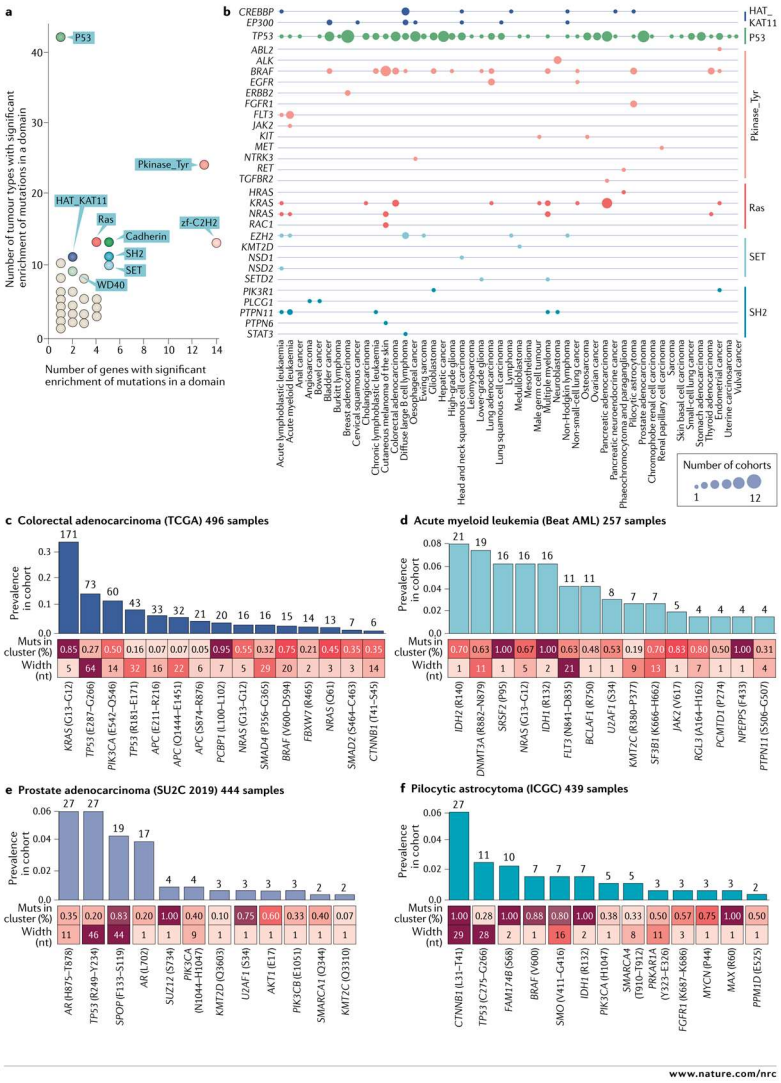
Mutational features of driver genes

We propose that the mutational features (exemplified in FIG. 3) of a driver gene provide a unique opportunity to shed light on its tumorigenic function. Below, we describe the mutational features of six driver genes as examples of the information they provide on their role in cell transformation.

The oncogene protein tyrosine phosphatase non-receptor type 11 (*PTPN11*) shows excessive missense mutations across multiple myelomas[130] (FIG. 6a) and other tumour types[131,132], which significantly

Wilms tumours
A rare type of kidney cancer that affects mostly children.

Paralogues
Genes within the same genome that have evolved from a common ancestor.


a
Number of tumour types with significant enrichment of mutations in a domain
Number of genes with significant enrichment of mutations in a domain
P53
Pkinase_Tyr
HAT_KAT11
Ras
Cadherin
zf-C2H2
SH2
SET
WD40
b
CREBBP
EP300
TP53
ABL2
ALK
BRAF
EGFR
ERBB2
FGFR1
FLT3
JAK2
KIT
MET
NTRK3
RET
TGFBR2
HRAS
KRAS
NRAS
RAC1
EZH2
KMT2D
NSD1
NSD2
SETD2
PIK3R1
PLCG1
PTPN11
PTPN6
STAT3
HAT_KAT11
P53
Pkinase_Tyr
Ras
SET
SH2
Acute lymphoblastic leukaemia
Acute myeloid leukaemia
Anal cancer
Adrenocortical
Bowel cancer
Bladder cancer
Brain lower grade glioma
Breast adenocarcinoma
Cervical squamous cancer
Cholangiocarcinoma
Chronic lymphoblastic leukaemia
Cutaneous melanoma of the skin
Ocular melanoma
Diffuse large B cell lymphoma
Oesophageal cancer
Glioblastoma
Glioma
Hepatic cancer
High-risk neuroblastoma
Head and neck squamous cell carcinoma
Leiomyosarcoma
Liposarcoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Lymphoma
Medulloblastoma
Mesothelioma
Male germ cell tumour
Multiple myeloma
Neuroblastoma
Non-Hodgkin lymphoma
Non-small cell lung cancer
Osteosarcoma
Ovarian cancer
Pancreatic adenocarcinoma
Pancreatic neuroendocrine cancer
Phaeochromocytoma and paraganglioma
Pilocytic astrocytoma
Prostate adenocarcinoma
Chromophobe renal cell carcinoma
Renal papillary cell carcinoma
Sarcoma
Skin basal cell carcinoma
Small cell lung cancer
Stomach adenocarcinoma
Thyroid adenocarcinoma
Endometrial cancer
Uterine carcinosarcoma
Vulval cancer
Number of cohorts
1
12
c Colorectal adenocarcinoma (TCGA) 496 samples
Prevalence in cohort
Muts in cluster (%)
Width (nt)
171 73 60 43 33 32 21 20 16 16 15 14 13 7 6
0.85 0.27 0.50 0.16 0.07 0.07 0.05 0.95 0.55 0.32 0.75 0.21 0.45 0.35 0.35
5 64 14 32 16 22 6 7 5 29 20 2 3 3 14
KRAS (G13–G12)
TP53 (E287–G266)
PIK3CA (E542–Q546)
TP53 (R181–E171)
APC (E211–R216)
APC (Q1444–E1451)
APC (S874–R876)
PCBP1 (L100–L102)
NRAS (G13–G12)
SMAD4 (P356–G365)
BRAF (V600–D594)
FBXW7 (R465)
NRAS (Q61)
SMAD2 (S464–C463)
CTNNB1 (T41–S45)
d Acute myeloid leukemia (Beat AML) 257 samples
Prevalence in cohort
Muts in cluster (%)
Width (nt)
21 19 16 16 16 11 11 8 7 7 5 4 4 4 4
0.70 0.63 1.00 0.67 1.00 0.63 0.48 0.53 0.19 0.70 0.83 0.80 0.50 1.00 0.31
1 11 1 5 2 21 1 1 9 13 1 7 1 1 4
IDH2 (R140)
DNMT3A (R882–N879)
SRSF2 (P95)
NRAS (G13–G12)
IDH1 (R132)
FLT3 (N841–D835)
BCLAF1 (R750)
U2AF1 (S34)
KMT2C (R380–P377)
SF3B1 (K666–H662)
JAK2 (V617)
RGL3 (A164–H162)
PCMT1D (P274)
NPEPPS (F433)
PTPN11 (S506–G507)
e Prostate adenocarcinoma (SU2C 2019) 444 samples
Prevalence in cohort
Muts in cluster (%)
Width (nt)
27 27 19 17 4 4 3 3 3 3 2 2
0.35 0.20 0.83 0.20 1.00 0.40 0.10 0.75 0.60 0.33 0.40 0.07
11 46 44 1 1 9 1 1 1 1 1 1
AR (H875–T878)
TP53 (R249–Y234)
SPOP (F133–S119)
AR (L702)
SUZ12 (S734)
PIK3CA (N1044–H1047)
KMT2D (Q3603)
U2AF1 (S34)
AKT1 (E17)
PIK3CB (E1051)
SMARCA1 (Q344)
KMT2C (Q3310)
f Pilocytic astrocytoma (ICGC) 439 samples
Prevalence in cohort
Muts in cluster (%)
Width (nt)
27 11 10 7 7 7 5 5 3 3 3 3 2
1.00 0.28 1.00 0.88 0.80 1.00 0.38 0.33 0.50 0.57 0.75 1.00 0.50
29 28 2 1 16 2 1 8 11 3 1 1 1
CTNNB1 (L31–T41)
TP53 (C275–G266)
FAM174B (S68)
BRAF (V600)
SMO (V411–C416)
IDH1 (R132)
PIK3CA (H1047)
SMARCA4 (T910–T912)
PRKAR1A (Y323–E326)
FGFR1 (K687–K686)
MYCN (P44)
MAX (R60)
PPM1D (E525)

◀ Fig. 7 | **Recurrent cancer driver domains and mutational clusters. a** | Dots represent all domains with significant enrichment of mutations in a number of different driver genes across a number of different tumour types. Selected domains with very significant enrichment are coloured and denoted with the domain acronym, while the rest appear in light grey. **b** | Genes with significant enrichment of mutations in domains of their protein products coloured in part **a** across tumour types. **c**–**f** | Number of mutations and prevalence of linear mutational clusters identified in several drivers across the cohorts of colorectal adenocarcinomas, obtained from The Cancer Genome Atlas (TCGA) (part **c**), acute myeloid leukaemias (obtained from the Beat AML project[184]) (part **d**), prostate adenocarcinomas (obtained from a Stand Up To Cancer (SU2C) 2019 publication[185]) (part **e**) and pilocytic astrocytomas (obtained from the International Cancer Genome Consortium (ICGC)) (part **f**). The fraction of mutations of each protein in each cohort that appear in clusters and the width of these clusters in the gene sequence appear in the heatmaps below each graph. The numbers at the top of each column represent the number of samples with mutations falling in each cluster. Cadherin, cadherin domain; HAT_KAT11, histone acetylation protein; muts, mutations, nt, nucleotides; P53, p53 DNA-binding domain; Pkinase_Tyr, protein tyrosine and serine/threonine kinase domain; Ras, RAS family domain; SET, SET domain; SH2, Src homology 2; WD40, WD domain Gβ repeat; zf-C2H2, C2H2 zinc-finger.

cluster within the SH2 domain of its protein product. Inhibitory contacts between this domain and the phosphatase domain are abrogated on phosphorylation by a receptor tyrosine kinase in the wild type or by mutations in the domain[133]. The activated PTPN11 then dephosphorylates inhibitors of several signalling pathways, such as the MAPK or AKT pathways[134]. Nuclear factor erythroid 2-related factor 2 (*NFE2L2*), another classic oncogene, encodes a transcription factor that is key in the control of the redox state of the cell and its response to stress[135–137]. Across lung squamous cell carcinomas[138], two narrow clusters of missense mutations appear at its N-terminal portion (FIG. 6b). These mutations affect sequences recognized by the cognate E3-ubiquitin ligase Kelch-like ECH-associated protein 1 (KEAP1) (that is, degrons), and cause the abnormal stabilization of NFE2L2, as do *KEAP1* mutations affecting the domain that recognizes the NFE2L2 degrons. This, in turn, results in the constitutive activation of NFE2L2-regulated genes[100].

The mutational features are radically different for tumour suppressors such as *RB1* across bladder adenocarcinomas[139] (FIG. 6c), with greater excess of nonsense mutations and mutations affecting splicing than of missense mutations. Most nonsense mutations trigger nonsense-mediated decay of the *RB1* mRNA[140], thus causing depletion of the protein and abrogating its functions in the regulation of cell cycle progression and the cell division cycle, the response to cellular stress, differentiation, cellular senescence, programmed cell death and maintenance of chromatin structure[141–143]. *PTEN*, another tumour suppressor, shows an excess of both nonsense and missense mutations across glioblastomas[72,144] (FIG. 6d). Like nonsense mutations in *RB1*, nonsense mutations in *PTEN* trigger nonsense-mediated decay, reducing the production of a functional *PTEN* protein product, while missense mutations hinder either its enzymatic activity or its recruitment to the membrane, or increase its susceptibility to ubiquitylation for proteasome-mediated degradation[145,146]. These outcomes, in turn, interfere with its role in the regulation of a host of cellular functions, such as cell cycle progression, apoptosis and protein synthesis[147–149].

Different tumorigenic mechanisms of the same driver across tumour types may also be revealed by their mutational features. For example, in glioblastomas[72], missense mutations of *EGFR* (an oncogene whose protein product is involved in the activation of multiple signalling pathways) tend to cluster in the extracellular domains of its protein product (FIG. 6e). These act as gain-of-function alterations, likely through the stabilization of the open conformation of the receptor, which stimulates its autophosphorylation in the absence of a ligand[150,151]. By contrast, across lung adenocarcinomas[152], missense mutations tend to cluster in the tyrosine kinase domain of the protein product of *EGFR* (FIG. 6f), altering its 'on–off' equilibrium and increasing its activity at the expense of reduced affinity for ATP[153,154].

Overall, several domains across the protein products of multiple genes appear to be preferentially affected by mutations across more than ten different tumour types (FIG. 7a,b). The p53 DNA-binding domain (P53 in FIG. 7a,b) appears significantly enriched for somatic mutations across cohorts of 42 different cancer types, a greater number than any other protein domain, although this is driven only by *TP53*. In another example, the tyrosine kinase domain of 13 different genes is significantly enriched for mutations across cohorts of 24 tumour types. Of these 13 genes, *BRAF* is the one exhibiting a significant enrichment of somatic mutations within the tyrosine kinase domain across most tumour types (14). The RAS, cadherin and C2H2 zinc-finger domains each exhibit significant enrichment of mutations across 13 cancer types.

An overview of significant clusters reveals that those in tumour suppressors tend to be wider, while those in oncogenes are narrow and tend to accumulate a larger share of the mutations observed in the gene (FIGS 7c–f, 8). Particularly narrow clusters are observed, for example in *KRAS* (five nucleotides overlapping codons 12 and 13 of the protein) accumulating 85% of the mutations in the gene across a cohort of 496 colorectal adenocarcinomas (FIG. 7c), or in isocitrate dehydrogenase 1 (*IDH1*) with all mutations in a cohort of 257 acute myeloid leukaemias affecting two nucleotides of codon 132 (FIG. 7d). Wider clusters accumulate 83% of the mutations of speckle-type POZ protein (*SPOP*) (44 nucleotides between codons 119 and 133) across a cohort of 444 prostate adenocarcinomas (FIG. 7e) or 28% of mutations of *TP53* (28 nucleotides between codons 266 and 275) across a cohort of 439 pilocytic astrocytomas (FIG. 7f). The width of clusters and the fraction of mutations of the gene that fall within them differ depending on the mode of action of cancer genes in tumorigenesis (FIG. 8). The relatively narrow clusters of oncogenes reflect the existence of relatively few available gain-of-function mutations along their sequence. This is also the reason why these clusters tend to concentrate large shares of all the mutations observed in oncogenes across a cohort of tumours. Wider clusters in tumour suppressor genes are observed because as a rule more loss-of-function mutations are available in their sequence (for example, mutations affecting several amino acids of a functionally important domain).

Degrons
Short sequences (4–10 amino acids) within a protein that are specifically recognized and bound by enzymes responsible for the conjugation of ubiquitin moieties.

Nonsense-mediated decay
Surveillance mechanism charged with the elimination of mRNA transcripts with premature stop codons.

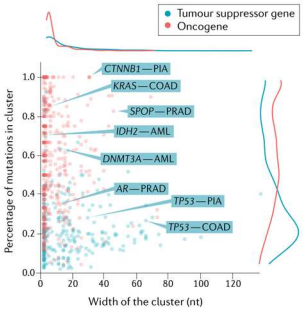


Fig. 8 | **Linear clusters detected in tumour suppressors and oncogenes in the compendium.** Every cluster detected in genes of the compendium in a particular tumour type is represented as a dot. Oncogenes are depicted in red and tumour suppressors are depicted in blue. The separate distribution of the two variables in the plot is represented through 1D histograms above and to the right of the graph. The intensity of the colour of each dot reflects the number of dots in the same position. AML, acute myeloid leukaemia; COAD, colon adenocarcinoma; nt, nucleotides; PIA, pilocytic astrocytoma; PRAD, prostate adenocarcinoma.

Conclusions and perspectives

Much like ancient manuscripts, in which newer layers of writing have been superimposed onto older texts, or cities with a long history of human dwelling, such as Rome, in which certain edifices exhibit rows of brick and mortar dating from different ages, the somatic mutations in tumour genomes constitute a record of their history. Therefore, borrowing the name given to these ancient scripts, somatic mutations in a tumour may be considered a palimpsest[106], the study of which may provide extremely useful information about the tumour and its environment. These palimpsests contain the footprints of all the mutational processes to which somatic cells in the tumour have been exposed during the life of the patient, as well as the signals of positive selection reminiscent of successive selective sweeps caused by driver mutations. Cleverly designed bioinformatics analyses applied to tumour genomes are able to reveal such footprints and traces. This Review has shown that the systematic application of such bioinformatics analyses to the detection of positive selection from the palimpsest of tumour somatic mutations is able to begin to reveal the compendium of cancer driver genes.

Before the inception of cancer genomics, a few dozen cancer driver genes were identified (FIG. 9). In the span of two or three decades, these genes were intensively studied and functionally characterized through an array of biochemical assays and the laborious dedication of several research groups. By contrast, in less than the two decades that have elapsed since the sequencing of the first tumour genomes, several hundred more cancer genes have been identified. This 'era' of cancer genomics has been made possible by advances in DNA sequencing and the development of bioinformatics methods to handle the challenges that analysis of genomics data poses. As shown herein, the compendium of mutational driver genes derived from the analysis of the cancer exomes currently in the public domain (~28,000) comprises between 500 and 600 mutational drivers. The completion of the compendium will constitute a milestone on the road to our understanding of tumour biology. To date, it is very likely that genes mutated at frequencies above 10% have already been discovered[96], and systematic analyses, such as those made possible with the IntOGen platform, reveal their involvement in tumorigenesis across cancer types.

We are also now in a position to project the evolution of the compendium into the future. The number of datasets of tumour somatic mutations deposited in the public domain is foreseen to increase quickly as initiatives to share data generated internationally, such as the Global Alliance for Genomics and Health and the 1+ Million Genomes initiative[155], come to fruition. As new snapshots of the compendium are uncovered with use of these data, the trend described above is predicted to continue into the future, with the identification of (1) new drivers mutated at frequencies below 10% across malignancies (owing to increased statistical power[96]), (2) drivers of conditions not profiled before, (3) drivers in diverse populations or ethnicities that have so far been biased against in tumour genome sequencing projects and (4) drivers of new clinical entities, such as metastatic or relapse tumours, which have been comparatively underexplored to date. For instance, a search through the current snapshot of the compendium shows that oestrogen receptor (*ESR1*) and androgen receptor (*AR*), while rarely mutated across primary breast and prostate tumours, respectively, are clear mutational drivers of resistance to treatment.

In this Review, we have purposefully focused on driver mutations affecting protein-coding genes. As mentioned in the Introduction, this excludes other types of somatic alterations affecting driver genes. While short indels are included within point mutations for the purpose of revealing mutational driver genes, their probability of occurrence likely involve features beyond their immediate sequence context, and thus their background rate is more difficult to model[109,110,156]. It also excludes the potential role in tumorigenesis of mutations affecting non-coding genomic elements, of which recent studies have identified few in comparison with coding genes[77,102]. Focusing on known cancer genes and their *cis*-regulatory regions, one of these surveys revealed that non-coding driver mutations are much less frequent than protein-coding ones, with the exception of mutations in telomerase reverse transcriptase (*TERT*), even after correcting for differences in statistical power between whole-genome and whole-exome sequencing datasets[77]. Nevertheless, it has also become apparent from whole-genome-sequenced tumours that our current knowledge of the distribution of mutations in non-coding regions is not comprehensive enough to allow the correct modelling of their background

Cis-regulatory regions
DNA sequences involved in the regulation of the expression of genes, such as transcription factor binding sites that may be found in promoters or enhancers.

Clonal haematopoiesis
Ageing-related clonal expansion of specific haematopoietic stem cells (HSCs) or other early blood cell progenitors which contributes to the appearance of genetically distinct subpopulations of blood cells.

mutation rate. Furthermore, our knowledge of the biological function of most of the non-coding areas of the genome still lags far behind that of coding genes[76]. Solving these issues will be key to fully exploring the catalogue of driver non-coding genomic elements. Furthermore, a holistic compendium of all types of driver alterations (coding and non-coding somatic point mutations, structural variants, epigenetic silencing events and germline susceptibility variants) is needed to uncover their panorama across tumours (reported in a preprint article[102]).

A detailed description of the precise involvement of each gene in tumour development is absent from the current snapshot of the compendium of driver genes. Thus, understanding the precise mode of alteration of each driver gene (that is, which of its mutations has the potential to drive tumorigenesis and why) and the specific biological function it perturbs in tumorigenesis is one of the major challenges of cancer genomics in the near future.

A first challenge is to precisely identify the mechanisms that alter the function of driver genes making them capable of driving tumorigenesis. This is the same as identifying all of the mutations of cancer driver genes that are capable of driving the malignancy and understanding their role in cell transformation[7,97,102]. As explained already, we propose that the mutational features computed within the compendium may aid in this endeavour. Furthermore, while the perturbation of several key biological processes (the hallmarks of cancer detailed above) are required for tumorigenesis, the specific process — for example, evading apoptosis, maintaining proliferative signalling and escaping the immune system — affected by mutations in many of the genes in the compendium is still unknown. The interpretation of the significance of driver mutations is also confounded by intratumoural heterogeneity and by the complexity of the ecology of the microenvironment of cancer cells[157,158]. Profiling other dimensions of tumours, such as by transcriptomics, proteomics and methylomics (as performed, for example, in REF.[73]), as well as systematic assays of the function of individual genes and their interactions[159–161] and single-cell profiling approaches[162–165], will contribute to bridging this gap.

A second challenge arises from the fact that while driver genes are identified in isolation by their signals of positive selection, it is in fact a set of driver mutations that drives tumorigenesis[97,102]. For example, driver mutations affecting four specific pathways are known to occur in the vast majority of colorectal adenocarcinomas and are required for the progression of a healthy cell to an invasive carcinoma[68]. Furthermore, while the signals of positive selection in all driver genes in a tumour cohort are equivalent, driver mutations probably occur at different stages of the evolution of a tumour. Again, the clever application of bioinformatics to the analysis of the cancer genome palimpsest has enabled researchers to start resolving this temporal order[166]; nevertheless, more work is needed to understand it.

Finally, there is the challenge of fully understanding how other features besides somatic mutations cooperate in tumorigenesis. While virtually all tumours contain genomic driver mutations[102], these are not sufficient to explain the complete history of cell transformation. Studies of somatic mutations from healthy donors have shown that many cancer drivers are already mutated in non-transformed cells across somatic tissues[167–171]. The same has been shown in other scenarios[172,173] (for example, in clonal haematopoiesis) or benign tumours[169,174,175]. This has led to the conclusion that a certain level of positive selection is present in healthy somatic tissues in a continuum, without reaching the level of cell transformation. In this continuum, positive selection occurs on mutations that confer a fitness advantage, which likely vary between somatic tissues and over time. Thus, a mutation can be a driver only when presented against a background of specific selective constraints. In some cases to reach the level of cell transformation, non-genetic phenotypic changes, such as the stochasticity of gene expression, errors in protein synthesis or certain epigenetic modifications[176], may also be important. Such changes have been documented in processes such as resistance to drugs and metastasis[176–180].

In summary, closing the gap between the list of genes in the compendium and our complete knowledge of the process of tumorigenesis is one of the big challenges of cancer genomics for the near future. In turn, gaining this insight into tumorigenesis will be fundamental to translate our knowledge of cancer genomics into precision cancer medicine.

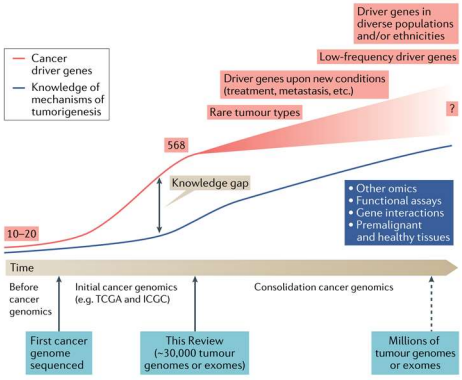


Fig. 9 | **The past, present and future of cancer genomics.** A conceptual representation of the evolution of the compendium of mutational driver genes starting from the identification of the first cancer genes before the start of the cancer genomics era through sequencing of the first tumours to the publication of this Review. It also provides an outlook on the consolidation of cancer genomics (with cancer genomics as a well-established knowledge area) and future trends in cancer genomics research. ICGC, International Cancer Genome Consortium; TCGA, The Cancer Genome Atlas.

Published online: 10 August 2020

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553–1558 (2011).
3. Mwenifumbo, J. C. & Marra, M. A. Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14**, 321–332 (2013).
4. ICGC. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
5. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
This article describes the use of an ensemble of bioinformatics methods to identify mutational cancer driver genes across a large pan-cancer cohort as well as an approach to combine their outputs into a unified list of driver genes.
6. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
7. Tamborero, D. et al. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
8. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
9. Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
10. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* **6**, 107–116 (2006).
11. Kuenzi, B. M. & Ideker, T. A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020).
12. Gonzalez-Perez, A. et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
13. Rubio-Perez, C. et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382–396 (2015).
This article describes an in silico approach to bridging the gap between the identification of driver genes across a large cohort of patients with cancer and describing targeted anticancer therapies potentially benefiting each of the patients.
14. Faguet, G. B. A brief history of cancer: age-old milestones underlying our current knowledge database. *Int. J. Cancer* **136**, 2022–2036 (2015).
15. Greenberg, M. & Selikof, I. J. Lung cancer in the Schneeberg mines: a reappraisal of the data reported by Harting and Hesse in 1879. *Ann. Occup. Hyg.* **37**, 5–14 (1993).
16. Waldron, H. A. A brief history of scrotal cancer. *Occup. Environ. Med.* **40**, 390–401 (1985).
17. Rahman, N. Realising the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
18. Martin, G. S. The road to Src. *Oncogene* **23**, 7910–7917 (2004).
19. Morgan, T. H. The mechanism of Mendelian heredity. http://www.columbia.edu/cu/lweb/digital/collections/cul/texts/ldpd_5998129_000/ (Henry Holt and Company, 1915).
20. Boveri, T. *Zur Frage der Entstehung Maligner Tumoren* (Gustav Fischer, 1914).
21. Bouck, N. & di Mayorca, G. Somatic mutation as the basis for malignant transformation of BHK cells by chemical carcinogens. *Nature* **264**, 722–727 (1976).
22. Shih, C., Shilo, B. Z., Goldfarb, M. P., Dannenberg, A. & Weinberg, R. A. Passage of phenotypes of chemically transformed cells via transfection of DNA and chromatin. *Proc. Natl Acad. Sci. USA* **76**, 5714–5718 (1979).
23. Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc. Natl Acad. Sci. USA* **78**, 1181–1184 (1981).
24. Cooper, G. M., Okenquist, S. & Silverman, L. Transforming activity of DNA of chemically transformed and normal cells. *Nature* **284**, 418–421 (1980).
25. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).
26. Huebner, R. J. & Todaro, G. J. Oncogenes of RNA tumor viruses as determinants of cancer. *Proc. Natl Acad. Sci. USA* **64**, 1087–1094 (1969).
27. Parada, L. F., Tabin, C. J., Shih, C. & Weinberg, R. A. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. *Nature* **297**, 474–478 (1982).
28. Santos, E., Tronick, S. R., Aaronson, S. A., Pulciani, S. & Barbacid, M. T24 human bladder carcinoma oncogene is an activated form of the normal human homologue of BALB- and Harvey-MSV transforming genes. *Nature* **298**, 343 (1982).
29. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149–152 (1982).
This is one in a series of pioneering articles published in 1982 that thoroughly describes the identification of *HRAS* and one of its precise tumorigenic point mutations.
30. Tabin, C. J. et al. Mechanism of activation of a human oncogene. *Nature* **300**, 143–149 (1982).
31. Klein, G. & Klein, E. Evolution of tumours and the impact of molecular oncology. *Nature* **315**, 190–195 (1985).
32. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820 (1971).
33. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat. Rev. Cancer* **8**, 976–990 (2008).
34. Eng, C. & Mulligan, L. M. Mutations of the RET proto-oncogene in the multiple endocrine neoplasia type 2 syndromes, related sporadic tumours, and Hirschsprung disease. *Hum. Mutat.* **9**, 97–109 (1997).
35. Zhuang, Z. et al. Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. *Nat. Genet.* **20**, 66–69 (1998).
36. Hirota, S. et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science* **279**, 577 (1998).
37. Gilliland, D. G. & Griffin, J. D. The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100**, 1532–1542 (2002).
38. Wong, A. J. et al. Structural alterations of the epidermal growth factor receptor gene in human gliomas. *Proc. Natl Acad. Sci. USA* **89**, 2965–2969 (1992).
39. Davies, H. et al. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
40. Laken, S. J. et al. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat. Genet.* **17**, 79–83 (1997).
41. Nigro, J. M. et al. Mutations in the p53 gene occur in diverse human tumour types. *Nature* **342**, 705–708 (1989).
42. Baker, S. et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244**, 217–221 (1989).
43. Grady, W. M. et al. Mutational inactivation of transforming growth factor β receptor type II in microsatellite stable colon cancers. *Cancer Res.* **59**, 320 (1999).
44. Friend, S. H. et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646 (1986).
45. Dunn, J. M., Phillips, R. A., Becker, A. J. & Gallie, B. L. Identification of germline and somatic mutations affecting the retinoblastoma gene. *Science* **241**, 1797–1800 (1988).
46. Li, J. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
47. Kamb, A. et al. A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264**, 436–440 (1994).
48. Meijers-Heijboer, H. et al. Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* **31**, 55–59 (2002).
49. Tavtigian, S. V. et al. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat. Genet.* **12**, 333–337 (1996).
50. Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).
51. Hall, J. M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
52. Wooster, R. et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**, 2088–2090 (1994).
53. Malkin, D. et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233–1238 (1990).
54. Merlo, A. et al. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.* **1**, 686–692 (1995).
55. Powers, M. P. The ever-changing world of gene fusions in cancer: a secondary gene fusion and progression. *Oncogene* **38**, 7197–7199 (2019).
56. Futreal, A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
This study describes the first systematic census of 291 human cancer driver genes collected from the scientific literature; this CGC has since been extensively used as a gold standard to test bioinformatics driver identification methods.
57. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
The first model of the hallmarks that define malignant cell transformation and cancer is developed and described in this review.
58. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
59. Stephens, P. et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* **37**, 590–592 (2005).
60. Sjöblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
This article describes the results of one of the earliest tumour whole-exome sequencing efforts, which paved the way for the development of cancer genomics.
61. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
62. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
63. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
64. Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
65. Pleasance, E. D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
66. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
67. Shah, S. P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
68. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
This review describes frequent alterations to genes, and biochemical pathways driving cancer as well as a simple approach to identify recurrently mutated cancer driver genes across a cohort of tumours.
69. Forrest, W. F. & Cavet, G. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500; author reply 1500 (2007).
70. Getz, G. et al. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500 (2007).
71. Rubin, A. F. & Green, P. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* **317**, 1500 (2007).
72. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
This is the first systematic analysis of genomic alterations affecting tumours across a large cohort of patients with a common malignancy (glioblastoma) performed within the framework of the TCGA consortium.
73. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018).
74. Sanchez-Vega, F. et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337.e10 (2018).
75. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320.e10 (2018).
76. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
77. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

78. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
79. Pleasance, E. et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
80. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
81. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
82. Dees, N. D. et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
83. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
84. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
85. Kamburov, A. et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl Acad. Sci. USA* **112**, E5486 (2015).
86. Porta-Pardo, E. & Godzik, A. e-driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
87. Tokheim, C. et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* **76**, 3719–3731 (2016).
88. Niu, B. et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).
89. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
90. Dietlein, F. et al. Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
91. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
92. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci. USA* **113**, 14330 (2016).
93. Porta-Pardo, E. et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat. Methods* **14**, 782–788 (2017).
94. Tamborero, D. et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
95. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
96. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
97. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
This article describes a method that used the excess of observed somatic mutations with different consequences (which are among the mutational features presented here) to identify cancer driver genes, as well as its application to a large pan-cancer cohort.
98. Mularoni, L. et al. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* https://doi.org/10.1186/s13059-016-0994-0 (2016).
99. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019).
100. Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer* https://doi.org/10.1038/s43018-019-0001-2 (2019).
101. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
102. Sabarinathan, R. et al. The whole-genome panorama of cancer drivers. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/190330v2 (2017).
103. Zhang, J. et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database* **2011**, bar026–bar026 (2011).
104. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
105. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
This article describes a repository of cancer genomics data essential for the discovery of the compendium of mutational driver genes.
106. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
107. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018).
108. Phillips, D. H. Mutational spectra and mutational signatures: insights into cancer aetiology and mechanisms of DNA damage and repair. *DNA Repair* **71**, 6–11 (2018).
109. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
This article presents a bioinformatics approach to identify mutational signatures de novo and describes the first systematic compendium of mutational signatures active across cancer types, and the cause of some of them.
110. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
111. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696 (2018).
112. Hayashi, T. et al. RASA1 and NF1 are preferentially co-mutated and define a distinct genetic subset of smoking-associated non–small cell lung carcinomas sensitive to MEK inhibition. *Clin. Cancer Res.* https://doi.org/10.1158/1078-0432.CCR-17-2343 (2017).
113. Sung, H. et al. Inactivation of RASA1 promotes melanoma tumorigenesis via R-Ras activation. *Oncotarget* **7**, 23885–23896 (2016).
114. Suárez-Cabrera, C. et al. A transposon-based analysis reveals *RASA1* is involved in triple-negative breast cancer. *Cancer Res.* **77**, 1357–1368 (2017).
115. Simanshu, D. K., Nissley, D. V. & McCormick, F. RAS proteins and their regulators in human disease. *Cell* **170**, 17–33 (2017).
116. An, M.-J. et al. Histone demethylase KDM3B regulates the transcriptional network of cell-cycle genes in hepatocarcinoma HepG2 cells. *Biochem. Biophys. Res. Commun.* **508**, 576–582 (2019).
117. Li, J. et al. KDM3 epigenetically controls tumorigenic potentials of human colorectal cancer stem cells through Wnt/β-catenin signalling. *Nat. Commun.* **8**, 1–15 (2017).
118. Mahamdallie, S. et al. Identification of new Wilms tumour predisposition genes: an exome sequencing study. *Lancet Child Adolesc. Health* **3**, 322–331 (2019).
119. Smith, B. et al. The mutational spectrum of FOXA2 in endometrioid endometrial cancer points to a tumor suppressor role. *Gynecologic Oncol.* **143**, 398–405 (2016).
120. Neff, R. et al. Functional characterization of recurrent FOXA2 mutations seen in endometrial cancers. *Int. J. Cancer* **143**, 2955–2961 (2018).
121. Song, Y., Washington, M. K. & Crawford, H. C. Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer Res.* **70**, 2115–2125 (2010).
122. Zhang, Z. et al. FOXA2 attenuates the epithelial to mesenchymal transition by regulating the transcription of E-cadherin and ZEB2 in human breast cancer. *Cancer Lett.* **361**, 240–250 (2015).
123. Zhang, X. et al. Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. *Cancer Discov.* **8**, 108–125 (2018).
124. Jia, L. et al. KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2. *Oncogene* **35**, 2040–2051 (2016).
125. Liu, T. et al. Tumor suppressor bromodomain-containing protein 7 cooperates with Smads to promote transforming growth factor-β responses. *Oncogene* **36**, 362–372 (2017).
126. Cowling, V. H., Turner, S. A. & Cole, M. D. Burkitt's lymphoma-associated c-Myc mutations converge on a dramatically altered target gene response and implicate Nol5a/Nop56 in oncogenesis. *Oncogene* **33**, 3519–3527 (2014).
127. Schmitz, R. et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**, 116–120 (2012).
128. Van Raamsdonk, C. D. et al. Mutations in GNA11 in uveal melanoma. *N. Engl. J. Med.* **363**, 2191–2199 (2010).
129. Radovich, M. et al. The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell* **33**, 244–258. e10 (2018).
130. Hoang, P. H. et al. Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia* **32**, 2459–2470 (2018).
131. Prahallad, A. et al. PTPN11 is a central node in intrinsic and acquired resistance to targeted cancer drugs. *Cell Rep.* **12**, 1978–1985 (2015).
132. Hill, K. S. et al. PTPN11 plays oncogenic roles and is a therapeutic target for BRAF wild-type melanomas. *Mol. Cancer Res.* **17**, 583–593 (2019).
133. Keilhack, H., David, F. S., McGregor, M., Cantley, L. C. & Neel, B. G. Diverse biochemical properties of Shp2 mutants implications for disease phenotypes. *J. Biol. Chem.* **280**, 30984–30993 (2005).
134. Östman, A., Hellberg, C. & Böhmer, F. D. Protein-tyrosine phosphatases and cancer. *Nat. Rev. Cancer* **6**, 307–320 (2006).
135. Qian, Z. et al. Nuclear factor, erythroid 2-like 2-associated molecular signature predicts lung cancer survival. *Sci. Rep.* **5**, 1–10 (2015).
136. Kerins, M. J. & Ooi, A. A catalogue of somatic NRF2 gain-of-function mutations in cancer. *Sci. Rep.* **8**, 1–13 (2018).
137. Stewart, P. A. et al. Proteogenomic landscape of squamous cell lung cancer. *Nat. Commun.* **10**, 1–17 (2019).
138. Hammerman, P. S. et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
139. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
140. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
141. Di Fiore, R., D'Anneo, A., Tesoriere, G. & Vento, R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J. Cell. Physiol.* **228**, 1676–1687 (2013).
142. Goodrich, D. W. The retinoblastoma tumor-suppressor gene, the exception that proves the rule. *Oncogene* **25**, 5233–5243 (2006).
143. Dick, F. A., Goodrich, D. W., Sage, J. & Dyson, N. J. Non-canonical functions of the RB protein in cancer. *Nat. Rev. Cancer* **18**, 442–451 (2018).
144. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
145. Yang, J.-M. et al. Characterization of PTEN mutations in brain cancer reveals that pten mono-ubiquitination promotes protein stability and nuclear localization. *Oncogene* **36**, 3673–3685 (2017).
146. Nguyen, H.-N. et al. A new class of cancer-associated PTEN mutations defined by membrane translocation defects. *Oncogene* **34**, 3737–3745 (2015).
147. Hollander, M. C., Blumenthal, G. M. & Dennis, P. A. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat. Rev. Cancer* **11**, 289–301 (2011).
148. Yin, Y. & Shen, W. H. PTEN: a new guardian of the genome. *Oncogene* **27**, 5443–5453 (2008).
149. Keniry, M. & Parsons, R. The role of PTEN signaling perturbations in cancer and in targeted therapy. *Oncogene* **27**, 5477–5485 (2008).
150. Furnari, F. B., Cloughesy, T. F., Cavenee, W. K. & Mischel, P. S. Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. *Nat. Rev. Cancer* **15**, 302–310 (2015).
151. Xu, H. et al. Epidermal growth factor receptor in glioblastoma. *Oncol. Lett.* **14**, 512–516 (2017).
152. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
153. Gazdar, A. F. Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene* **28**, S24–S31 (2009).

154. Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169–181 (2007).
155. Saunders, G. et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* **20**, 693–701 (2019).
156. Degasperi, A. et al. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).
157. Reiter, J. G. et al. An analysis of genetic heterogeneity in untreated cancers. *Nat. Rev. Cancer* **19**, 639–650 (2019).
158. Maley, C. C. et al. Classifying the evolutionary and ecological features of neoplasms. *Nat. Rev. Cancer* **17**, 605–619 (2017).
159. Dempster, J. M. et al. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nat. Commun.* **10**, 1–14 (2019).
160. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
161. Behan, F. M. et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).
162. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **20**, 1349–1360 (2018).
163. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer* **17**, 557–569 (2017).
164. Levitin, H. M., Yuan, J. & Sims, P. A. Single-cell transcriptomic analysis of tumor heterogeneity. *Trends Cancer* **4**, 264–268 (2018).
165. Wagner, J. et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345.e18 (2019).
166. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
Using clever bioinformatics analyses, this article reconstructs the evolution of mutational processes and driver mutation sequences of 2,658 tumours of 38 cancer types from a single biopsy of each of them.
167. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
168. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
169. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
170. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
171. Colom, B. et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nat. Genet.* **52**, 604–614 (2020).
172. Weaver, J. M. J. et al. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).
173. Gregson, E. M., Bornschein, J. & Fitzgerald, R. C. Genetic progression of Barrett's oesophagus to oesophageal adenocarcinoma. *Br. J. Cancer* **115**, 403–410 (2016).
174. Kanojia, D. et al. Identification of somatic alterations in lipoma using whole exome sequencing. *Sci. Rep.* **9**, 14370 (2019).
175. Ye, L. et al. The genetic landscape of benign thyroid nodules revealed by whole exome and transcriptome sequencing. *Nat. Commun.* **8**, 1–8 (2017).
176. Payne, J. L. & Wagner, A. The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24–38 (2019).
177. Pisco, A. O. & Huang, S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer* **112**, 1725–1732 (2015).
178. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* **13**, 714–726 (2013).
179. Nguyen, D. X., Bos, P. D. & Massagué, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **9**, 274–284 (2009).
180. Ganesh, K. et al. L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat. Cancer* **1**, 28–45 (2020).
181. Tanaga, K. et al. LRP1B attenuates the migration of smooth muscle cells by reducing membrane localization of urokinase and PDGF receptors. *Arterioscler. Thromb. Vasc. Biol.* **24**, 1422–1428 (2004).
182. Li, Y. et al. Low density lipoprotein (LDL) receptor-related protein 1B impairs urokinase receptor regeneration on the cell surface and inhibits cell migration. *J. Biol. Chem.* **277**, 42366–42371 (2002).
183. Wang, Z. et al. Down-regulation of LRP1B in colon cancer promoted the growth and migration of cancer cells. *Exp. Cell Res.* **357**, 1–8 (2017).
184. Tyner, J. W. et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
185. Abida, W. et al. Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl Acad. Sci.* **116**, 11428–11436 (2019).
186. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
187. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
188. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
189. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local determinants of the mutational landscape of the human genome. *Cell* **177**, 101–114 (2019).
This article reviews the effect of several small-scale (spanning up to a few hundred nucleotides) genomic elements and different mutational processes.
190. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).
191. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
This article describes the relationship between the first recognized large-scale genomic feature (replication timing) and the mutation rate in humans.
192. Polak, P. et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
193. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
194. Pich, O. et al. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell* **175**, 1074–1087. e18 (2018).
195. Brown, A. J., Mao, P., Smerdon, M. J., Wyrick, J. J. & Roberts, S. A. Nucleosome positions establish an extended mutation signature in melanoma. *PLOS Genet.* **14**, e1007823 (2018).
196. Sabarinathan, R. et al. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
197. Frigola, J. et al. Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).
198. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547.e23 (2017).
199. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaav2872 (2019).
200. Hess, J. M. et al. Passenger hotspot mutations in cancer. *Cancer Cell* **36**, 288–301.e14 (2019).
201. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 1264–1271 (2018).

Acknowledgements
First and foremost, the authors acknowledge the contribution of patients with cancer, their families and a myriad of medical doctors and cancer genomics researchers who laboriously gather, process and sequence tens of thousands of tumour samples. Without them, the compendium of mutational driver genes would not be possible. They are also greatly indebted to generations of researchers who laid the foundations of cancer genomics, generated and shared data, and developed methods for driver identification. N.L-B. acknowledges funding from the European Research Council (consolidator grant 682398), the Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, European Regional Development Fund) and the Asociación Española Contra el Cáncer (GC16173697BIGA). C.A.-P. is supported by a "la Caixa" Foundation (ID 100010434) fellowship (LCF/BQ/ES18/11670011). H.K. and J.B. are supported by CONTRA innovative training network European Union Horizon 2020 grant MSCA-ITN-2017-766030. O.P. is supported by a Barcelona Institute of Science and Technology Ph.D. fellowship supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the Barcelona Institute of Science and Technology, which is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (Government of Spain) and is supported by CERCA (Generalitat de Catalunya). The results shown here are in whole or in part based on data generated by the TCGA Research Network, the Pan-Cancer Analysis of Whole Genomes, cBioPortal, the Hartwig Medical Foundation, the International Cancer Genome Consortium, St Jude Children's Research Hospital, PedcBioPortal, TARGET projects, the BEAT AML study and several other studies scattered throughout the scientific literature. Finally, the authors state the specific contributions of different authors to the development of IntOGen. IntOGen pipeline conceptualization: F.M-J., F.M, A.G-P. and N.L-B. Combination approach development: F.M. and F.M-J. Reimplementation of driver identification methods: C.A-P., L.M. and F.M-J. Downstream analyses: F.M-J., F.M., O.P., H.K., J.B. and C.A-P. Analysis and discussion of the snapshot of the compendium: F.M-J., F.M., O.P., A.G-P. and N.L-B. Data collection and annotation: I.S., F.M-J. and L.M. IntOGen pipeline development and maintenance: J.D-P., F.M-J., L.M. and I.R-S. IntOGen website development and maintenance: I.R-S. and F.M-J. Project supervision: A.G-P. and N.L-B.

Author contributions
F.M-J., F.M, A.G-P., N.L-B., C.A-P., L.M., O.P., H.K., J.B., I.S., J.D-P. and I.R-S. researched data for the article. F.M-J., A. G-P. and N.L-B. contributed to discussion of the content. F.M-J., A. G-P. and N.L-B. wrote the article. F.M-J., F.M, A.G-P., N.L-B., C.A-P., O.P., H.K., J.B., I.R-S., L.M. and I.S. reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

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

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41568-020-0290-x.

RELATED LINKS
Global Alliance for Genomics and Health: https://www.ga4gh.org
Integrative OncoGenomics: https://www.intogen.org

nature research

https://doi.org/10.1038/s41568-020-0290-x

Supplementary information

A compendium of mutational cancer driver genes

In the format provided by the authors and unedited

Supplementary Information

This document contains Supplementary Information to Martinez-Jimenez *et al.*, *Nat. Rev. Cancer*, 2020, and is composed of three main sections. The first, a document of **Supplementary Methods** to the main manuscript contains technical details of the development of the IntOGen pipeline and its application to collected and annotated datasets of tumor somatic mutations from the public domain. Secondly, two **Supplementary Figures** illustrate specific aspects of the IntOGen pipeline, i.e., the combination of the output of driver identification methods and the comparison of the performance of this combination with that of individual driver identification methods. Finally, a **Supplementary Table** lists relevant information on the cohorts employed to produce the snapshot of the compendium of mutational cancer genes that is described in the main manuscript.

Table of Contents

Supplementary Methods

Data collection and annotation

TCGA

TCGA somatic mutations (mc3.v0.2.8 version) were downloaded from (https://gdc.cancer.gov/about-data/publications/pancanatlas). We then grouped mutations according to their patient's cancer type into 32 different cohorts. Additionally, we kept somatic mutations passing the somatic filtering from TCGA (i.e., column FILTER == "PASS").

PCAWG

PCAWG somatic mutations were downloaded from the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel/). Note that only mutations of ICGC samples can be freely downloaded from this site. The TCGA portion of the callsets is controlled data. To obtain them, we followed the instructions to dowload them that can be found in the same webpage.

cBioPortal

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) cohorts uploaded into cBioPortal that were not part of any other projects included in the analysis (i.e., TCGA, PCAWG, St. Jude or HARTWIG) were downloaded on 2020/01/15 (http://www.cbioportal.org/datasets). We then created cohorts following these criteria:

1. Cohorts with a limited number of samples (i.e., lower than 30 samples) associated to cancer types with extensive representation (such as Breast cancer, Prostate cancer or Colorectal adenocarcinoma) across the compendium of cohorts were removed.
2. Samples were uniquely mapped to a cohort. If the same sample was originally included in two cohorts, we removed the sample from one of them.
3. Mutations from samples that were not obtained from human tumor biopsies were discarded (cell lines, xenografts, normal tissue, etc.).
4. When patient information was available, only one sample of each patient was selected. The criteria to prioritize samples from the same patient was: WXS over WGS; untreated over treated, primary over metastasis or relapse and, finally, by alphabetical order. When there is no patient information we assumed that all patients have only one sample in the cohort.
5. When sequencing platform information was available, samples from the same study but with different sequencing platforms were further subclassified into WXS and WGS datasets (only if the resulting cohorts fulfilled the requirements herein described; otherwise, the samples were discarded).
6. When variant calling information was available, samples from the same cohort and sequencing type were further classified according to their calling algorithm (VarScan, MuTect, etc.). If the resulting cohorts for each subclass fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When variant calling information was not available we assumed that all the samples went through the same calling pipeline.
7. When treatment information was available, samples from the same cohort, sequencing type, calling algorithm were further classified according to their treatment status (i.e, treated versus untreated). If the resulting cohorts from the subclassification fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When information was not available we assumed that samples had not been treated.

8. When biopsy information was available, samples from the same cohort, sequencing type, calling algorithm, treatment status were further classified according to their biopsy type (i.e, primary, relapse or metastasis). If the resulting datasets from the subclassification fulfilled the requirements herein described, the samples were included; otherwise, the samples were discarded. When information was not available we assumed that the biopsy type of the sample was primary.

Hartwig Medical Foundation

Somatic mutations of metastatic WGS from Hartwig Medical Foundation https://www.hartwigmedicalfoundation.nl/en/database/ were downloaded on 2020/01/17 through their platform. Datasets were split according to their primary site. Samples from unknown primary sites (i.e., None, Nan, Unknown, Cup, Na), double primary or aggregating into cohorts of fewer than 7 samples were not considered. A total of 30 different cohorts were thus created.

ICGC

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) studies uploaded in ICGC Data Portal (https://dcc.icgc.org/repositories) not overlapping with other projects included in the analysis (i.e., TCGA, PCAWG, CBIOP or St. Jude) were downloaded on 2018/01/09. We then created cohorts following the criteria used for the cBioPortal datasets (cBioPortal).

St. Jude

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) of Pediatric Cancer Genome Project uploaded in the St. Jude Cloud

(https://www.stjude.cloud/data.html) were downloaded on 2018/07/16. Cohorts were created according to their primary site and their biopsy type (i.e., primary, metastasis and relapse). Resulting datasets with fewer than 5 samples were discarded.

PedcBioPortal

Somatic mutations from Whole Exome Sequencing (WXS) and Whole Genome Sequencing (WGS) studies uploaded in PedcBioPortal that were not part of any other projects included in the analysis (i.e., St. Jude or CBIOP) were downloaded on 2020/01/15 (http://www.pedcbioportal.org/datasets). We then created cohorts following the criteria described in the cBioPortal dataset (cBioPortal).

TARGET

Somatic SNVs from WXS and WGS of two TARGET studies, Neuroblastoma (NB) and Wilms Tumor (WT), from the TARGET consortium were downloaded on 2019/03/07 from the Genomic Data Commons Porta (https://gdc.cancer.gov/).

Beat AML

We downloaded unfiltered somatic mutations from samples included in the Beat AML study from the Genomic Data Commons Porta (https://gdc.cancer.gov/). We next applied the following criteria to create our Beat AML cohort:

1. We focused on somatic single nucleotide variants from VarScan2 using skin as normal control. All samples that did not belong to this class were discarded.
2. Samples from relapses were filtered out.
3. Samples from bone-marrow transplants were discarded.
4. If there were several samples per patient fulfilling the points 1-3, we selected the first in chronological order.

257 independent samples of Beat AML tumors composed our Beat AML cohort.

Literature

We also manually collected publicly available cohorts from the literature. Each cohort was filtered following the same steps mentioned above for the cBioPortal dataset (see above).

Preprocessing

Given the heterogeneity of the datasets analyzed in the current release of intOGen (resulting from differences in the genome aligners, variant calling algorithms, sequencing coverage, sequencing strategy, etc.), we implemented a pre-processing strategy aiming at reducing possible biases. Specifically, we conducted the following filtering steps:

1. The pipeline is configured to run using GRCh38 as reference genome. Therefore, for each input dataset the pipeline requires that the reference genome is defined. Datasets using GRCh37 as reference genome were lifted over using PyLiftover (https://pypi.org/project/pyliftover/; version 0.3) to GRCh38. Mutations failing to liftover from GRCh37 to GRCh38 were discarded.
2. We removed mutations with equal alternate and reference alleles, duplicated mutations within the sample sample, mutations with 'N' as reference or alternative allele, mutations with a reference allele not matching the nucleotide in the reference genome and mutations outside autosomes or sexual chromosomes.
3. Additionally, we removed mutations with low pileup mappability, i.e. mutations in regions that could potentially map elsewhere in the genome. For each position of the genome we computed the pileup mappability, defined as the average uniqueness of all the possible reads of 100bp overlapping a position and allowing up to 2 mismatches. This value is equal to 1 if all the reads overlapping a

mutation are uniquely mappable while it is close to 0 if most mapping reads can map elsewhere in the genome. Positions with pileup mappability lower than 0.9 were removed from further analyses.

4. We filtered out multiple samples from the same donor. The analysis of positive selection in tumors requires that each sample in a cohort is independent from the other samples. That implies that if the input dataset includes multiple samples from the same patient –resulting from different biopsy sites, time points or sequencing strategies– the pipeline automatically selects the first according to its alphabetical order. Therefore, all mutations in the discarded samples are not considered anymore.
5. We also filtered out hypermutated samples. WXS samples carrying more than 1000 mutations or WGS samples with more than 10000 mutations were filtered out if they also exhibited a mutation count greater than 1.5 times the interquartile range above the third quartile of the mutation burden of the cohort were considered hypermutated and therefore removed from further analyses.
6. Datasets without synonymous variants were discarded. Most cancer driver identification methods require synonymous variants to fit a background mutation model. Therefore, datasets with less than 5 synonymous and datasets with a missense/synonymous ratio greater than 10 were excluded .
7. When the Variant Effect Predictor (VEP) mapped one mutation into multiple transcripts associated with different HUGO symbols, we selected the canonical transcript of the first HUGO symbol in alphabetical order.
8. We also discarded mutations mapping into genes without canonical transcript in VEP.92[1].

Methods for cancer driver gene identification

The current version of the intOGen pipeline uses seven cancer driver identification methods to identify cancer driver genes from somatic point mutations: dNdScv[2], cBaSE[3] and MutPanning[4] which test for mutation count bias in genes while correcting for regional genomic covariates[5], mutational processes and coding consequence type; OncodriveCLUSTL[6], which tests for significant clustering of mutations in the protein sequence; smRegions[7], which tests for enrichment of mutations in protein functional domains; HotMAPS[8], which tests for significant clustering of mutations in the 3D protein structure; and OncodriveFML[9], which tests for functional impact bias of the observed mutations. Next, we briefly describe the rationale and the configuration used to run each driver identification method.

dNdScv

dNdScv assesses gene-specific positive selection by inferring the ratio of non-synonymous to synonymous substitutions (dN/dS) in the coding region of each gene. The method resorts to a Poisson-based hierarchical count model that can correct for: i) the mutational processes operative in the cohort determined by the mutational profile of single-nucleotide substitutions with its flanking nucleotides; ii) the regional variability of the background mutation rate explained by histone modifications – it incorporates information about 10 histone marks from 69 cell lines within the ENCODE project[5]; iii) the abundance of sites per coding consequence type across in the coding region of each gene.

We downloaded (release date 2018/10/12) and built a new reference database based on the list canonical transcripts defined by VEP.92 (GRCh38). We then used this

reference database to run dNdScv on all datasets of somatic mutations using the default setting of the method.

OncodriveFML

OncodriveFML is a tool that aims to detect genes under positive selection by analysing the functional impact bias of observed somatic mutations. Briefly, OncodriveFML consists of three steps: in the first step, it computes the average Functional Impact (FI) score (in our pipeline we used CADD[10] v1.4) of coding somatic mutations observed in a gene across a cohort of tumor samples. In the next step, sets of mutations of the same size as the number of mutations observed in the gene of interest are randomly sampled following trinucleotide context conditional probabilities consistent with the relative frequencies of the mutational profile of the cohort. This sampling is repeated N times (with $N = 10^6$ in our configuration) to generate expected average scores across all mutated genes. Finally, it compares the observed average FI score with the distribution expected from the simulations in the form of an empirical p-value. The p-values are then adjusted with a multiple testing correction using the Benjamini–Hochberg (FDR).

OncodriveCLUSTL

OncodriveCLUSTL is a sequence-based clustering algorithm to detect significant linear clustering bias of the observed somatic mutations. Briefly, OncodriveCLUSTL first maps somatic single nucleotide variants (SNVs) observed in a cohort to the genomic element under study. After smoothing the mutation count per position along its genomic sequence using a Tukey kernel-based density function, clusters are identified and scored taking into account the number and distribution of mutations observed. A score for each genomic element is obtained by adding up the scores of its clusters. To estimate the significance of the observed clustering signals, mutations are locally randomized using tri- or penta-nucleotide context conditional probabilities consistent with the relative frequencies of the mutational profile of the cohort.

Within the IntOGen pipeline, OncodriveCLUSTL version 1.1.2 is run for the set of defined canonical transcripts bearing 2 or more SNVs mapping the mutations file. Tuckey-based smoothing is conducted with 11bp windows. The different consecutive coding sequences contained on each transcript are concatenated to allow the algorithm to detect clusters of 2 or more SNVs expanding two exons in a transcript. Simulations are carried out using pre-computed mutational profiles. All cohorts are run using tri-nucleotide context SNVs profiles except for cutaneous melanomas, where penta-nucleotide profiles are calculated. Default randomization windows of 31bp are not allowed to expand beyond the coding sequence boundaries (e.g., windows overlapping part of an exon and an intron are shifted to fit inside the exon). A total number of $N = 10^3$ simulations per transcript are performed. Clustering signals are assessed using analytical p-values.

cBaSE

cBaSE asserts gene-specific positive and negative selection by measuring mutation count bias with Poisson-based hierarchical models. The method allows six different models based on distinct prior alternatives for the distribution of the regional mutation rate. As in the case of dNdScv, the method allows for correction by i) the mutational processes operative in the tumor, with either tri- or penta- nucleotide context; ii) the site count per consequence type per gene; iii) regional variability of the neutral mutation rate.

We run a modified version of the cBaSE script to fit the specific needs of our pipeline. The main modification is adding a rule to automatically select a regional mutation rate prior distribution. Based on the total mutation burden in the dataset, the method runs either an inverse-gamma (mutation count < 12,000), an exponential-inverse-gamma mixture (12,000 < mutation count < 65,000) or a gamma-inverse-gamma mixture (mutation count > 65,000) as mutation rate prior distributions – after communication with

Donate Weghorn, cBaSE's first author). We also skip the negative selection analysis part, as it is not needed for downstream analyses.

Mutpanning

Mutpanning resorts to a mixture signal of positive selection based on two components: i) the mutational recurrence realized as a Poisson-based count model reminiscent to the models implemented at dNdScv or cBaSE; ii) a measure of deviance from the characteristic tri-nucleotide contexts observed in neutral mutagenesis; specifically, an account of the likelihood that a prescribed count of non-synonymous mutations occur in their observed given a context-dependent mutational likelihood attributable to the neutral mutagenesis.

HotMaps3D

HotMAPS asserts gene-specific positive selection by measuring the spatial clustering of mutations in the 3D structure of the protein. The original HotMAPS method assumes that all amino acid substitutions in a protein structure are equally likely. We employed HotMAPS-1.1.3 and modified it to incorporate a background model that more accurately represents the mutational processes operative in a cohort of tumors.

Specifically, we implemented a modified version of the method where the trinucleotide context probability of mutation is compatible with the mutational processes operative in the cohort. Briefly, for each analyzed protein structure harbouring missense mutations, the same number of simulated mutations are randomly generated within the protein structure considering the precomputed mutation frequencies per tri-nucleotide in the cohort. This randomization is performed N times ($N = 10^5$ in our configuration) thereby leading to a background with which to compare the observed mutational data. The rest of the HotMAPS algorithm was not modified.

We downloaded the pre-computed mapping of GRCh37 coordinates into structure residues from the Protein Data Bank (PDB) (http://karchinlab.org/data/HotMAPS/mupit_modbase.sql.gz). We also downloaded (on 2019/09/20) all protein structures from the PDB alongside all human protein 3D models from Modeller (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/H_sapiens_2013.tar.xz). and (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/ModBase_H_sapiens_2013_refseq.tar.xz). We then annotated the structures following the steps described in HotMAPS tutorial (https://github.com/KarchinLab/HotMAPS/wiki/Tutorial-(Exome-scale)).

Since HotMAPS configuration files are pre-built in GRCh37 coordinates and our pipeline is designed to run using GRCh38, for each input cohort, we first converted input somatic mutations to GRCh37, executed the HotMAPS algorithm and transformed the output to coordinates to GRCh38. All conversions were done using the PyLiftover tool (https://pypi.org/project/pyliftover/).

smRegions

SmRegions is a method developed to detect linear enrichment of somatic mutations in user-defined regions of interest. Briefly, smRegions first counts the number of non-synonymous mutations overlapping a Pfam domain in a particular protein. Next, these non-synonymous variants are randomized N times ($N = 10^3$ in our configuration) along the nucleotide sequence of the gene, following the trinucleotide context probability derived from precomputed mutation frequencies per tri-nucleotide in the cohort. The observed and average number of simulated mutations in the Pfam domain and outside of it are compared using a G-test of goodness-of-fit, from which the smRegions p-value is derived. Within the IntOGen pipeline, smRegions discards domains with a number of observed mutations lower than the average from the randomizations. The p-values are adjusted with a multiple testing correction using the Benjamini–Hochberg procedure.

Therefore, the analysis is confined to Pfam domains with a number of observed mutations higher than or equal to the mean simulated number of mutations in the re-sampling.

To create the database of genomic coordinates of Pfam domains we followed the next steps: i) we gathered the first and last amino acid positions of all Pfam domains for canonical transcripts (VEP.92) from BioMart; ii) for each Pfam domain we mapped the first and last amino acid positions into genomic coordinates using TransVar –using GRCh38 as reference genome–; iii) we discarded Pfam domains failing to map either the first or last amino acid positions into genomic coordinates.

smRegions was conceptually inspired by e-driver[11], although significant enhancements to the approach have been introduced. Particularly, i) our background model accounts for the observed tri-nucleotide frequencies rather than assuming that all mutations are equally likely; ii) the statistical test is more conservative; iii) Pfam[12] domains are part of the required input and can be easily updated by downloading the last Pfam release; iv) the method can be configured to any other setting that aims to detect genes possibility selected by enrichment of mutations in pre-defined gene regions.

Combining the outputs of driver identification methods

Rationale

The IntOGen pipeline aims to provide a compendium of driver genes which appropriately reflects the consensus from these seven driver identification methods.

To combine the results of individual statistical tests, p-value combination methods continue to be a standard approach in the field: e.g., Fisher's[13], Brown's[14], and Stouffer's Z-score methods have been used for this purpose. These methods are useful for combining probabilities in meta-analyses, in order to provide a ranking based on combined significance under statistical grounds. However, the application of these methods may bear some caveats:

1. The ranking resulting from p-value combination may lead to inconsistencies when compared to the individual rankings, i.e., they may yield a consensus ranking that does not preserve recurrent precedence relationships found in the individual rankings.
2. Some methods, like Fisher's or Brown's method, may bear anti-conservative performance, thus leading to many likely false discoveries.
3. Balanced (non-weighted) p-value combination methods may lead to faulty results just because of the influence of one or more driver identification method performing poorly for a given dataset.

Weighted methods to combine p-values, like the weighted Stouffer's Z-score, provide some extra room for proper balancing, in the sense of incorporating the relative credibility of each driver identification method. We reasoned that in the context of the combination of the output of driver identification methods, the allocation of weights should account for differences in credibility between methods and across cohorts.

Our combination approach works independently for each cohort. To create a consensus list of driver genes for each cohort, we first determine how credible each driver identification method is when applied to this specific cohort (see Supplementary Figure 1 for a representation of the combinatorial workflow). We do so by tuning a voting weight for each driver identification method that yields a good enrichment of bona-fide cancer genes -- reported in the COSMIC Cancer Gene Census database[15] (CGC) -- in the highly ranked positions of the resulting consensus ranking upon letting each driver

identification method vote. Once a credibility score has been established, we use a weighted method for combining the p-values that each driver identification method gives for each candidate gene: this combination takes the driver identification methods credibility into account. Based on the combined p-values, we conduct FDR correction to conclude a ranking of candidate driver genes alongside q-values.

Weight Estimation by Voting

The relative credibility awarded to each method is based on the ability of the method to give precedence to well-known genes already collected in the CGC catalog of validated driver genes. As each method yields a ranking of driver genes, these lists can be combined using a voting system –Schulze's voting method. The method allows us to consider each method as a voter with some voting rights (weighting) which casts ballots containing a list of candidates sorted by precedence. Schulze's method takes information about precedence from each individual method and produces a new consensus ranking[16].

Instead of conducting balanced voting, we tune the voting rights of the methods so that the enrichment of CGC genes at the top positions of the consensus list is maximized. We limit the share each method can attain in the credibility simplex –up to a uniform threshold. The resulting voting rights are deemed the relative credibility of each method.

Ranking Score

Upon selection of a catalog of bona-fide known driver elements (the Cancer Gene Census, or CGC) we can provide a score for each ranking R of genes as follows:

$$E(R) = \sum_{i=1}^{N} \frac{p_i}{\log(i+1)}$$

where p(i) is the proportion of elements with rank smaller (closer to top) than i which belong to CGC and N is a suitable threshold to consider only the N top ranked

elements. Using E(R) we can define a function that maps each weighting vector w (in the simplex of methods weights) to a value E(R(w)) where R(w) denotes the consensus ranking obtained by applying Schulze's voting with voting rights given by the weighting vector w.

Optimization with constraints

Finally we are bound to find a good candidate for

$$\hat{w} = argmax\ E(R(w))$$

For each method to have chances to contribute to the consensus score, we impose the mild constraint of limiting the share of each method to 0.3.

Optimization is then carried out in two steps: we first find a good candidate $\hat{w}_0$ by exhaustive search in a rectangular grid satisfying the constraints defined above (with grid step=0.05); in the second step we take $\hat{w}_0$ as the seed for a stochastic hill-climbing procedure (we resort to Python's scipy.optimize "basinhopping", method=SLSQP and stepsize=0.05).

Estimation of combined p-values using weighted Stouffer's Z-score

Using the relative weight estimate that yields a maximum value of the objective function *f* we combined the p-values resorting to the weighted Stouffer's Z-score method. Thereafter we performed Benjamini-Hochberg FDR correction with the resulting combined p-values, yielding one q-value for each genomic element. If the element

belongs to the CGC, we computed its q-value using only the collection of p-values of CGC genes. Otherwise, we computed the q-value using the p-values computed for all genes.

Tiers of driver genes from sorted list of combined rankings and p-values

To finalize the analysis we considered only genes with at least two mutated samples in the cohort under analysis. These genes were classified into four groups according to the level of evidence in that cohort that the gene harbours signal of positive selection.

For the sake of simplicity, we give some conventions before proceeding to describe the groups. For each gene G we have defined a rank r(G) and a significance q-value q(G) according to the voting and p-value combinations described above. Given the final ranked list of genes we can define two rank cutoffs that depend on a prescribed significance level t:

$$R = \min_{G} \{r(G) \mid q(G) > t\} - 1$$

$$r = \max_{G} \{r(G) \mid q(G) < t\}$$

It is readily seen that r < R+1. By default the significance level t is set to 0.05.

1. The first group of genes, TIER1, contains genes showing high confidence and agreement in their positive selection signals. TIER1 comprises all the genes G such that r(G) < r.

2. The second group, TIER2, was devised to contain known cancer driver genes, showing mild signals of positive selection, that were not included in TIER1. More in detail, we defined TIER2 genes as those CGC genes, not included in TIER1, whose CGC q-value was lower than a prescribed significance level (default CGC q-value=0.25). The CGC q-value is computed by performing FDR of the combined p-values albeit restricted to CGC genes.
3. The third group, TIER3, encompasses genes G that are not included in TIER1 or TIER2 which fulfill that r(G) < R.
4. All genes not included in the aforementioned classes are considered non-driver genes.

Combination benchmark

We have assessed the performance of the combination compared to i) each of the seven individual methods and ii) other strategies to combine the output from cancer driver identification methods.

Finally, we evaluated the contribution of each of the individual methods to the consensus list of driver genes.

Datasets for evaluation

We decided to perform an evaluation based on the 32 Whole-Exome cohorts of the TCGA PanCanAtlas project (downloaded from *https://gdc.cancer.gov/about-data/publications/pancanatlas*). These cohorts sequence coverage, sequence alignment and somatic mutation calling were performed using the same methodology guaranteeing that biases due to technological and methodological artifacts are minimal.

The Cancer Genes Census –version v87– was downloaded from the COSMIC data portal (*https://cancer.sanger.ac.uk/census*) and used as a positive set of known cancer driver genes.

We created a catalog of genes with evidence of not involvement in cancerogenesis. This set includes very long genes (HMCN1, TTN, OBSCN, GPR98, RYR2 and RYR3), and a list of olfactory receptors from Human Olfactory Receptor Data Exploratorium (HORDE) (https://genome.weizmann.ac.il/horde/; download date 14/02/2018). In addition, for all TCGA cohorts, we added non-expressed genes, defined as genes where at least 80% of the samples showed a RSEM expressed in $\log_2$ scale smaller or equal to 0. Expression data for TCGA was downloaded from *https://gdc.cancer.gov/about-data/publications/pancanatlas*.

Metrics for evaluation

We defined a metric, referred to as CGC-Score, that is intended to measure the quality of a ranking of genes as the enrichment of CGC elements in the top positions of the ranking; specifically given a ranking R mapping each element to a rank, we define the CGC-Score of R as:

$$S(R) = \sum_{i=1}^{N} \frac{p(i)}{\log(i+1)} \Big/ \sum_{i=1}^{N} \frac{1}{\log(i+1)}$$

where p(i) is the proportion of elements with rank $\leq i$ that belong to CGC and N is a convenient threshold to consider just the top elements in the ranking (by default N=40). We estimated the CGC-Score across TCGA cohorts for all the rankings given by individual methods and by the consensus ranking.

Similarly, we defined a metric, referred to as Negative-Score, that aims to measure the proportion of non-cancer genes among the top positions in the ranking. Specifically, given a ranking R, we define the Negative-Score of R as:

$$N(R) = \sum_{i=1}^{N} \frac{n(i)}{\log(i+1)} \; / \; \sum_{i=1}^{N} \frac{1}{\log(i+1)}$$

where n(i) is the proportion of elements with rank $\leq i$ that belong to the negative set and N is a suitable threshold to consider just the top elements in the ranking (by default N = 40). We estimated the Negative-Score across TCGA cohorts for all the rankings given by individual methods and by the consensus ranking.

Comparison with individual methods

We compared the CGC-Score and Negative-Score of the combined lists of drivers with the individual outputs of the seven driver discovery methods integrated in the pipeline.

We observed a consistent increase in CGC-Score of the combinatorial strategy compared to any individual method across 23/32 (71%) of the TCGA cohorts (Supplementary Figure 2a and 2b). Similarly, we observed a consistent decrease in Negative-Score across TCGA cohorts, where the combinatorial strategy ranked the least enriched in non-cancer genes in 14 (43%) cohorts and in none of them was the most enriched in non-cancer genes (Supplementary Figure 2c).
In summary, the evaluation shows that the combinatorial strategy increases the True Positive Rate (measured using the CGC-Score) preserving lower False Positive Rate (measured using the Negative-Score) than the seven individual methods included in the pipeline.

Comparison with other combinatorial selection methods

We then computed the CGC-Score and Negative-Score based on the consensus ranking from the aforementioned combinatorial methods and compared them to our Schulze's weighted combination ranking across all TCGA cohorts. Our combinatorial approach met a larger enrichment in known cancer genes for 30/32 (93%) TCGA cohorts (Supplementary Figure 2d).

Leave-one-out combination

We aimed to estimate the contribution of each method's ranking to the final ranking after Schulze's weighted combination. To tackle this question, we measured the effect of removing one method from the combination by, first, calculating the CGC-Score of the combination excluding the aforementioned method and, next, obtaining its ratio with the original combination (i.e., including all methods). This was iteratively calculated for all methods across TCGA cohorts. Methods that positively contributed to the combined ranking quality show a ratio below one, while methods that negatively contributed to the combined ranking show a ratio greater than one.

We observed that across TCGA cohorts most of the methods contributed positively (i.e., ratio above one) to the weighted combination performance (Supplementary Figure 2e). Moreover, there is substantial variability across TCGA cohorts in the contribution of each method to the combination performance. In summary, all methods contributed positively to the combinatorial performance across TCGA supporting our methodological choice for the individual driver discovery methods (Supplementary Figure 2e).

Drivers postprocessing

The intOGen pipeline outputs a ranked list of driver genes for each input cohort. We aimed to create a comprehensive compendium of driver genes per tumor type from all the cohorts included in this version.

Then, we performed a filtering on automatically generated driver gene lists per cohort. This filtering is intended to reduce artifacts from the cohort-specific driver lists, due to e.g. errors in calling algorithms, local hypermutation effects, undocumented filtering of mutations.

We first created a collection of candidate driver genes by selecting either: i) significant non-CGC genes (q-value < 0.05) with at least two significant bidders (methods rendering the genes as significant); ii) significant CGC genes (either q-value < 0.05 or CGC q-value < 0.25) from individual cohorts. All genes that did not fulfill these requirements were discarded.

Additionally, candidate driver genes were further filtered using the following criteria:

1. We discarded non-expressed genes using TCGA expression data. For tumor types directly mapping to cohorts from TCGA –including TCGA cohorts– we removed non-expressed genes in that tumor type. We used the following criterion for non-expressed genes: genes where at least 80% of the samples showed a negative log2 RSEM. For those tumor types which could not be mapped to TCGA cohorts this filtering step was not done.
2. We also discarded genes highly tolerant to Single Nucleotide Polymorphisms (SNP) across human populations. Such genes are more susceptible to calling errors and should be taken cautiously. More specifically, we downloaded transcript specific constraints from gnomAD (release 2.1; 2018/02/14) and used the observed-to-expected ratio score (oe) of missense (mys), synonymous (syn) and loss-of-function (lof) variants to detect genes highly tolerant to SNPs. Genes enriched in SNPs (oe_mys > 1.5 or oe_lof > 1.5 or oe_syn > 1.5) with a number of mutations per sample greater than 1 were discarded. Additionally, we discarded mutations overlapping with germline variants (germline count > 5) from a panel of normals (PON) from Hartwig Medical Foundation

(https://nextcloud.hartwigmedicalfoundation.nl/s/LTiKTd8XxBqwaiC?path=%2FHMFTools-Resources%2FSage).

3. We also discarded genes that are likely false positives according to their known function from the literature. We convened that the following genes are likely false positives: i) known long genes such as TTN, OBSCN, RYR2, etc.; ii) olfactory receptors from HORDE (http://bioportal.weizmann.ac.il/HORDE/; download date 2018/02/14); iii) genes not belonging to Tier1 CGC genes lacking literature references according to CancerMine[17] (http://bionlp.bcgsc.ca/cancermine/).
4. We also removed non CGC genes with more than 3 mutations in one sample. This abnormally high number of mutations in a sample may be the result of either a local hypermutation process or cross contamination from germline variants.
5. Finally we discarded genes whose mutations are likely the result of local hypermutation activity. More specifically, some coding regions might be the target of mutations associated with COSMIC Signature 9 (https://cancer.sanger.ac.uk/cosmic/signatures) which is associated with non-canonical AID activity in lymphoid tumours. In those cancer types were Signature 9 is known to play a significant mutagenic role (i.e., AML, Non-Hodgkin Lymphomas, B-cell Lymphomas, CLL and Myelodysplastic syndromes) we discarded genes where more than 50% of mutations in a cohort of patients were associated with Signature 9.

Candidate driver genes that were not discarded composed the compendium of driver genes.

Classification according to annotation level from CGC

We then annotated the catalog of highly confident driver genes according to their annotation level in CGC version 87. Specifically, we created a three-level annotation: i) the first level included driver genes with a reported involvement in the source tumor type according to the CGC; ii) the second group included CGC genes lacking reported

association with the tumor type; iii) the third group included genes that were not present in CGC.

To match the tumor type of our analyzed cohorts and the nomenclature/acronyms of cancer types reported in the CGC we manually created a dictionary comprising all the names of tumor types from CGC and cancer types defined in our study, according to the following rules:

1. All the equivalent terms for a cancer type reported in the CGC using the Somatic Tumor Type field (e.g. “breast”, “breast carcinoma”, “breast cancer”), were mapped into the same tumor type.
2. CGC terms with an unequivocal mapping into our cancer types were automatically linked (e.g., “breast” with “BRCA”).
3. CGC terms representing fine tuning classification of a more prevalent cancer type that did not represent an independent cohort in our study, were mapped to their closest parent tumor type in our study (e.g., “malignant melanoma of soft parts” into “cutaneous melanoma” or “alveolar soft part sarcoma” into “sarcoma”).
4. Adenomas were mapped to carcinomas of the same cell type (e.g.,”hepatic adenoma” into “hepatic adenocarcinoma”, “salivary gland adenoma” into “salivary gland adenocarcinoma”).
5. CGC parent terms mapping to several tumor types from our study were mapped to each of the potential child tumor types. For instance, the term “non small cell lung cancer” was mapped to “LUAD” (lung adenocarcinoma) and “LUSC” (lung squamous cell carcinoma).
6. Finally, CGC terms associated with benign lesions, with unspecified tumor types (e.g., “other”, “other tumor types”, “other CNS”) or with tumor types with missing parents in our study were left unmatched.

Mode of action of driver genes

We computed the mode of action for highly confident driver genes. To do so, we first performed a pan-cancer run of dNdScv across all TCGA cohorts. We then applied the aforementioned algorithm (see Mode of action section below for more details on how the algorithm determines the role of driver genes according to their distribution of mutations in a cohort of samples) to classify driver genes into the three possible roles: Act (activating or oncogene), LoF (loss-of-function or tumor suppressor) or Amb (ambiguous or non-defined). We then combined these predictions with prior knowledge from the Cancer Genome Interpreter[18] according to the following rules: i) when the inferred mode of action matched the prior knowledge, we used the consensus mode of action; ii) when the gene was not included in the prior knowledge list, we selected the inferred mode of action; iii) when the inferred mode of action did not match the prior knowledge, we selected that of the prior knowledge list.

Repository of mutational features

Linear clusters

Linear clusters for each gene and cohort were identified by OncodriveCLUSTL. We defined as significant those clusters in a driver gene with a p-value lower than 0.05. The start and end of the clusters were retrieved from the first and last mutated amino acid overlapping the cluster, respectively.

3D clusters

Information about the positions involved in the 3D clusters defined by HotMAPS were retrieved from the gene specific output of each cohort. We defined as significant those amino acids in a driver gene with a q-value lower than 0.05.

Pfam Domains

Pfam domains for each driver gene and cohort were identified by smRegions. We defined as significant those domains in driver genes with a q-value lower than 0.1 and with positive log ratio of observed-to-simulated mutations (observed mutations / simulated mutations > 1). The first and last amino acids are defined from the start and end of the Pfam domain, respectively.

Excess of mutations

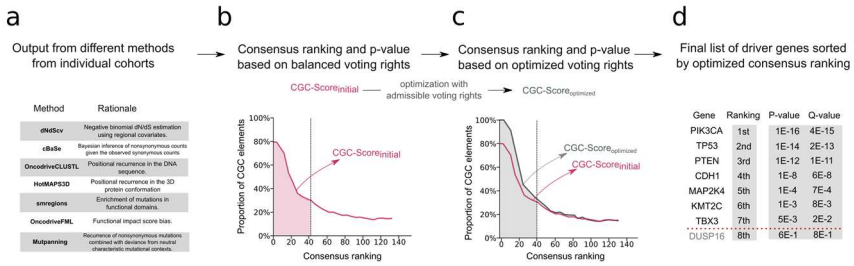
The so-called excess of mutations for a given coding consequence-type quantifies the proportion of observed mutations at this consequence-type that are not explained by the neutral mutation rate. The excess is computed from the consequence-type specific dN/dS estimates ω_c as $(\omega_c - 1) / \omega_c$. We computed the excess for missense, nonsense and splicing-affecting mutations according to the canonical transcript.

Mode of action

Upon the consequence-type specific dN/dS estimates for nonsense and missense mutations computed at each gene, denoted ω_{mis} and ω_{non}, we deemed a gene activating or Act (resp. Loss-of-function or LoF) if $\omega_{mis} - \omega_{non} > \varepsilon$ (resp. $\omega_{non} - \omega_{mis} > \varepsilon$) with $\varepsilon = 0.1$. Genes with $|\omega_{mis} - \omega_{non}| < \varepsilon$ as well as genes with $\omega_{mis} < 1$ were deemed to have an “ambiguous” mode of action.

Supplementary Figures

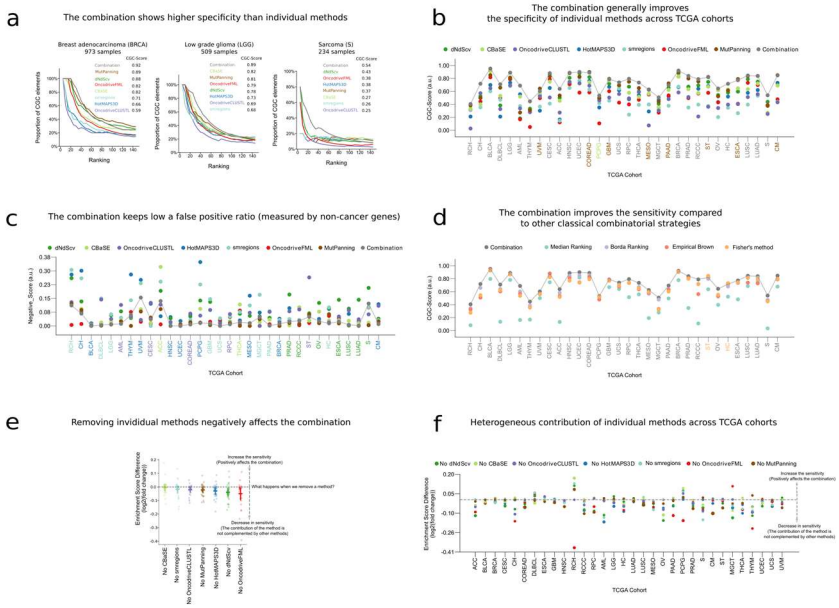
Figure Supplementary 1



Supplementary Figure 1. Schematic representation of the approach to combine the output of driver discovery methods.

a) Given the output of the seven driver discovery methods integrated in intOGen, b) the pipeline dynamically estimates the credibility of the output of each method based on its enrichment for Cancer Gene Census genes. Then in c) it performs the combination of the outputs weighting each method output according to the credibility previously allocated. Finally in d), the resulting list of drivers is sorted by the optimized consensus ranking and their associated combined p-value.

Figure Supplementary 2



Supplementary Figure 2. Benchmark of the IntOGen combination using TCGA cohorts.

a) The proportion of CGC drivers among the top ranking genes in the combined list is greater than that of the lists of individual driver identification methods in three exemplary TCGA cohorts (BRCA, LGG and Sarcoma). The proportion of CGC drivers in each list of genes is measured across growing top ranked genes (x-axis). To summarize the proportion of CGC drivers obtained throughout all values of rank tested, a numeric value (CGC score) is derived (see Supplementary Methods).

b) CGC score of the output of all driver discovery methods and the combined list across 32 TCGA cohorts. Systematically, the combined list exhibits a CGC score which is at least equal to that of the best performing individual method. In most cases, the combined list exhibits a higher CGC score than that of any individual method.

c) For any drivers list we can also compute a potential false positives score or Negative Score, tracking the proportion of a set of non driver genes (known “fishy” genes of driver identification, and not expressed genes in each tissue) within the top-ranking elements of the list. The Negative Score of the combined list across all TCGA cohorts is comparable to that of methods with the lowest Negative Score. This means that the increase in sensitivity of drivers identification in the combined list that is documented in a) and b) does not come at the cost of a reduction of specificity.

d) Comparison of the CGC Score of the combined list with that obtained using classic combination strategies across all TCGA cohorts. The combination approach developed in the pipeline exhibits higher sensitivity than any other strategy across all cohorts.

e) To assess the contribution of each individual method to the combined list of drivers, we carried out a systematic leave-one-out analysis across all TCGA cohorts (dots in each distribution). We then evaluated the sensitivity of the new combination using the CGC Score. In most cohorts, the elimination of a method from the combination causes a decrease of sensitivity.

f) The effect of eliminating one method on the sensitivity of the combination changes across cohorts.

Supplementary Table

Supplementary Table 1. Summarized list of cohorts employed to produce the snapshot of the compendium of cancer genes described in the main manuscript.
The list of cohorts collected from the public domain and employed in the construction of subsequent snapshots of the compendium will be updated and published regularly in the IntOGen website (www.intogen.org).

Bibliography

1. McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 1–14 (2016).
2. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
3. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
4. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
5. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
6. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019).
7. Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nat. Cancer* **1**, 122–135 (2020).
8. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res.* **76**, 3719–3731 (2016).
9. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with

cancer driver mutations. *Genome Biol.* **17**, 128 (2016).

10. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
11. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinforma. Oxf. Engl.* **30**, 3109–3114 (2014).
12. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
13. Mosteller, F. & Fisher, R. A. Questions and Answers. *Am. Stat.* **2**, 30–31 (1948).
14. Brown, M. B. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* **31**, 987–992 (1975).
15. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
16. Schulze, M. The Schulze Method of Voting. *ArXiv180402973 Cs* (2019).
17. Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R. & Jones, S. J. M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* **16**, 505–507 (2019).
18. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).

4. DISCUSSION

In the Chapter 1 project, we had the opportunity to analyze a unique cohort of cancer patients suffering from a rare adult disease called T-cell acute lymphoblastic leukemia to shed some light into their treatment response and for a better understanding of the leukemogenic process. There are few landscape genomic studies focused on T-ALL adult patients and none (to our knowledge) looking at the clonal evolution of their relapsed leukemias [238,242,246,249]. We have studied 19 T-ALL adult patients from their cancer evolution to their genomic characteristics in comparison with other ALL forms.

According to our findings, the relapsed leukemia of the majority of these patients arises from a population of relapse-fated cells already existing at time of diagnosis. Prior studies have also pointed out the therapy-preexisting origin of relapse in B-ALL [112,182,254,255]. Inspired by one of them, Li et al., 2020 [112], we computed the doubling time of the T-lymphoblasts through the percentage counts in serial measurements during treatment (remission and relapse timepoints) assessed by the pathologist. Therefore, given the estimated doubling time we could infer the number of relapse cells expected at time of diagnosis. The precise relapse population size at diagnosis is something that must be taken with caution since we only had two measures of blasts per patient, thus, we actually have an aggregated doubling time. However, these estimates are robust enough to say that the majority of the studied cases had a preexisting clone assumed to be larger than 1 cell and that only three patients fit more with a relapse-population emerging during treatment. With more timepoints of the emergence of relapse plus a bigger cohort size, it would be interesting to see whether there are any differences between immunophenotypes of T-ALL. Another important line of evidence that, indeed, there is mostly a preexisting relapse

subpopulation at diagnosis, is the fact that we cannot detect thiopurine treatment associated signature in relapse samples as shown recently in B-ALL [112]. Imagine a single or few leukemic persistent cells, damaged by the 6-mercaptopurine, that after treatment, make a clonal expansion. Then, therapy-induced mutations would have been fixed in the relapse population leaving a traceable pattern and allowing the detection of the signature. However, given the short time between diagnosis and relapse (lots of early relapsers in the in-house cohort[21]) and the lack of evidence of this signature, it also points towards an already quite large relapse-fated population at diagnosis.

In general, the initial clone from where the relapse arises has therefore diverged before diagnosis and has accumulated private mutations becoming a "branch" of the evolving leukemia. In order to have an estimate of when the divergence between the primary and relapse happened we modeled the contribution of signature 5 mutations which are considered to accumulate in a clock-like manner. We noticed that although there is an overall good linear correlation between the activity of signature 5 and the age of patients in our in-house cohort, a direct conversion of the shared number of mutations between primary and relapse to the corresponding chronological age of each patient was overestimating the time of divergence. We also observed that the intercept of the linear regression between the number of mutations of signature 5 and the age of patients was exceeding from 0 (Figure 4.a of the manuscript) which we believed it might be due to some acceleration in the mutation rate of this signature. For this reason, we considered the accumulation of the signature 5 mutations to be constant through time until a certain point (e.g. leukemic transformation) in which

[21] median 9.1 months and mean 9.3 months which is a bit early compared to median 11 reported here [296] and within "early" relapse category in Li et al., 2020 paper [112]

most likely it starts to increase. For that, we considered as reference, the constant mutation rate estimated for healthy hematopoietic stem cells published by others [109] and then assumed different moments from which the mutation rate started incrementing to fit the actual observations. We tried two increments of the mutation rate (a constant one and a linear one) and simulated them under a certain plausible variability departing at different moments (i.e years) before diagnosis. Each one of these simulated outputted models give an estimate of the divergence but still only the ones with bigger likelihoods are used in the final estimation of it. Therefore, each one of them contributes to a final estimate and grants a reasonable error margin making it robust enough. The majority of the patients showed that the divergence between primary and relapse clones likely happened the previous year to the diagnosis of the leukemia. Therefore, our findings suggest that in the majority of the studied cases, the relapse clone lineage started within the year prior to primary detection and constitutes a subclonal population at time of diagnosis of the primary.

In light of the results, one of the obvious questions is whether we can detect the relapse before it creates a full grown second leukemia. For that, we have checked which are the relapse-enriched mutations that can help predict the relapse and are suspicious of providing treatment-resistance advantage. We have identified relapse-specific alterations in known treatment-resistant associated genes such as NT5C2, ABCB1 and also a new candidate: SMARCA4. Little is known about the involvement of SMARCA4 in leukemogenesis and/or resistance but it has been detected exclusively in T-ALL in primary samples [234,248] and also it has appeared mutated in a relapse-specific manner [254,297]. We decided to see whether we could detect the relapse SMARCA4 mutations of the two affected patients at low allele frequency in the corresponding primary samples with a dPCR that would give us deeper resolution than the WGS. This way, we wanted to

check whether we can trail the relapse clone at time of diagnosis. Results were negative which are in concordance with the estimated relapse clone size being below the limit of detection for the two patients tested (Figure 5 a). On the one hand, that gave us more confidence that the estimate of the doubling time and the corresponding computation of the number of relapse cells at diagnosis is pretty accurate. On the other hand, it evidenced the need for finding ways for relapse prevention. In relation to that, it seems that relapse subclones at diagnosis are not always detectable at primary samples so a close monitoring of MRD along the treatment seems the best solution for relapse prevention [298]. In fact, the assessment of the MRD at the end of induction has proven to be of high value to stratify patients according to risk and is now widely used [281,298]. Therefore, the analysis of genomic markers such as, mutations in genes with bad prognosis, in serial aspirates or blood extractions during treatment can help to early detect a change in the clonal dynamics. This way, helping to anticipate the emergence of the resistant relapse clone, especially for those slow-responders to treatment with persistent MRD as shown in here [254]. This type of tracking requires the highest sensitivity and quantification of the mutations which can be very costly. Therefore, a combination of techniques for this type of monitoring such as keeping the morphological assessment at the end of induction but trying droplet dPCR or ultra-deep sequencing for the rest of the checkpoints seems more reasonable, as well as, the importance of treating patients within specialized centers which are more likely the ones guaranteeing such tracking. Apart from that, it would be interesting to perform some functional analysis with SMARCA4 to understand its involvement in ALL. Linking these results with the previous paragraph, given that our estimates dated the relapse divergence within the prior year to primary diagnosis, it seems crucial to early detect the primary. In other words, the sooner we stop the primary leukemia progression the better we avoid the relapse-resistant clone evolution.

Another thing that drove our attention is the way in which the NOTCH1 pathway is mutated. First, we detected cases of convergent evolution in which primary and relapse clones harbor different mutations in the same pathway genes (two different NOTCH1 mutations or NOTCH1 and FBXW7 private mutations to one or the other leukemia). In these cases, perhaps, the relapse clone is able to tolerate or resist the treatment by any genomic mechanism and the success of its progression is due to mutations in the NOTCH1 pathway as it is one of the most important signaling pathways for proliferation in T-ALL. In the past, clinical studies checking the prognostic value of NOTCH1 and FBXW7 have reported different results so further studies must be done to clarify it [241,299–301]. The combination of a clonal and a subclonal mutation of NOTCH1 pathway genes at time of diagnosis should serve to early detect a shift of the clonal dominance if a close MRD monitoring is settled and may help prevent refractory or relapsed patients. Moreover, we detected patients with multiple mutations in the same gene simultaneously. It has been observed that co-ocurrant mutations in HD and PEST domain of NOTCH1 in *cis* (same allele) cause a synergistic effect and overactivation of NOTCH1 [241,302]. It would be interesting to study the implications of NOTCH1 double mutants in adult T-ALL. In a recent study [302], the authors used CisChecker which is an algorithm that can be used for NGS data to infer whether the mutations are in *cis* or *trans*. Otherwise using Nanopore technology would also allow to sequence NOTCH1 gene and check it. Again, with a bigger cohort and serial samples of each patient, it would be interesting to see to what extent double mutants of NOTCH1 can decrease the doubling time and therefore, increase proliferation, and compare these types of patients with other non-double mutants in T-ALL. In addition, it would be interesting to relate this with clinical data of the patients, such as incidence of CNS relapse, survival and other metrics. Another related line

of investigation discussed along the project was to quantify the fitness advantage of mutations in NOTCH1 and determine the fitness effects of T-ALL main driver genes found such as PHP6 and RAS mutations in a similar way as it has been shown recently with AML drivers appearing in CH cases [303]. Unfortunately, with the size of our cohort it was not possible to accurately compute that but this is some analysis that is interesting to pursue in the future.

One of the main goals of our study in Chapter 1 was to find mechanisms of resistance. Obvious candidates have not become apparent for all the patients. As a consequence, we asked ourselves whether we could distinguish a relapse driven by a non-resistant survivor cell population (*non-resistant scenario*) from a relapse driven by a genetic resistance to the treatment (*resistant scenario*) regardless of the specific mechanism. Since there is an increasing evidence that the bone marrow niche can provide protection to the leukemic blasts [178,293] against the treatment, perhaps a non-resistance scenario would be a niche-protected group of leukemic cells not harboring any genomic-resistant mechanism that manage to survive and trigger a relapse. Whereas a resistant scenario would be, for example, a leukemic cell/s harboring NT5C2 and avoiding mercaptopurine damage. After simulating both scenarios we looked at the CCF at time of diagnosis of the clonal relapse mutations (those fixed in the relapse cell population) and compared them to our real data. In the non-resistant scenario, under different parameters but simulated with the observed elapsed time between the two leukemias, any of the undetectable mutations in the primary is able to get fixed in relapse making it an unrealistic situation. On the contrary, resistant simulations where a subgroup of cells carrying a resistant mutation are selected, generate a similar scenario to the one observed in our patients. Therefore, we are inclined to believe that, in this particular cohort, all patients must have a relapse driven by therapy resistance, regardless of the

concrete treatment-resistant mechanism. In other words, the relapses of our cohort seem to be driven by genetic resistance which implies that there are still resistant genomic mechanisms (unusual driver mutations, altered co-occurrences, epigenetic changes…) to be discovered. In this sense, this project evidenced the importance to generate tumoral data for those tumor types with low incidence and to study cases with cancer conditions less explored, such as relapse tumors or metastasis, to be able to increase the compendium of mutational driver genes. The pipeline of IntOGen and the whole system presented in Chapter 2 represents an important step towards the completion of the list of all cancer driver genes. The implementation of this framework has been optimized to facilitate the analysis of new data. As more datasets of understudied malignancies and conditions are available and fitted into the workflow, more complete snapshots of this compendium will be generated. Nevertheless, still many challenges are to be solved, as the detection of driver mutations in genes is necessary but not sufficient to understand the whole picture of tumorigenesis.

Another lesson learned from Chapter 1 study is the confirmation of T-ALL being a different entity from B-ALL. Although the active mutational processes in primary leukemias are the same between T and B-ALL, we can appreciate important differences in the pathways and altered genes driving each one of these ALL forms. In addition, we have computed a different doubling time for T-ALL compared to the one previously computed for B-ALL. Although it might be that these differences are due to age (B-ALL samples were pediatric and ours are adults), differences can also be caused by the different biology behind the lineage or cell of origin which might also greatly influence the tumoral growth and dynamics. Future studies would enlarge our knowledge regarding these two ALL diseases and with a better understanding there will come improvements in the therapeutic opportunities to cure these patients. In fact, it seems that with the advances

CAR-T for both leukemic lineages, more patients will benefit from this therapy, most likely leading to increased survival rates.

5. CONCLUSIONS

- In most adult T-ALL cases that recur, the relapse clone diverged from the primary within the year prior to primary diagnosis, by which time, its size ranges between a few and millions of cells, but below the limit of detection by bulk sequencing.

- The relapse clone most likely harbors genomic alterations that confer therapy resistance.

- The progression of T-ALL in some of the studied cases is characterized by convergent evolution of mutations affecting NOTCH1 pathway genes.

- The mutational processes detected in primary leukemias of B-ALL and T-ALL are very similar; moreover, there is no evidence of chemotherapy-related signatures in relapse adult T-ALLs, unlike in the pediatric malignancy.

- We identify well-known resistant mechanisms such as mutations in NT5C2 and also potential resistance alterations in less studied genes such as SMARCA4 and ABCB1 which appear in a relapse-specific manner in adult T-ALL cases.

- We have identified 568 mutational cancer driver genes across 66 cancer types; whereas some of these drive tumorigenesis across many cancer types (widespread), the majority are specific of one or two malignancies

6. BIBLIOGRAPHY

1. Cancer [Internet]. [cited 2020 Mar 28]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer
2. Hajdu SI. A note from history: Landmarks in history of cancer, part 1. Cancer. 2011;117:1097–1102.
3. A to Z List of Cancer Types - National Cancer Institute [Internet]. [cited 2020 Mar 28]. Available from: https://www.cancer.gov/types
4. WHO-Cancer [Internet]. [cited 2020 Mar 28]. Available from: http://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf
5. Balmain A. Cancer genetics : from Boveri. Nat Rev Cancer. 2001;1:77–82.
6. Bignold LP, Coghlan BLD, Jersmann HPA. Hansemann, Boveri, chromosomes and the gametogenesis-related theories of tumours. Cell Biol Int. 2006;30:640–644.
7. Hansemann D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und derenbiologische Bedeutung. Arch Pathol Anat. 1890;119:299–326.
8. Boveri T. Uber die konstitution der chromatischen kernsubstanz. Verhandlungen Dtsch Zool Ges. 1903;13.
9. Boveri T. In Zur Frage der Entstehung Maligner Tumoren. Gustav Fish Jena. 1914;1–64.
10. Loeb LA, Harris CC. Advances in chemical carcinogenesis: A historical review and prospective. Cancer Res. 2008;68:6863–6872.
11. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260:170–173.
12. Tabin CJ, Bradley SM, Bargmann CI, Weinberg RA, Papageorge AG, Scolnick EM, et al. Mechanism of activation of a human oncogene. Nature. 1982;300:143–149.
13. Fisher JC. Multiple-mutation theory of carcinogenesis. Nature. 1958;181.
14. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100:57–70.
15. Kurzrock R, Giles FJ. Precision oncology for patients with advanced cancer: the challenges of malignant snowflakes. Cell Cycle. 2015;14:2219–21.
16. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. Elsevier Inc.; 2011;144:646–674.
17. Lazebnik Y. What are the hallmarks of cancer? Nat Rev Cancer. Nature Publishing Group; 2010;10:232–233.
18. Sonnenschein C, Soto AM. The aging of the 2000 and 2011 Hallmarks of Cancer reviews: A critique. J Biosci. 2013;38:651–663.
19. Baker SG. Paradox-Driven Cancer Research. Disruptive Sci Technol. 2013;1:143–148.

20. Sigston EAW, Williams BRG. An emergence framework of carcinogenesis. Front Oncol. 2017;7:1–14.
21. Maley CC, Aktipis A, Graham TA, Sottoriva A, Boddy AM, Janiszewska M, et al. Classifying the evolutionary and ecological features of neoplasms. Nat Rev Cancer. Nature Publishing Group; 2017;17:605–619.
22. Darwin C 1809-1882. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life [Internet]. London : John Murray, 1859; 1859. Available from: https://search.library.wisc.edu/catalog/9934839413602122
23. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194:23–8.
24. Mcgranahan N, Swanton C. Review Clonal Heterogeneity and Tumor Evolution : Past , Present , and the Future. Cell. Elsevier Inc.; 2017;168:613–628.
25. Mazor T, Pankov A, Song JS, Costello JF. Intratumoral Heterogeneity of the Epigenome. Cancer Cell. Elsevier Inc.; 2016;29:440–451.
26. DeVita VT. The "War on Cancer" and its impact. Nat Clin Pract Oncol. 2004;1:55–55.
27. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science. American Association for the Advancement of Science; 1986;231:1055–6.
28. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.
29. Cancer genomics - Latest research and news | Nature [Internet]. [cited 2020 Jun 2]. Available from: https://www.nature.com/subjects/cancer-genomics
30. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. Nature Publishing Group; 2009;458:719–724.
31. Garraway LA, Lander ES. Lessons from the Cancer Genome. Cell. 2013;153:17–37.
32. The Cost of Sequencing a Human Genome [Internet]. Genome.gov. [cited 2020 Jun 16]. Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost
33. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107:1–8.
34. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature. Nature Publishing Group; 1977;265:687–95.
35. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
36. Voelkerding KV, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin Chem.

2009;55:641–58.
37. Mignardi M, Nilsson M. Fourth-generation sequencing in the cell and the clinic. Genome Med. 2014;6:31.
38. Snapshot [Internet]. [cited 2020 Jun 16]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/bioinformatics
39. Richter BG, Sexton DP. Managing and Analyzing Next-Generation Sequence Data. Bourne PE, editor. PLoS Comput Biol. 2009;5:e1000369.
40. Illumina | Sequencing and array-based solutions for genetic research [Internet]. [cited 2020 Jun 18]. Available from: https://www.illumina.com/
41. GATK [Internet]. [cited 2020 Jun 20]. Available from: https://gatk.broadinstitute.org/hc/en-us
42. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6:10001.
43. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18:286.
44. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLOS Comput Biol. Public Library of Science; 2019;15:e1007069.
45. Ye K, Hall G, Ning Z, Trust W, Campus WG. Structural Variation Detection from Next Generation Sequencing Journal of Next Generation Sequencing & Applications.
46. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW, et al. Cancer Genome Landscapes. Science. 2013;339:1546–1558.
47. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990;250:1684–9.
48. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives - Springer. BMC Bioinformatics. 2013;14 Suppl 1:S1.
49. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14:305–14.
50. Kotarbinski T. Hematopoietic Cancers. Genome.gov [Internet]. [cited 2020 Jun 25]. Available from: https://www.genome.gov/genetics-glossary/Polymorphism
51. Cai L, Yuan W, Zhang Z, He L, Chou K-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep. Nature Publishing Group; 2016;6:36540.
52. Wiuf C, Andersen CL. Statistics and Informatics in Molecular Cancer

Research. OUP Oxford; 2009.
53. The Cancer Genome Atlas Program - National Cancer Institute [Internet]. 2018 [cited 2020 Jun 29]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
54. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–339.
55. Herbst RS, Gandara DR, Hirsch FR, Redman MW, LeBlanc M, Mack PC, et al. Lung Master Protocol (Lung-MAP)--A Biomarker-Driven Protocol for Accelerating Development of Therapies for Squamous Cell Lung Cancer: SWOG S1400. Clin Cancer Res. 2015;21:1514–24.
56. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
57. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–1120.
58. The Pan-Cancer Atlas [Internet]. [cited 2020 Jun 29]. Available from: http://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html
59. International Cancer Genome Consortium [Internet]. [cited 2020 Jun 29]. Available from: https://icgc.org/
60. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012;149:994–1007.
61. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93.
62. Rheinbay E, Morten MN, Abascal F, Tiao G, Hornshøj H, Hess JM, et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. 2017;
63. Sabarinathan R, Pich O, Martincorena I, Rubio-Perez C, Juul M, Wala J, et al. The whole-genome panorama of cancer drivers. bioRxiv. 2017;190330.
64. Reference GH. What is the difference between precision medicine and personalized medicine? What about pharmacogenomics? [Internet]. Genet. Home Ref. [cited 2020 Jul 1]. Available from: https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvspersonalized
65. Stratton MR. Exploring the Genomes of Cancer Cells: Progress and Promise. Science. 2011;331:1553–8.
66. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn. 2015;17:251–64.

67. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.
68. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–218.
69. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science. 2006;314:268–74.
70. Forrest WF, Cavet G. Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." Science. American Association for the Advancement of Science; 2007;317:1500–1500.
71. Getz G, Höfling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, et al. Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." Science. American Association for the Advancement of Science; 2007;317:1500–1500.
72. Rubin AF, Green P. Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." Science. American Association for the Advancement of Science; 2007;317:1500–1500.
73. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns Of Selection In Cancer And Somatic Tissues. Cell [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/04/29/132324
74. Weghorn D, Sunyaev S. Bayesian inference of negative and positive selection in human cancers. Nat Genet. Nature Publishing Group; 2017;49:1785–8.
75. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. Oxford University Press; 2012;40:e169.
76. Mularoni L, Sabarinathan R, Deu-pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML : a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. Genome Biology; 2016;1–13.
77. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29:2238–2244.
78. Arnedo-Pac C, Mularoni L, Muinos F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. bioRxiv. 2018;500132.
79. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016;76:3719–31.
80. Martínez-Jiménez F, Muiños F, López-Arribillaga E, Lopez-Bigas N, Gonzalez-Perez A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer.

Nat Cancer. Nature Publishing Group; 2020;1:122–35.
81. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, et al. Identification of cancer driver genes based on nucleotide context. Nat Genet. Nature Publishing Group; 2020;52:208–18.
82. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. Nature Publishing Group; 2020;1–18.
83. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Proto-Oncogenes and Tumor-Suppressor Genes. Mol Cell Biol 4th Ed [Internet]. W. H. Freeman; 2000 [cited 2020 Jul 5]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21662/
84. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. Nature Publishing Group; 2004;4:177–83.
85. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. Nature Publishing Group; 2018;18:696–705.
86. Hainaut P, Pfeifer GP. Patterns of p53 G→T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis. Oxford Academic; 2001;22:367–74.
87. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. Nature Publishing Group; 2002;21:7435–51.
88. Pfeifer GP, You Y-H, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res Mol Mech Mutagen. 2005;571:19–31.
89. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. Nature Publishing Group; 2010;463:191–6.
90. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. Nature Publishing Group; 2010;463:184–90.
91. Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. Galli A, editor. PLOS ONE. 2019;14:e0221235.
92. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell. 2012;149:979–93.
93. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3:246–259.

94. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. Nature Research; 2013;500:415–421.
95. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
96. COSMIC. https://cancer.sanger.ac.uk/cosmic/signatures/SBS/.
97. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. Genome Biology; 2016;17:1–11.
98. Gori K, Baez-Ortega A. sigfit: flexible Bayesian inference of mutational signatures. bioRxiv. Cold Spring Harbor Laboratory; 2020;372896.
99. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. Nature Publishing Group; 2015;47:1402–1407.
100. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. Cell. Elsevier; 2012;150:264–78.
101. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun [Internet]. Springer US; 2019;10. Available from: http://dx.doi.org/10.1038/s41467-019-11037-8
102. Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. N Engl J Med. Massachusetts Medical Society; 2014;371:2477–87.
103. Martincorena I, Roshan A, Gerstung M, Ellis P, Loo PV, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. American Association for the Advancement of Science; 2015;348:880–6.
104. Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. Springer US; 2018;561:473–478.
105. Coorens THH, Treger TD, Al-Saadi R, Moore L, Tran MGB, Mitchell TJ, et al. Embryonal precursors of Wilms tumor. Science. 2019;366:1247–51.
106. Behjati S, Huch M, Boxtel R van, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. Nature Publishing Group; 2014;513:422–5.
107. Blokzijl F, Ligt J de, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. Nature Publishing Group; 2016;538:260–4.

108. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362:911–7.
109. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. ElsevierCompany.; 2018;25:2308–2316.e4.
110. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. Elsevier; 2019;177:821-836.e16.
111. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51:1732–40.
112. Li B, Brady SW, Ma X, Shen S, Zhang Y, Li Y, et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood. 2020;135:41–55.
113. Loeb LA. Mutator Phenotype May Be Required for Multistage Carcinogenesis. Cancer Res. American Association for Cancer Research; 1991;51:3075–9.
114. Loeb LA. Mutator phenotype in cancer: Origin and consequences. Semin Cancer Biol. 2010;20:279–80.
115. Tomlinson IPM, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci. 1996;93:14800–3.
116. Bodmer W, Loeb LA. Genetic Instability Is Not a Requirement for Tumor Development. Cancer Res. 2008;68:3558–61.
117. Campbell PJ, Martinocorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015;349:1483–1489.
118. Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome Biol. 2018;19:67.
119. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat Publ Group. Nature Publishing Group; 2015;47:209–216.
120. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Genet. Nature Publishing Group; 2016;48:1–9.
121. Tarabichi M, Martincorena I, Gerstung M, Leroi AM, Markowetz F, Dentro SC, et al. Neutral tumor evolution? Nat Genet. 2018;50:1630–1633.
122. Heide T, Zapata L, Williams MJ, Werner B, Caravagna G, Barnes CP, et al. Reply to 'Neutral tumor evolution?' Nat Genet. Nature Publishing Group; 2018;50:1633–7.
123. Gupta GP, Massagué J. Cancer Metastasis: Building a Framework. Cell. Elsevier; 2006;127:679–95.
124. Valastyan S, Weinberg RA. Tumor metastasis: Molecular insights and

evolving paradigms. Cell. Elsevier Inc.; 2011;147:275–292.
125. Massagué J, Obenauf AC. Metastatic colonization by circulating tumour cells. Nature. 2016;529:298–306.
126. Kreso A, Dick JE. Evolution of the cancer stem cell model. Cell Stem Cell. Elsevier Inc.; 2014;14:275–291.
127. Avgustinova A, Benitah SA. The epigenetics of tumour initiation: cancer stem cells and their chromatin. Curr Opin Genet Dev. 2016;36:8–15.
128. Campbell LL, Polyak K. Breast Tumor Heterogeneity: Cancer Stem Cells or Clonal Evolution? Cell Cycle. 2007;6:2332–8.
129. Batlle E, Clevers H. Cancer stem cells revisited. Nat Med. 2017;23:1124–34.
130. Jordan CT. Cancer Stem Cells: Controversial or Just Misunderstood? Cell Stem Cell. 2009;4:203–5.
131. Visvader JE, Lindeman GJ. Cancer Stem Cells: Current Status and Evolving Complexities. Cell Stem Cell. 2012;10:717–28.
132. Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J, et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. Nature. 1994;367:645–8.
133. Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat Med. 1997;3:730–7.
134. Notta F, Mullighan CG, Wang JCY, Poeppl A, Doulatov S, Phillips LA, et al. Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. Nature. Nature Publishing Group; 2011;469:362–367.
135. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 2016;22:105–13.
136. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nat Rev Cancer. 2013;13:714–26.
137. Triller N, Korošec P, Kern I, Košnik M, Debeljak A. Multidrug resistance in small cell lung cancer: Expression of P-glycoprotein, multidrug resistance protein 1 and lung resistance protein in chemo-naive patients and in relapsed disease. Lung Cancer. 2006;54:235–40.
138. Doyle LA, Yang W, Abruzzo LV, Krogmann T, Gao Y, Rishi AK, et al. A multidrug resistance transporter from human MCF-7 breast cancer cells. Proc Natl Acad Sci. National Academy of Sciences; 1998;95:15665–70.
139. Kosztyu P, Bukvova R, Dolezel P, Mlejnek P. Resistance to daunorubicin, imatinib, or nilotinib depends on expression levels of ABCB1 and ABCG2 in human leukemia cells. Chem Biol Interact. Elsevier Ireland Ltd; 2014;219:203–210.
140. Steinbach D, Legrand O. ABC transporters and drug resistance in leukemia: Was P-gp nothing but the first head of the Hydra? Leukemia.

2007;21:1172–1176.
141. Gorre ME. Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification. Science. 2001;293:876–80.
142. McMillin DW, Negri JM, Mitsiades CS. The role of tumour–stromal interactions in modifying drug response: challenges and opportunities. Nat Rev Drug Discov. 2013;12:217–28.
143. Zhao J. Cancer stem cells and chemoresistance: The smartest survives the raid. Pharmacol Ther. 2016;160:145–58.
144. Zhang H, Li H, Xi HS, Li S. HIF1␣ is required for survival maintenance of chronic myeloid leukemia stem cells. 2012;119:13.
145. Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F. Cancer Evolution : Mathematical Models and Computational Inference. 2015;64.
146. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet [Internet]. Nature Publishing Group; 2017; Available from: http://www.nature.com/doifinder/10.1038/nrg.2016.170
147. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. PLoS Comput Biol. 2014;10:1–11.
148. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11:396–398.
149. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling: Evolutionary trajectories of ovarian cancers. J Pathol. 2013;231:21–34.
150. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
151. Jiang Y, Qiu Y, Minn AJ, Zhang NR. (Canopy) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci. 2016;201522203.
152. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biol. 2016;17:86.
153. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21:751–759.
154. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I, et al. Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell. 2017;32:169–184.e7.
155. Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harb Perspect Med. 2017;7:a026625.

156. Jolly C, Van Loo P. Timing somatic events in the evolution of cancer. Genome Biol. 2018;19:95.
157. Makishima H, Yoshizato T, Yoshida K, Sekeres MA, Radivoyevitch T, Suzuki H, et al. Dynamics of clonal evolution in myelodysplastic syndromes. Nat Genet [Internet]. 2016; Available from: http://www.nature.com/doifinder/10.1038/ng.3742
158. Colom B, Alcolea MP, Piedrafita G, Hall MWJ, Wabik A, Dentro SC, et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. Nat Genet [Internet]. Springer US; 2020; Available from: http://dx.doi.org/10.1038/s41588-020-0624-3
159. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci Transl Med. 2015;7:283ra54-283ra54.
160. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. Nature Publishing Group; 2020;578:122–8.
161. Hu Z, Li Z, Ma Z, Curtis C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. Nat Genet. 2020;52:701–8.
162. Bispo JAB, Pinheiro PS, Kobetz EK. Epidemiology and Etiology of Leukemia and Lymphoma. Cold Spring Harb Perspect Med. 2020;10:a034819.
163. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127:2391–405.
164. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127:2375–90.
165. Leukaemia (all subtypes combined) survival statistics [Internet]. Cancer Res. UK. 2015 [cited 2020 Aug 29]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia/survival
166. Belson M, Kingsley B, Holmes A. Risk Factors for Acute Leukemia in Children: A Review. Environ Health Perspect. 2007;115:138–45.
167. Greaves M. Leukaemia “firsts” in cancer research and treatment. Nat Rev Cancer. 2016;16:163–172.
168. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. Nat Rev Cancer. Springer US; 2018;1.
169. Kampen KR. The discovery and early understanding of leukemia. Leuk Res. Elsevier Ltd; 2012;36:6–13.
170. Cooper B. The origins of bone marrow as the seedbed of our blood: from antiquity to the time of Osler. Proc Bayl Univ Med Cent.

2011;24:115–8.
171. Nowell PC, Hungerford DA. A minute chromosome in human chronic granulocytic leukemia. Science. 1960;132.
172. Awong G, Zúñiga-Pflücker JC. Development of Human T Lymphocytes. Ref Module Biomed Sci [Internet]. Elsevier; 2014 [cited 2020 Aug 30]. p. B978012801238300115X. Available from: https://linkinghub.elsevier.com/retrieve/pii/B978012801238300115X
173. Carroll WL, Loh M, Biondi A, Willman C. The Biology of Acute Lymphoblastic Leukemia. In: Reaman GH, Smith FO, editors. Child Leuk Prat Handb [Internet]. Berlin and Heidelberg: Springer Berlin Heidelberg; 2011. p. 29–62. Available from: http://www.springerlink.com/index/10.1007/978-3-642-13781-5
174. Marti Cavalheiro L, Strachman Bacal N, Camerão Bento L, Patussi Correia R, Agostini Rocha F. Lymphoid Hematopoiesis and Lymphocytes Differentiation and Maturation. IntechOpen [Internet]. 2017; Available from: https://www.intechopen.com/books/lymphocyte-updates-cancer-autoimmunity-and-infection/lymphoid-hematopoiesis-and-lymphocytes-differentiation-and-maturation
175. Jaffe ES, Harris NL, Stein H, Isaacson PG. Classification of lymphoid neoplasms: the microscope as a tool for disease discovery. Blood. 2008;112:4384–99.
176. Bell JJ, Bhandoola A. The earliest thymic progenitors for T cells possess myeloid lineage potential. Nature. Nature Publishing Group; 2008;452:764–7.
177. Wada H, Masuda K, Satoh R, Kakugawa K, Ikawa T, Katsura Y, et al. Adult T-cell progenitors retain myeloid potential. Nature. Nature Publishing Group; 2008;452:768–72.
178. Belver L, Ferrando A. The genetics and mechanisms of T cell acute lymphoblastic leukaemia. Nat Rev Cancer. Nature Publishing Group; 2016;16:494–507.
179. Terwilliger T, Abdul-Hay M. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer J. 2017;7:577.
180. Pui CH. Acute lymphoblastic leukemia. Child Leuk Third Ed. 2010;332–366.
181. Ding LW, Sun QY, Tan KT, Chien W, Thippeswamy AM, Yeoh AEJ, et al. Mutational landscape of pediatric acute lymphoblastic leukemia. Cancer Res. 2017;77:390–400.
182. Dobson SM, García-Prat L, Vanner RJ, Wintersinger J, Waanders E, Gu Z, et al. Relapse-Fated Latent Diagnosis Subclones in Acute B Lineage Leukemia Are Drug Tolerant and Possess Distinct Metabolic Programs. Cancer Discov. American Association for Cancer Research; 2020;10:568–87.
183. Spinella JF, Richer C, Cassart P, Ouimet M, Healy J, Sinnett D. Mutational dynamics of early and late relapsed childhood ALL: rapid clonal

expansion and long-term dormancy. Blood Adv. 2018;2:177–188.
184. Hoelzer D, Bassan R, Dombret H, Fielding A, Ribera JM, Buske C. Acute lymphoblastic leukaemia in adult patients: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up †. Ann Oncol. Elsevier; 2016;27:v69–82.
185. Feng H, Stachura DL, White RM, Gutierrez A, Zhang L, Sanda T, et al. T-Lymphoblastic Lymphoma Cells Express High Levels of BCL2, S1P1, and ICAM1, Leading to a Blockade of Tumor Cell Intravasation. Cancer Cell. Elsevier; 2010;18:353–66.
186. Hamid GA. Classification of Acute Leukemia, Acute Leukemia - The Scientist's Perspective and Challenge. InTech. 2011;
187. Spinella J-F, Cassart P, Richer C, Saillour V, Ouimet M, Langlois S, et al. Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations. Oncotarget. 2016;7:65485–65503.
188. Omman RA, Kini AR. Acute leukemias. Rodak's Hematol Clin Princ Appl. Sixth Edit. Elsevier Inc.; 2017. p. 540–54.
189. Hamid GA. Acute Leukemia Clinical Presentation. Leukemia [Internet]. IntechOpen; 2013 [cited 2020 Sep 18]; Available from: https://www.intechopen.com/books/leukemia/acute-leukemia-clinical-presentation
190. Bene MC, Castoldi G, Knapp W, Ludwig WD, Matutes E, Orfao A, et al. Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). Leukemia. 1995;9:1783–6.
191. Coustan-Smith E, Mullighan CG, Onciu M, Behm FG, Raimondi SC, Pei D, et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol. Elsevier; 2009;10:147–56.
192. Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC, et al. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell. 2002;1:75–87.
193. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. The Lancet. Elsevier Ltd; 2013;381:1943–1955.
194. Hunger SP, Mullighan CG. Redefining ALL classification : toward detecting high-risk ALL and implementing precision medicine. Blood. 2015;125:3977–3988.
195. Mullighan CG. The genomic landscape of acute lymphoblastic leukemia in children and young adults. Hematology. 2014;2014:174–180.
196. Paulsson K, Lilljebjörn H, Biloglav A, Olsson L, Rissler M, Castor A, et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2015;47:672–677.
197. Bateman CM, Alpar D, Ford AM, Colman SM, Wren D, Morgan M, et al. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins.

Leukemia. 2015;29:58–65.
198. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:242–52.
199. Daniel MG, Rapp K, Schaniel C, Moore KA. Induction of developmental hematopoiesis mediated by transcription factors and the hematopoietic microenvironment. Ann N Y Acad Sci. 2020;1466:59–72.
200. Alpar D, Wren D, Ermini L, Mansur MB, van Delft FW, Bateman CM, et al. Clonal origins of ETV6-RUNX1+ acute lymphoblastic leukemia: studies in monozygotic twins. Leukemia. 2015;29:839–46.
201. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2014;46:116–25.
202. Iacobucci I, Mullighan CG. Genetic basis of acute lymphoblastic leukemia. J Clin Oncol. 2017;35:975–983.
203. Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J, et al. BCR–ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. Nature. 2008;453:110–4.
204. Boer MLD, Slegtenhorst M van, Menezes RXD, Cheok MH, Buijs-Gladdines JG, Peters ST, et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol. Elsevier; 2009;10:125–34.
205. Jain N, Roberts KG, Jabbour E, Patel K, Eterovic AK, Chen K, et al. Ph-like acute lymphoblastic leukemia: a high-risk subtype in adults. Blood. 2017;129:572–81.
206. Iacobucci I, Papayannidis C, Lonetti A, Ferrari A, Baccarani M, Martinelli G. Cytogenetic and Molecular Predictors of Outcome in Acute Lymphocytic Leukemia: Recent Developments. Curr Hematol Malig Rep. 2012;7:133–43.
207. Lilljebjörn H, Fioretos T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. Blood. 2017;130:1395–401.
208. Zhang J, Mccastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet. 2016;48.
209. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet. 2019;51:296–307.
210. Moorman AV. New and emerging prognostic and predictive genetic biomarkers in B-cell precursor acute lymphoblastic leukemia. Haematologica. 2016;101:407–16.
211. Gu Z, Churchman M, Roberts K, Li Y, Liu Y, Harvey RC, et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. Nat Commun. 2016;7:13331.

212. Hirabayashi S, Ohki K, Nakabayashi K, Ichikawa H, Momozawa Y, Okamura K, et al. ZNF384-related fusion genes define a subgroup of childhood B-cell precursor acute lymphoblastic leukemia with a characteristic immunotype. Haematologica. 2017;102:118–29.
213. Ferrando AA, Look AT. Gene expression profiling in T-cell acute lymphoblastic leukemia. Semin Hematol. 2003;40:274–80.
214. Girardi T, Vicente C, Cools J, De Keersmaecker K. The genetics and molecular biology of T-ALL. Blood. 2017;129:1113–1123.
215. Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic leukemia. J Clin Invest. 2012;122:3398–3406.
216. on behalf of the GFCH (Groupe Francophone de Cytogénétique Hématologique) and the BCGHO (Belgian Cytogenetic Group for Hematology and Oncology), Graux C, Stevens-Kroef M, Lafage M, Dastugue N, Harrison CJ, et al. Heterogeneous patterns of amplification of the NUP214-ABL1 fusion gene in T-cell acute lymphoblastic leukemia. Leukemia. 2009;23:125–33.
217. Homminga I, Pieters R, Langerak AW, de Rooi JJ, Stubbs A, Verstegen M, et al. Integrated Transcript and Genome Analyses Reveal NKX2-1 and MEF2C as Potential Oncogenes in T Cell Acute Lymphoblastic Leukemia. Cancer Cell. Elsevier Inc.; 2011;19:484–497.
218. Colomer-Lahiguera S, Pisecker M, König M, Nebral K, Pickl WF, Kauer MO, et al. MEF2C-dysregulated pediatric T-cell acute lymphoblastic leukemia is associated with *CDKN1B* deletions and a poor response to glucocorticoid therapy. Leuk Lymphoma. 2017;58:2895–904.
219. Luskin MR, DeAngelo DJ. T-cell acute lymphoblastic leukemia: Current approach and future directions. Adv CELL GENE Ther [Internet]. 2019 [cited 2020 Sep 25];2. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acg2.70
220. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LAA, Miller CB, et al. Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. 2009;
221. Perez-Andreu V, Roberts KG, Xu H, Smith C, Zhang H, Yang W, et al. A genome-wide association study of susceptibility to acute lymphoblastic leukemia in adolescents and young adults. 2015;125:7.
222. Tijchon E, Havinga J, van Leeuwen FN, Scheijen B. B-lineage transcription factors and cooperating gene lesions required for leukemia development. Leukemia. 2013;27:541–52.
223. Shah S, Schrader KA, Waanders E, Timms AE, Vijai J, Miething C, et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:1226–1231.
224. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature. 2007;446:758–64.

225. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. Nat Rev Cancer. 2003;3:639–49.
226. Hiebert SW, Sun W, Davis JN, Golub T, Shurtleff S, Buijs A, et al. The t(12;21) translocation converts AML-1B from an activator to a repressor of transcription. Mol Cell Biol. 1996;16:1349–55.
227. Zhang J, Mullighan CG, Harvey RC, Wu G, Chen X, Edmonson M, et al. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood. 2011;118:3080–7.
228. Lilljebjörn H, Henningsson R, Hyrenius-Wittsten A, Olsson L, Orsmark-Pietras C, von Palffy S, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. Nat Commun. 2016;7:11790.
229. Guo X, Zhang R, Liu J, Li M, Song C, Dovat S, et al. Characterization of LEF1 High Expression and Novel Mutations in Adult Acute Lymphoblastic Leukemia. Bandapalli OR, editor. PLOS ONE. 2015;10:e0125429.
230. Hof J, Krentz S, Van Schewick C, Körner G, Shalapour S, Rhein P, et al. Mutations and deletions of the TP53 gene predict nonresponse to treatment and poor outcome in first relapse of childhood acute lymphoblastic leukemia. J Clin Oncol. 2011;29:3185–3193.
231. Russell LJ, Capasso M, Vater I, Akasaka T, Bernard OA, Calasanz MJ, et al. Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia. Blood. 2009;114:2688–98.
232. Vainchenker W, Constantinescu SN. JAK/STAT signaling in hematological malignancies. Oncogene. 2013;32:2601–13.
233. Tasian SK, Loh ML, Hunger SP. Philadelphia chromosome–like acute lymphoblastic leukemia. 2017;130:9.
234. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature [Internet]. Nature Publishing Group; 2018; Available from: http://www.nature.com/doifinder/10.1038/nature25795
235. Li J-F, Dai Y-T, Lilljebjörn H, Shen S-H, Cui B-W, Bai L, et al. Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. Proc Natl Acad Sci. 2018;115:E11711–20.
236. Mullighan CG, Downing JR. Global Genomic Characterization of Acute Lymphoblastic. Semin Hematol. 2009;46:3–15.
237. Weng AP, Ferrando AA, Lee W, Iv JPM, Silverman LB, Sanchez-irizarry C, et al. Activating Mutations of NOTCH1 in Human T Cell Acute Lymphoblastic Leukemia. 2004;306:269–272.
238. Neumann M, Vosberg S, Schlee C, Heesch S, Schwartz S, Gökbuget

N, et al. Mutational spectrum of adult T-ALL. Oncotarget. 2015;6:2754–2766.
239. Bigas A, Robert-Moreno À, Espinosa L. The Notch pathway in the developing hematopoietic system. Int J Dev Biol. 2010;54:1175–1188.
240. Tosello V, Ferrando AA. The NOTCH signaling pathway: role in the pathogenesis of T-cell acute lymphoblastic leukemia and implication for therapy. Ther Adv Hematol. 2013;4:199–210.
241. Clappier E, Collette S, Grardel N, Girard S, Suarez L, Brunie G, et al. NOTCH1 and FBXW7 mutations have a favorable impact on early response to treatment, but not on outcome, in children with T-cell acute lymphoblastic leukemia (T-ALL) treated on EORTC trials 58881 and 58951. Leukemia. 2010;24:2023–2031.
242. Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat Genet. 2017;49:1211–1218.
243. Lahortiga I, De Keersmaecker K, Van Vlierberghe P, Graux C, Cauwelier B, Lambert F, et al. Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. Nat Genet. 2007;39:593–595.
244. De Keersmaecker K, Ferrando AA. TLX1-induced T-cell acute lymphoblastic leukemia. Clin Cancer Res. 2011;17:6381–6386.
245. Wendorff AA, Quinn SA, Rashkovan M, Madubata CJ, Ambesi-Impiombato A, Litzow MR, et al. Phf6 loss enhances HSC self-renewal driving tumor initiation and leukemia stem cell activity in T-All. Cancer Discov. 2019;9:436–451.
246. Neumann M, Heesch S, Schlee C, Schwartz S, Gökbuget N, Hoelzer D, et al. Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. Blood. 2013;121:4749–4752.
247. Bowman RL, Busque L, Levine RL. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell. 2018;22:157–70.
248. Raetz EA, Teachey DT. T-cell acute lymphoblastic leukemia. Hematol Am Soc Hematol Educ Program. 2016;580–8.
249. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:186–90.
250. Li Z, Abraham BJ, Berezovskaya A, Farah N, Liu Y, Leon T, et al. APOBEC signature mutation generates an oncogenic enhancer that drives LMO1 expression in T-ALL. Leukemia. 2017;31:2057–2064.
251. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med. 2015;373:2336–2346.
252. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood

cancers. Nature. 2018;555:321–327.
253. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009;41:1001–5.
254. Waanders E, Gu Z, Dobson SM, Antić Ž, Crawford JC, Ma X, et al. Mutational Landscape and Patterns of Clonal Evolution in Relapsed Pediatric Acute Lymphoblastic Leukemia. Blood Cancer Discov. 2020;
255. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, et al. Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. Science. 2008;322:1377–1380.
256. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, Rusch M, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun. Nature Publishing Group; 2015;6:1–12.
257. Oshima K, Khiabanian H, da Silva-Almeida AC, Tzoneva G, Abate F, Ambesi-Impiombato A, et al. Mutational landscape, clonal evolution patterns, and role of RAS mutations in relapsed acute lymphoblastic leukemia. Proc Natl Acad Sci U S A. National Academy of Sciences; 2016;113:11306–11311.
258. Ferrando AA, López-Otín C. Clonal evolution in leukemia. Nat Med. Nature Publishing Group; 2017;23:1135–1145.
259. Eguchi-Ishimae M, Eguchi M, Kempski H, Greaves M. NOTCH1 mutation can be an early, prenatal genetic event in T-ALL. Blood. 2008;111:376–8.
260. Rampersaud E, Ziegler DS, Iacobucci I, Payne-Turner D, Churchman ML, Schrader KA, et al. Germline deletion of ETV6 in familial acute lymphoblastic leukemia. Blood Adv. 2019;3:1039–46.
261. Bie JD, Alberti-servera L, Geerdens E, Segers H, Broux M, Keersmaecker KD, et al. Single-cell sequencing reveals the origin and the order of mutation acquisition in T-cell acute lymphoblastic leukemia. 2018;1358–1369.
262. Mansour MR, Duke V, Foroni L, Patel B, Allen CG, Ancliff PJ, et al. NOTCH1 mutations are secondary events in some patients with T-cell acute lymphoblastic leukemia. Clin Cancer Res. 2007;13:6964–6969.
263. Potter N, Jones L, Blair H, Strehl S, Harrison CJ, Greaves M, et al. Single-cell analysis identifies CRLF2 rearrangements as both early and late events in Down syndrome and non-Down syndrome acute lymphoblastic leukaemia. Leukemia. 2019;33:893–904.
264. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci. 2014;111:17947–17952.
265. Bhatla T, Jones CL, Meyer JA, Vitanza NA, Raetz EA, Carroll WL. The Biology of Relapsed Acute Lymphoblastic Leukemia. J Pediatr Hematol Oncol. NIH Public Access; 2014;36:413–418.

266. Yang J, Bhojwani D, Yang W. Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia. \ldots. 2008;112:4178–4183.
267. Schroeder MP, Bastian L, Eckert C, Gökbuget N, James AR, Tanchez JO, et al. Integrated analysis of relapsed B-cell precursor Acute Lymphoblastic Leukemia identifies subtype-specific cytokine and metabolic signatures. Sci Rep. 2019;9:1–11.
268. Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat Med. Nature Publishing Group; 2013;19:368–71.
269. Meyer JA, Wang J, Hogan LE, Yang JJ, Dandekar S, Patel JP, et al. Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet. Nature Publishing Group; 2013;45:290–294.
270. Li B, Li H, Bai Y, Kirschner-Schwabe R, Yang JJ, Chen Y, et al. Negative feedback–defective PRPS1 mutants drive thiopurine resistance in relapsed childhood ALL. Nat Med. 2015;21:563–71.
271. Xiao H, Wang LM, Luo Y, Lai X, Li C, Shi J, et al. Mutations in epigenetic regulators are involved in acute lymphoblastic leukemia relapse following allogeneic hematopoietic cell transplantation. Oncotarget. 2015;7.
272. Mullighan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. NIH Public Access; 2011;471:235–9.
273. Evensen NA, Madhusoodhan PP, Meyer J, Saliba J, Chowdhury A, Araten DJ, et al. MSH6 haploinsufficiency at relapse contributes to the development of thiopurine resistance in pediatric B-lymphoblastic leukemia. Haematologica. 2018;103:830–9.
274. Anderson K, Lutz C, van Delft FW, Bateman CM, Guo Y, Colman SM, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature. Nature Research; 2011;469:356–361.
275. Wynn R, Bhat R, Monagle P. Acute Lymphoblastic Leukemia. Pediatr Hematol Pract Guide [Internet]. Cambridge: Cambridge University Press; 2017 [cited 2020 Sep 27]. Available from: http://ebooks.cambridge.org/ref/id/CBO9781139942430
276. Malard F, Mohty M. Acute lymphoblastic leukaemia. 2020;395:17.
277. Forero-Castro M, Robledo C, Benito R, Bodega-Mayor I, Rapado I, Hernández-Sánchez M, et al. Mutations in TP53 and JAK2 are independent prognostic biomarkers in B-cell precursor acute lymphoblastic leukaemia. Br J Cancer. 2017;1–10.
278. Carrasco Salas P, Fernández L, Vela M, Bueno D, González B, Valentín J, et al. The role of CDKN2A/B deletions in pediatric acute lymphoblastic leukemia. Pediatr Hematol Oncol. 2016;33:415–22.
279. Zhang W, Kuang P, Liu T. Prognostic significance of CDKN2A/B

deletions in acute lymphoblastic leukaemia: a meta-analysis. Ann Med. 2019;51:28–40.
280. Childhood Acute Lymphoblastic Leukemia Treatment (PDQ®)–Patient Version - National Cancer Institute [Internet]. 2020 [cited 2020 Sep 28]. Available from: https://www.cancer.gov/types/leukemia/patient/child-all-treatment-pdq
281. Pui CH, Pei D, Coustan-Smith E, Jeha S, Cheng C, Bowman WP, et al. Clinical utility of sequential minimal residual disease measurements in the context of risk-based therapy in childhood acute lymphoblastic leukaemia: A prospective study. Lancet Oncol. 2015;16:465–474.
282. Litzow MR, Ferrando AA. How I treat T-cell acute lymphoblastic leukemia in adults. Blood. 2015;833–841.
283. Samra B, Jabbour E, Ravandi F, Kantarjian H, Short NJ. Evolving therapy of adult acute lymphoblastic leukemia: state-of-the-art treatment and future directions. J Hematol OncolJ Hematol Oncol. 2020;13:70.
284. Aldoss IT, Marcucci G, Vinod Pullarkat. Treatment of Acute Lymphoblastic Leukemia in Adults: Applying Lessons Learned in Children [Internet]. Cancer Netw. 2016 [cited 2020 Sep 29]. Available from: https://www.cancernetwork.com/view/treatment-acute-lymphoblastic-leukemia-adults-applying-lessons-learned-children
285. Dieck CL, Tzoneva G, Forouhar F, Carpenter Z, Ambesi-Impiombato A, Sánchez-Martín M, et al. Structure and Mechanisms of NT5C2 Mutations Driving Thiopurine Resistance in Relapsed Lymphoblastic Leukemia. 2018;19.
286. Tzoneva G, Dieck CL, Oshima K, Ambesi-Impiombato A, Sánchez-Martín M, Madubata CJ, et al. Clonal evolution mechanisms in NT5C2 mutant-relapsed acute lymphoblastic leukaemia. Nature. Nature Publishing Group; 2018;553:511–514.
287. Follini E, Marchesini M, Roti G. Strategies to Overcome Resistance Mechanisms in T-Cell Acute Lymphoblastic Leukemia. Int J Mol Sci. 2019;20:3021.
288. Jing D, Huang Y, Liu X, Sia KCS, Zhang JC, Tai X, et al. Lymphocyte-Specific Chromatin Accessibility Pre-determines Glucocorticoid Resistance in Acute Lymphoblastic Leukemia. Cancer Cell. 2018;34:906-921.e8.
289. Delgado-Martin C, Meyer LK, Huang BJ, Shimano KA, Zinter MS, Nguyen JV, et al. JAK/STAT pathway inhibition overcomes IL7-induced glucocorticoid resistance in a subset of human T-cell acute lymphoblastic leukemias. Leukemia. 2017;31:2568–76.
290. Ankathil R. ABCB1 genetic variants in leukemias: current insights into treatment outcomes. Pharmacogenomics Pers Med. Dove Press; 2017;Volume 10:169–181.
291. Demir S, Boldrin E, Sun Q, Hampp S, Tausch E, Eckert C, et al. Therapeutic targeting of mutant p53 in pediatric acute lymphoblastic

leukemia. Haematologica. 2020;105:170–81.
292. Ariës IM, Bodaar K, Karim SA, Chonghaile TN, Hinze L, Burns MA, et al. PRC2 loss induces chemoresistance by repressing apoptosis in T cell acute lymphoblastic leukemia. J Exp Med. 2018;215:3094–114.
293. Meyer LK, Hermiston ML. The bone marrow microenvironment as a mediator of chemoresistance in acute lymphoblastic leukemia. Cancer Drug Resist [Internet]. 2019 [cited 2020 Sep 30]; Available from: https://cdrjournal.com/article/view/3233
294. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Research. 2020;9:63.
295. gerstung-lab/clonex [Internet]. Cancer Data Science; 2019 [cited 2020 Oct 5]. Available from: https://github.com/gerstung-lab/clonex
296. Oriol A, Vives S, Hernández-Rivas JM, Tormo M, Heras I, Rivas C, et al. Outcome after relapse of acute lymphoblastic leukemia in adult patients included in four consecutive risk-adapted trials by the PETHEMA study group. Haematologica. 2010;95:589–596.
297. Kunz JB, Rausch T, Bandapalli OR, Eilers J, Pechanska P, Schuessele S, et al. Pediatric T-cell lymphoblastic leukemia evolves into relapse by clonal selection, acquisition of mutations and promoter hypomethylation. Haematologica. 2015;100:1442–1450.
298. Meleveedu KS, Litzow M. Advances in measurable residual disease monitoring for adult acute lymphoblastic leukemia. Adv CELL GENE Ther [Internet]. 2019 [cited 2020 Oct 8];2. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/acg2.67
299. Asnafi V, Buzyn A, Le Noir S, Baleydier F, Simon A, Beldjord K, et al. NOTCH1/FBXW7 mutation identifies a large subgroup with favorable outcome in adult T-cell acute lymphoblastic leukemia (T-ALL): A Group for Research on Adult Acute Lymphoblastic Leukemia (GRAALL) study. Blood. 2009;113:3918–3924.
300. Jenkinson S, Koo K, Mansour MR, Goulden N, Vora A, Mitchell C, et al. Impact of NOTCH1/FBXW7 mutations on outcome in pediatric T-cell acute lymphoblastic leukemia patients treated on the MRC UKALL 2003 trial. Leukemia. 2013;27:41–7.
301. Ferrando A. NOTCH mutations as prognostic markers in T-ALL. Leuk Off J Leuk Soc Am Leuk Res Fund UK. Nature Publishing Group; 2010;24:2003–2004.
302. Saito Y, Koya J, Araki M, Kogure Y, Shingaki S, Tabata M, et al. Landscape and function of multiple mutations within individual oncogenes. Nature [Internet]. 2020 [cited 2020 May 26]; Available from: http://www.nature.com/articles/s41586-020-2175-2
303. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science. 2020;367:1449–1454.

7. APPENDIX

7.1 Collaboration

I have also been involved in the Liver Cancer Evolution Consortium (LCE). This is a collaboration of some researchers in the labs of Dr. Martin S. Taylor (Edinburgh University), Dr. Duncan T. Odom (DKFZ), Dr. Paul Flicek (EBI), Dr. Núria López-Bigas and Dr. Colin S Semple (Edinburgh University). The aim of the consortium was to shed some light into the mutagenesis of DEN-induced mouse liver tumors and fully understand the progression of hepatocellular carcinomas of a mouse model to get insights into the human counterpart. The LCE sequenced 371 whole-genomes from liver tumors from DEN-induced C3H and CAST mouse strains. Together with Claudia Arnedo-Pac and Oriol Pich, we have searched for driver mutations in coding and non-coding regions.

The first study of the consortium was published in Nature this year.

Aitken, S.J., Anderson, C.J., Connor, F. et al. Pervasive lesion segregation shapes cancer genome evolution. Nature 583, 265–270 (2020). https://doi.org/10.1038/s41586-020-2435-1

Article

Pervasive lesion segregation shapes cancer genome evolution

https://doi.org/10.1038/s41586-020-2435-1

Received: 6 December 2019

Accepted: 7 May 2020

Published online: 24 June 2020

Check for updates

Sarah J. Aitken[1,2,3], Craig J. Anderson[4,13], Frances Connor[1,13], Oriol Pich[5], Vasavi Sundaram[1,6], Christine Feig[1], Tim F. Rayner[1], Margus Lukk[1], Stuart Aitken[4], Juliet Luft[4], Elissavet Kentepozidou[6], Claudia Arnedo-Pac[5], Sjoerd V. Beentjes[7], Susan E. Davies[3], Ruben M. Drews[1], Ailith Ewing[4], Vera B. Kaiser[4], Ava Khamseh[4,8], Erika López-Arribillaga[5], Aisling M. Redmond[1], Javier Santoyo-Lopez[9], Inés Sentís[5], Lana Talmane[4], Andrew D. Yates[6], Liver Cancer Evolution Consortium*, Colin A. Semple[4], Núria López-Bigas[5,10,11], Paul Flicek[1,6], Duncan T. Odom[1,12] ✉ & Martin S. Taylor[4] ✉

Cancers arise through the acquisition of oncogenic mutations and grow by clonal expansion[1,2]. Here we reveal that most mutagenic DNA lesions are not resolved into a mutated DNA base pair within a single cell cycle. Instead, DNA lesions segregate, unrepaired, into daughter cells for multiple cell generations, resulting in the chromosome-scale phasing of subsequent mutations. We characterize this process in mutagen-induced mouse liver tumours and show that DNA replication across persisting lesions can produce multiple alternative alleles in successive cell divisions, thereby generating both multiallelic and combinatorial genetic diversity. The phasing of lesions enables accurate measurement of strand-biased repair processes, quantification of oncogenic selection and fine mapping of sister-chromatid-exchange events. Finally, we demonstrate that lesion segregation is a unifying property of exogenous mutagens, including UV light and chemotherapy agents in human cells and tumours, which has profound implications for the evolution and adaptation of cancer genomes.

Analysis of cancer genomes has led to the identification of many driver mutations and mutation signatures[1,3] that illustrate how environmental mutagens cause genetic damage and increase cancer risk[4,5]. The numerous patterns of mutations identified in cancer genomes reflects the temporal and spatial heterogeneity of exogenous and endogenous exposures, mutational processes and germline variation among patients. A study of diverse human cancers identified 49 distinct single-base-substitution signatures, with almost all tumours showing evidence of at least three such signatures[3].

This intrinsic heterogeneity leads to overlapping mutation signatures that make it difficult to accurately disentangle the biases of DNA damage and repair, or to interpret the dynamics of clonal evolution. We reasoned that a more controlled and genetically uniform cancer model system would overcome some of these limitations. By effectively re-running cancer evolution hundreds of times, we aimed to explore oncogenesis and mutation patterns at high resolution and with good statistical power.

We chemically induced liver tumours in postnatal day 15 (P15) male C3H/HeOuJ inbred mice (hereafter referred to as C3H mice) (Fig. 1a; $n = 104$) using a single dose of diethylnitrosamine (DEN)[6]. For comparison and validation, we replicated the study in the divergent mouse strain CAST/EiJ[7] (hereafter referred to as CAST mice) (Extended Data Fig. 1; $n = 54$).

Whole-genome sequencing (WGS) of 371 independently-evolved tumours from 104 C3H mice (Supplementary Table 1) revealed that each genome had about 60,000 (approximately 13 per Mb) somatic point mutations (Extended Data Fig. 1a), a similar level to that found in human cancers caused by exogenous mutagens such as tobacco[8] and UV exposure[9]. Insertion–deletion mutations and larger segmental changes were rare (Extended Data Fig. 1a–f). Point mutations were predominantly (76%) T→N or their complement A→N changes (where N represents any other nucleotide; Fig. 1b, Extended Data Fig. 1g–j), consistent with the long-lived thymine adduct O^4-ethyl-deoxythymidine being the principal mutagenic lesion[10]. Known driver mutations were in the EGFR–RAS–RAF pathway[6,11,12] (Fig. 1c) and usually mutually exclusive. Similar results were replicated in CAST mice (Extended Data Fig. 1j).

Chromosome-scale segregation of lesions

In each tumour, we observed multimegabase genomic segments with pronounced Watson-versus-Crick-strand asymmetry of mutations, frequently encompassing entire chromosomes (Fig. 2). We define

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [2]Department of Pathology, University of Cambridge, Cambridge, UK. [3]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [4]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [7]School of Mathematics and Maxwell Institute, University of Edinburgh, Edinburgh, UK. [8]Higgs Centre for Theoretical Physics, University of Edinburgh, Edinburgh, UK. [9]Edinburgh Genomics (Clinical), The University of Edinburgh, Edinburgh, UK. [10]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [12]German Cancer Research Center (DKFZ), Division of Regulatory Genomics and Cancer Evolution, Heidelberg, Germany. [13]These authors contributed equally: Craig J. Anderson, Frances Connor. *A list of members and their affiliations appears at the end of the paper. ✉e-mail: Duncan.Odom@cruk.cam.ac.uk; martin.taylor@igmm.ed.ac.uk

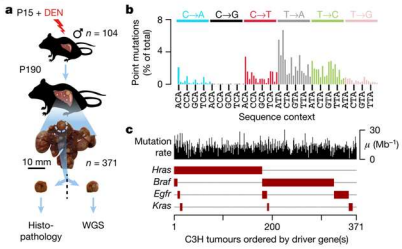


Fig. 1 | DEN-initiated tumours have a high burden of T→N and A→N mutations and driver changes in the EGFR–RAS–RAF pathway. a, P15 male C3H mice received a single dose of DEN; 371 tumours were isolated 25 weeks later (P190) and analysed by histopathology and WGS. **b**, Aggregated mutations showing the distribution of nucleotide substitutions; every fourth trinucleotide context is displayed (*x*-axis). **c**, Each tumour is shown as a column with its mutation rate (μ) per million base pairs (Mb) (black) and driver mutations (brown boxes).

Watson-strand bias as an excess of T→N over A→N mutations when called on the forward strand of the reference genome, and Crick-strand bias as the converse of this. With a median span of 55 Mb (Fig. 2a–d), these asymmetrically mutated segments are orders of magnitude longer than asymmetries generated by transcription-coupled nucleotide-excision repair (TCR)[13], APOBEC mutagenesis[14,15] or replication biases[13,16]. Total mutation load and DNA copy number remain uniform across the genome (Fig. 2e, f).

Pervasive, strand-asymmetric mutagenesis can be explained by DEN-induced lesions remaining unrepaired before genome replication. The first round of replication after DEN treatment results in two sister chromatids with independent lesions on each parent strand, and daughter strands containing misincorporation errors complementary to the lesions (Fig. 2i). The sister chromatids segregate into separate daughter cells during mitosis, and lesion–mutation duplexes are resolved into a mutated DNA base pair by later replication cycles. Asymmetric regions show a 23-fold excess (median) of their preferred mutation over its reverse complement, thus more than 95% of lesions that generate a mutation segregate for at least one mitotic division. We subsequently refer to this phenomenon as 'lesion segregation'.

The haploid X chromosome always contains segments with a strong strand bias (Fig. 2g). On autosomal chromosomes, when both allelic copies have the same bias, the genome shows that bias (for example, Watson bias on chromosome 15 in Fig. 2a–d); when one copy has Watson bias and the other has Crick bias, the chromosome appears unbiased (for example, chromosome 19 in Fig. 2a–d). A model based on random retention of Watson- or Crick-biased chromosomes accurately predicts that (1) around 50% of the autosomal genome and (2) 100% of the haploid X chromosome show mutational asymmetry (Fig. 2g, Extended Data Fig. 2). A few tumours (3.5%) have absent or muted asymmetry; cellularity estimates indicate that they are polyclonal or polyploid (Supplementary Table 1).

Resolving sister-chromatid exchange

The lesion segregation model predicts that mutational asymmetries should span whole chromosomes. However, we observe symmetry switches between multimegabase segments of Watson and Crick bias within chromosomes (Fig. 2a–d, g). These probably represent sister-chromatid exchanges (SCEs) from homologous-recombination-mediated DNA repair[17] (Extended Data Fig. 4a). SCEs are typically invisible to sequencing technologies because homologous recombination between sister chromatids is thought to be error-free[18].

SCE frequency per tumour positively correlates with point mutation rate (Extended Data Fig. 3a, b). With about 27 SCEs (median) in each tumour genome ($n = 371$), we had sufficient statistical power to detect recurrent exchange sites and biases in genomic context (Extended Data Fig. 3c, d). After removing three reference-genome misassemblies (Fig. 2g, Extended Data Fig. 3e, f), we found that SCEs occur with modest enrichment in transcriptionally inactive, late-replicating regions (Extended Data Fig. 4b). The fine mapping (approximately 20-kb resolution) of SCEs enabled us to test the fidelity of homologous recombination. The mutation rate appears locally elevated at SCEs, but the mutational spectrum matches the rest of the genome (Extended Data Fig. 4c–f). A model of Holliday-intermediate branch migration could explain these observations (Extended Data Fig. 4g).

Lesion segregation reveals selection

Cumulatively the tumours have equal Watson and Crick lesion-strand retention across most of the genome (Fig. 2h). However, we observe striking deviations at loci spanning known driver genes (Fig. 2h). The T→A mutation at codon 584 of the *Braf* driver gene[6] is observed in 153 out of 371 tumours in C3H mice, and we would expect the surrounding chromosomal segment to retain T lesions on the same strand. This is the case in 94% (144 out of 153) of tumours (Fisher's exact test, $P = 3.6 \times 10^{-19}$). By contrast, tumours lacking the *Braf* mutation do not show a retention bias (47% Crick, 53% Watson; $P = 0.88$, not rejecting the 50:50 null expectation). We applied this test for oncogenic selection at sites with sufficient recurrent mutations to have statistical power, which confirmed that there was significant oncogenic selection in *Hras*, *Braf* and *Egfr* (Fig. 1c, Extended Data Table 1).

DNA repair with lesion-strand resolution

Resolving DNA lesions to specific strands within a single cell cycle presents a unique opportunity to investigate strand-specific DNA damage and repair in vivo. For example, TCR (Fig. 3a) specifically removes DNA lesions from the RNA template strand[19,20].

We generated transcriptomes from the tissue of origin at the developmental time of DEN mutagenesis. Mutation rates were calculated for each gene in each tumour, stratified by both expression level and the strand containing lesions (Fig. 3b). As expected, TCR was highly specific to the template strand and correlated closely with gene expression. The mutation rate in non-expressed genes had no observable transcription-strand bias. By contrast, mutations in highly expressed genes were reduced by 79.8 ± 1.0% (mean ± s.d.) if the tumour had template-strand lesions.

To evaluate the specificity of TCR, we compared mutation rates for each trinucleotide context between template and non-template strands, stratified by expression level (Fig. 3c). For highly expressed genes, thymines have an 82 ± 6.8% (mean ± s.d. across sequence contexts) lower mutation rate on the template strand; the non-template mutation rate is indifferent to expression (Fig. 3c, dark blue lines are close to vertical), as expected[19]. Mutations from C and G show highly efficient TCR on the template strand; 70 ± 7.8% and 34 ± 21%, respectively. In contrast to T mutations, they also show an expression-dependent reduction in mutation rate on the non-template strand, suggesting that non-TCR repair processes are active. Rare mutations from adenine on the lesion-containing strand increase with transcription, possibly owing to activity of error-prone *trans*-lesion DNA polymerase Pol-η[21].

The ability to resolve the lesion strand unmasks the contribution of bidirectional transcription from active promoters[22] in shaping mutation patterns (Fig. 3d–f, Extended Data Fig. 5). Genic transcription is associated with a sharp, sustained reduction in mutation rate from template-strand lesions. A local increase in the mutation rate over

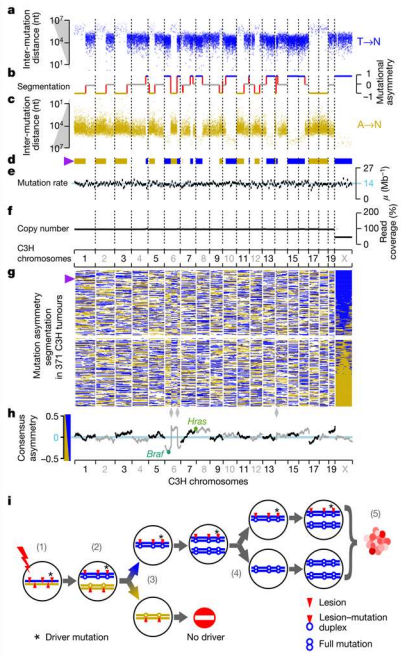


Fig. 2 | Chromosome-scale and strand asymmetric segregation of DNA lesions. a–**f**, An example DEN-induced C3H tumour (identifier: 94315_N8). **a**–**c**, Mutational asymmetry. Individual T→N mutations shown as blue (**a**; Watson strand) and gold (**c**; Crick strand) points; the *y*-axis shows distance to the nearest same-strand T→N mutation; nt, nucleotide. **b**, Segmentation of mutation strand asymmetry patterns. The *y*-axis shows degree of asymmetry (grey indicates no bias). Red indicates symmetry switches. **d**, Asymmetric segments shown as ribbon plot. **e**, Mutation rate in 10-Mb windows; blue line shows genome-wide average. **f**, DNA copy number in 10-Mb windows (grey) and for each asymmetric segment (black). **g**, Ribbon plots (as in **d**) for 371 C3H tumours ranked by X chromosome asymmetry. Purple triangle indicates the example tumour depicted in **a**–**f**. Grey diamonds mark reference genome misassemblies. **h**, Driver genes distort the balance of Watson and Crick asymmetries (see Methods). **i**, Mechanistic model of lesion segregation. (1) A mutagen generates lesions (red triangles) on both DNA strands. (2) If not removed, lesions will segregate into sister chromatids: one carrying only Watson-strand lesions (blue) and the second carrying only Crick-strand lesions (gold). Postmitotic daughter cells will have independent lesions and resulting replication errors (3), resolved into full mutations in later replication (4). (5) Only lineages containing driver changes (* in (1)) will expand into substantial populations.

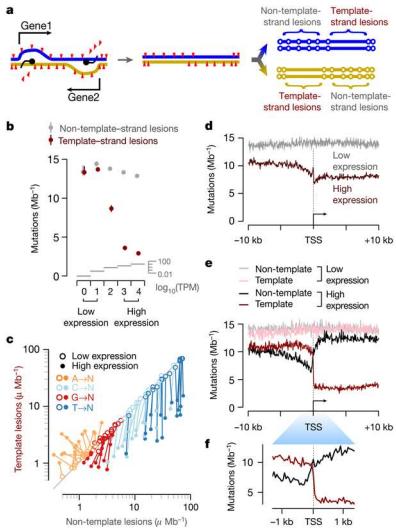


Fig. 3 | Identification of the lesion-containing DNA strand enables TCR to be quantified with strand specificity. a, TCR of DNA lesions is expected to reduce the mutation rate only when lesions are on the template strand of an expressed gene. Black spot shows RNA polymerase and black line the nascent RNA transcript. **b**, TCR of template-strand lesions is dependent on transcription level (P15 liver, median transcripts per million (TPM)). Estimates of mutation rate (circles) are the aggregate rates for expression level of binned genes across C3H tumours ($n = 371$). Expression level bin 0 contains $n = 2{,}835$ genes, all subsequent bins contain $n = 4{,}351 \pm 1$ genes (inclusion criteria, see Methods); empirical confidence intervals (99%) were calculated through bootstrap sampling ($n = 100$ replicates) of genes within each bin. **c**, Comparison of template versus non-template mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **d**, Sequence-composition-normalized profiles of mutation rate around transcription start sites (TSS). **e**, Stratifying by lesion strand reveals how bidirectional transcription initiation shapes the observed mutation patterns. **f**, Higher resolution of the TSS region from **e**.

approximately 200 nucleotides upstream of the transcription start site (Fig. 3d) is revealed to result from genic and upstream bidirectional transcription emerging from opposite edges of the core promoter[23] leading to a local depletion of TCR activity within the promoter (Fig. 3e, f).

An engine for genetic diversity

A segregating lesion may act as template for multiple rounds of replication in successive cell cycles (Fig. 2i). Each replication could incorporate different incorrectly or correctly paired nucleotides opposite a persistent lesion, resulting in multiple alleles at the same position. Consistent with this notion, multiallelic mutations have been reported in human cancers[24] and a cell-lineage-tracking system[25].

We evaluated multiallelic variation by identifying sites with multiple high-confidence—but conflicting—mutation calls. On average, 8% of mutated sites in DEN-induced tumours have multiallelic variants ($n = 1.8 \times 10^6$ sites in C3H tumours); per tumour, this value ranges from less than 1% to 26% (Fig. 4a). As a control, only 0.098% (95% confidence interval: 0.043–0.25%) of sites permuted between tumours show evidence of non-reference nucleotides. We further validated WGS multiallelic-variant calls using independently performed exome sequencing[6] (Fig. 4b).

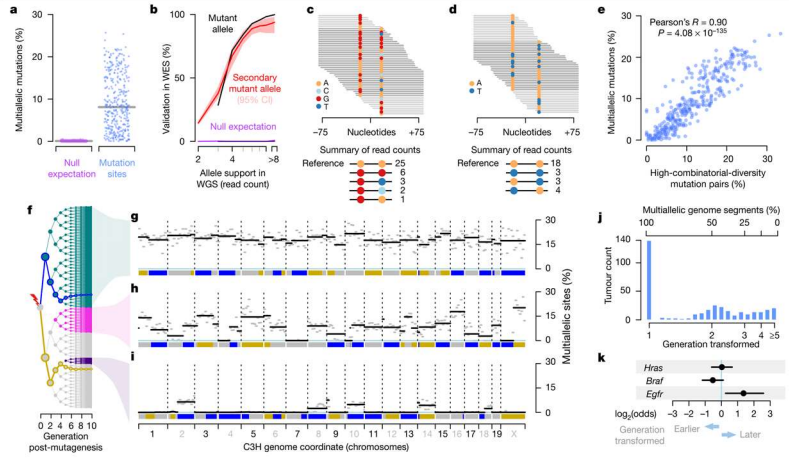


Fig. 4 | Lesion segregation generates multiallelic and combinatorial genetic diversity. a, Mutation sites per tumour with robust support for multiallelic variation; the grey line shows the median; null expectation is from permutation between tumours. **b**, Validation rate for mutations from WGS in independent whole-exome sequencing (WES); $n = 15$ tumours, collectively with $n = 20{,}683$ WGS mutations meeting inclusion criteria (Methods). Curves show validation rate stratified by WGS read support. Empirical 95% confidence interval (CI) from 100 bootstrap samplings of the aggregated WGS mutations. The null expectation permuted tumour identity between WGS and WES. **c**, Sequence reads spanning proximal mutated sites. **d**, As **c**, but showing combinatorial diversity between a pair of biallelic sites. **e**, Correlation between per-tumour multiallelic rate and combinatorially diverse mutation pairs (as in **c**, **d**), with one point per tumour. **f**, Tree of all possible progeny of a DEN-mutagenized cell for ten generations. Blue and gold lines trace simulated segregation of lesion-containing strands from a single haploid chromosome. Coloured nodes show hypothetical transformed daughter lineages with their multiallelic patterns (right). **g**–**i**, Mutation asymmetry summary ribbons for example C3H tumours that show high (**g**), variable (**h**) or low (**i**) rates of genetic diversity. The percentage of mutation sites with robust support for multiallelic variation calculated in 10-Mb windows (grey) and for each asymmetric segment (black). **j**, Histogram of the estimated cell generation post-DEN exposure from which tumours developed based on the proportion of multiallelic segments. **k**, Enrichment of specific driver gene mutations in earlier (generation 1) and later (generation >1) transforming tumours. $\log_2$ odds ratios (circles) from Fisher's exact test with 95% confidence intervals (whiskers) calculated from the hypergeometric distribution. All $n = 371$ tumours were included in the analysis for each gene.

The generation of multiallelic variation produces combinatorial genetic diversity that would not be expected under purely clonal expansion. This can be directly visualized in pairs of mutations spanned by individual sequencing reads (Fig. 4c, d). The observed combinations of biallelic sites require replication over lesions without the generation of mutations in some cell divisions (Fig. 4d). This directly demonstrates that non-mutagenic synthesis over DNA lesions occurs, and allele frequency analysis indicates it is common (Extended Data Fig. 6). The per-tumour rates of combinatorial diversity and multiallelic sites correlate closely and highlight the wide variation between tumours (Fig. 4e).

The explanation for such intertumour variance becomes evident when plotting the distribution of multiallelic sites along each genome (Fig. 4f–i). Tumours with high rates of genetic diversity have consistently high rates of multiallelism throughout their genome (Fig. 4g). They are likely to have expanded from a first-generation daughter of the original DEN-mutagenized cell, in which all DNA is a duplex of a lesion-containing and non-lesion-containing strand. Therefore, replication using lesion-containing strands as the template in subsequent generations produces multiallelic variation uniformly across the genome. Tumours with lower total levels of genetic diversity exhibit discrete genomic segments of high and low multiallelism (Fig. 4h, i). These tumours probably developed from a cell some generations after DEN treatment. Each mitosis following DEN exposure is expected to dilute the number of lesion-containing strands in each daughter cell by approximately 50%. Only lesion-retaining fractions of the genome generate multiallelic and combinatorial genetic diversity in the daughter lineages; consistent with this, the multiallelic segments mirror the mutational asymmetry segmentation pattern.

By estimating the fraction of multiallelic chromosomal segments, we can infer the cell generation, relative to DEN exposure, from which the tumour expanded (Fig. 4j). In 67% of C3H tumours and 21% of CAST tumours, the initial burst of mutations was instantly transformative. In the remainder of tumours, the observed fractions of multiallelic segments cluster around expectations for subsequent cell generations, suggesting that transformation required a specific combination of mutated alleles, an additional mutation or an external trigger. Of note, *Egfr*-driven tumours appear to transform significantly later ($P = 0.042$ after Bonferroni correction, Fisher's exact test), suggesting that driver gene identity influences the timing of tumour inception (Fig. 4k).

Lesion segregation is ubiquitous

Lesion segregation is a feature of DEN mutagenesis in mice. This raises two critical questions. Do other DNA-damaging agents induce lesion

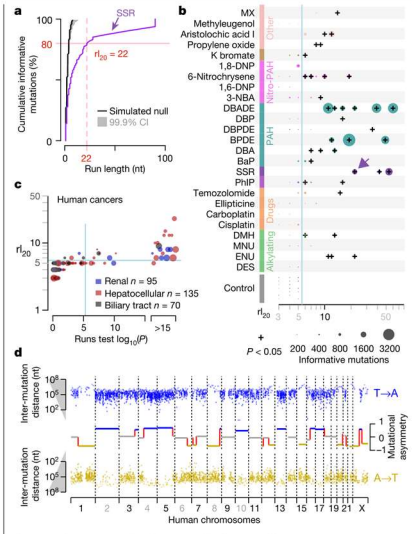


Fig. 5 | Lesion segregation is a pervasive feature of exogenous mutagens and is evident in human cancers. a, The runs-based rl_{20} metric, calculated for an example simulated solar radiation (SSR) clone (Extended Data Fig. 7a); 20% of informative mutations (C→T or their complement G→A) are in strand asymmetric runs of at least 22 consecutive mutations (for example, ≥22 C→T mutations without an intervening G→A). Simulated null based on 100,000 permutations of 1,000 mutations; black curve shows median. **b**, All robust mutagens in human iPSCs[5], mutagen classes indicated by coloured boxes; PAH indicates polycyclic aromatic hydrocarbons. Individual compound abbreviations expanded in Supplementary Table 2. The rl_{20} metric (*x*-axis) is plotted for each clone ($n = 325$), including multiple replicates per exposure. Data point size quantifies informative mutations; $^*P < 0.05$ (two-sided, Bonferroni-corrected). **c**, The rl_{20} metric and runs tests for human cancers[27]; $n = 18{,}850$ cancers screened, three cohorts plotted. Blue lines show Bonferroni adjusted $P = 0.05$ threshold for the runs test (two-sided) and an empirical threshold for rl_{20} (Methods). *x*-axis *P*-values $<1 \times 10^{-15}$ are rank-ordered. **d**, Mutational asymmetry (plotted as in Fig. 2a–c) in a human hepatocellular carcinoma (donor DO231953) with a dominant mutation signature for aristolochic acid exposure.

segregation? Does lesion segregation occur in human cells and cancers? Recently, a study in which human induced pluripotent stem cells (iPSCs) were exposed to 79 environmental mutagens revealed that 41 of the mutagens produced excess nucleotide substitutions[5]. Although not previously noted in these in vitro data, many of the exposures generated chromosome-scale lesion segregation patterns (Extended Data Fig. 7) similar to those observed in the in vivo DEN model. Applying runs-based tests (Fig. 5a, b, Extended Data Fig. 8), we detect marked mutational asymmetry in every sample with more than 1,000 'informative' mutations (Fig. 5b, Extended Data Fig. 8b; see Methods), including clinically relevant insults such as sunlight (simulated solar radiation), tobacco smoke (benzo[*a*]pyrene diol-epoxide (BPDE)) and chemotherapeutics (temozolomide). By contrast, mutations induced by perturbation of replication and repair pathways[26] independent of DNA lesions showed no detectable asymmetry, as expected (Extended Data Fig. 8c). We conclude that the chromosome-scale segregation of lesions and the resulting strand asymmetry of mutation patterns, are general features of all tested DNA-damaging mutagens.

The pronounced mutational asymmetry observed in both DEN-induced tumours and mutagen-exposed human iPSCs[5] occurs after a single mutagenic insult. By contrast, most human cancers accumulate mutations as a result of multiple damaging events over their history. Lesion segregation predicts that such tumours will acquire new waves of segregating lesions after each exposure, thus progressively masking their asymmetry patterns. Therefore, even though UV exposure causes substantial lesion segregation in human cells (Fig. 5a, b, Extended Data Figs. 7a, 8b), it is unlikely that skin cancers would show mutational asymmetry following repeated UV exposure.

Nevertheless, analysis of human cancer genomes[27] ($n = 18{,}850$ tumours, 22 primary sites) identified multiple cancers with the characteristic mutational asymmetry of lesion segregation (Fig. 5c, d). The majority of these tumours are renal, hepatic or biliary in origin, and show a high mutation rate and strand asymmetry of T→A or their complement A→T mutations, consistent with exposure to aristolochic acid[3] (Supplementary Table 2). Although it is seen most clearly in tumours subjected to a single dose of a mutagen, lesion segregation probably shapes all genomes subjected to DNA damage, with important implications for tumour evolution and heterogeneity.

Discussion

In this study, we have shown that most mutation-causing DNA lesions are not resolved as mutations within a single cell cycle. Instead, lesions segregate unrepaired into daughter cells for multiple cellular generations, resulting in chromosome-scale strand asymmetry of subsequent mutations. This suggests that lesion removal before replication has high fidelity and rarely results in mutations. Lesion segregation was initially discovered in an in vivo mouse model of oncogenesis; we have demonstrated that it is ubiquitous for all tested mutagens, also occurs in human cells and is evident in human cancers. Similar patterns of asymmetry in bacterial mutagenesis suggest that the underlying mechanisms are highly conserved[28,29].

Our discovery of lesion segregation challenges longstanding assumptions of cancer evolution[30]. For example, the widely used infinite sites model[31] does not allow for recurrent mutation at the same site. Our findings also provide new perspectives for understanding cancer evolution using mutational asymmetry and multiallelism patterns to track events during oncogenesis and to quantify selection. Perhaps most notably, lesion segregation is a previously unrecognized mechanism for a cancer to sample the fitness effects of mutation combinations, thus evading Muller's ratchet[32] and Hill–Robertson interference, which assumes low selection efficiency owing to the inability to separate mutations of opposing fitness[33,34]. Consequently, DNA-damaging chemotherapeutics, particularly large or closely spaced doses generating persistent lesions, could inadvertently provide an opportunity for cancer to efficiently select resulting mutations. This insight may guide the development of more effective chemotherapeutic regimens.

Once identified, lesion segregation is a deeply intuitive concept. Its practical applications provide new vistas for the exploration of genome maintenance and fundamental molecular biology. The discovery of pervasive lesion segregation profoundly revises our understanding of how the architecture of DNA repair and clonal proliferation can conspire to shape the cancer genome.

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/s41586-020-2435-1.

1. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
3. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
5. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
6. Connor, F. et al. Mutational landscape of a chemically-induced mouse model of liver cancer. *J. Hepatol.* **69**, 840–850 (2018).
7. Maronpot, R. R. Biological basis of differential susceptibility to hepatocarcinogenesis among mouse strains. *J. Toxicol. Pathol.* **22**, 11–33 (2009).
8. Wang, C. et al. Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat. Commun.* **9**, 2054 (2018).
9. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
10. Verna, L., Whysner, J. & Williams, G. M. *N*-nitrosodiethylamine mechanistic data and risk assessment: bioactivation, DNA-adduct formation, mutagenicity, and tumor initiation. *Pharmacol. Ther.* **71**, 57–81 (1996).
11. Maronpot, R. R., Fox, T., Malarkey, D. E. & Goldsworthy, T. L. Mutations in the ras proto-oncogene: clues to etiology and molecular pathogenesis of mouse liver tumors. *Toxicology* **101**, 125–156 (1995).
12. Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J. & Schwarz, M. Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutat. Res.* **638**, 66–74 (2008).
13. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
14. 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).
15. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294 (2019).
16. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
17. Perry, P. & Evans, H. J. Cytological detection of mutagen-carcinogen exposure by sister chromatid exchange. *Nature* **258**, 121–125 (1975).
18. Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Front. Genet.* **5**, 175 (2014).
19. Strick, T. R. & Portman, J. R. Transcription-coupled repair: from cells to single molecules and back again. *J. Mol. Biol.* **431**, 4093–4102 (2019).
20. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–960 (2015).
21. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
22. Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
23. Preker, P. et al. PROMoter uPstream transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179–7193 (2011).
24. 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).
25. Brody, Y. et al. Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Res.* **28**, 1901–1918 (2018).
26. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
27. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
28. Parkhomchuk, D., Amstislavskiy, V., Soldatov, A. & Ogryzko, V. Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proc. Natl Acad. Sci. USA* **106**, 20830–20835 (2009).
29. Chan, K. & Gordenin, D. A. Clusters of multiple mutations: incidence and molecular mechanisms. *Annu. Rev. Genet.* **49**, 243–267 (2015).
30. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).
31. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
32. Zhang, Y. et al. Genetic load and potential mutational meltdown in cancer cell populations. *Mol. Biol. Evol.* **36**, 541–552 (2019).
33. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
34. Tilk, S., Curtis, C., Petrov, D. & McFarland, C. D. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. Preprint at *bioRxiv* https://doi.org/10.1101/764340 (2019).

Liver Cancer Evolution Consortium

Sarah J. Aitken[1,2,3], Stuart Aitken[4], Craig J. Anderson[4], Claudia Arnedo-Pac[5], Frances Connor[1], Ruben M. Drews[1], Ailith Ewing[4], Christine Feig[1], Paul Flicek[1,6], Vera B. Kaiser[4], Elissavet Kentepozidou[6], Erika López-Arribillaga[5], Núria López-Bigas[5,10,11], Juliet Luft[4], Margus Lukk[1], Duncan T. Odom[1,12]✉, Oriol Pich[5], Tim F. Rayner[1], Colin A. Semple[4], Inés Sentís[5], Vasavi Sundaram[1,6], Lana Talmane[4] & Martin S. Taylor[4]✉

Methods

Statistical methods were used to predetermine sample size for the testing of oncogenic selection by biased strand retention; otherwise, no statistical methods were used to determine sample size. The investigators were blinded to allocation during histopathological assessment.

Mouse colony management

Animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical Review Body (AWERB): the maximum approved tumour burden was 10% body weight, which was not exceeded. Animals were maintained using standard husbandry: mice were group housed in Tecniplast GM500 IVC cages with a 12 h:12 h light:dark cycle and ad libitum access to water, food (LabDiet 5058), and environmental enrichments.

Chemical model of hepatocarcinogenesis

P15 male C3H and CAST mice were treated with a single intraperitoneal (IP) injection of DEN (Sigma-Aldrich N0258; 20 mg kg^{-1} body weight) diluted in 0.85% saline. Liver tumour samples were collected from DEN-treated mice 25 weeks (C3H) or 38 weeks (CAST) after treatment. All macroscopically identified tumours were isolated and processed in parallel for DNA extraction and histopathological examination. Non-tumour tissue from untreated P15 mice (ear, tail, and background liver) was sampled for control experiments.

Tissue collection and processing

Liver tumours of sufficient size (≥2 mm diameter) were bisected; one half was flash frozen in liquid nitrogen and stored at −80 °C for DNA extraction, and the other half was processed for histology. Tissue samples for histology were fixed in 10% neutral buffered formalin for 24 h, transferred to 70% ethanol, machine processed (Leica ASP300 Tissue Processor; Leica), and paraffin embedded. All formalin-fixed paraffin-embedded sections were 3 µm in thickness.

Histochemical staining

Formalin-fixed paraffin-embedded tissue sections were stained with haematoxylin and eosin (H&E) using standard laboratory techniques. Histochemical staining was performed using the automated Leica ST5020; mounting was performed on the Leica CV5030.

Imaging

Tissue sections were digitised using the Aperio XT system (Leica Biosystems) at 20× resolution; all H&E images are available in the BioStudies archive at EMBL-EBI under accession S-BSST383.

Tumour histopathology

H&E sections of liver tumours were blinded and assessed twice by a pathologist (S.J.A.); discordant results were reviewed by an independent hepatobiliary pathologist (S.E.D.). Tumours were classified according to the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) guidelines for lesions in rats and mice[35]. In addition, tumour grade, size, morphological subtype, nature of steatosis and mitotic index were assessed (Supplementary Table 1), as well as the presence of cystic change, haemorrhage, necrosis, or vascular invasion.

Sample selection for WGS

Tumours which met the following histological criteria were selected for WGS (C3H $n = 371$, CAST $n = 84$): (i) diagnosis of either dysplastic nodule (DN) or hepatocellular carcinoma (HCC), (ii) homogenous tumour morphology, (iii) tumour cell percentage >70%, and (iv) adequate tissue for DNA extraction. Neoplasms with extensive necrosis, mixed tumour types, a nodule-in-nodule appearance (indicative of an HCC arising within a DN), or contamination by normal liver tissue were excluded. Since carcinogen-induced tumours arising in the same liver are independent[6], multiple tumours were selected from each mouse to minimise the number of animals used. A subset of normal (non-tumour) samples from untreated mice were also sequenced (C3H $n = 13$, CAST $n = 7$).

Whole-genome sequencing

Genomic DNA was isolated from liver tissue and liver tumours using the AllPrep 96 DNA/RNA Kit (Qiagen, 80311) according to the manufacturer's instructions. DNA quality was assessed on a 1% agarose gel and quantified using the Quant-IT dsDNA Broad Range Kit (Thermo Fisher Scientific). Genomic DNA was sheared using a Covaris LE220 focused-ultrasonicator to a 450-bp mean insert size.

WGS libraries were generated from 1 µg of 50 ng μl^{-1} high molecular weight genomic DNA using the TruSeq PCR-free Library Prep Kit (Illumina), according to the manufacturer's instructions. Library fragment size was determined using a Caliper GX Touch with a HT DNA 1k/12K/Hi Sensitivity LabChip and HT DNA Hi Sensitivity Reagent Kit to ensure fragments of 300–800 bp (target ~450 bp).

Libraries were quantified by real-time PCR using the Kapa library quantification kit (Kapa Biosystems) on a Roche LightCycler 480. 0.75 nM libraries were pooled in 6-plex and sequenced on a HiSeq X Ten (Illumina) to produce paired-end 150-bp reads. Each pool of 6 libraries was sequenced over eight lanes (minimum of 40× coverage).

Variant calling and somatic mutation filtering

Sequencing reads were aligned to respective genome assemblies (C3H = C3H_HeJ_v1; CAST = CAST_EiJ_v1)[36] with bwa-mem (v.0.7.12)[37] using default parameters. Reads were annotated to read groups using the Picard (v.1.124)[38] tool AddOrReplaceReadGroups, and minor annotation inconsistencies corrected using the Picard CleanSam and FixMateInformation tools. Bam files were merged as necessary, and duplicate reads were annotated using the Picard tool MarkDuplicates.

Single-nucleotide variants were called using Strelka2 (v.2.8.4)[39] implementing default parameters. Initial variant annotation was performed with the GATK (v.3.8.0)[40] walker CalculateSNVMetrics (https://github.com/crukci-bioinformatics/gatk-tools). Genotype calls with a variant allele frequency <0.025 were removed. Although inbred strains were used, fixed genetic differences between the colonies and the reference genome, as well as small numbers of germline variants segregating within the colonies were identified. For each strain, fixed differences identified as homozygous changes present in 100% of genotyped samples were filtered out. Segregating variants were filtered based on the excess clustering of mutations to animals with shared mothers. To generate a null expectation taking into account the family structure of the colonies, the parent–offspring relationships were randomly permuted 1,000 times. For each count of recurrent mutation (range 5 to 371 inclusive), we determined the null distribution of expected distinct mothers. Comparing this to the observed count of distinct mothers for each recurrent ($n > 4$) mutation, those with a low probability ($P < 1x10^{-4}$, pnorm function from R (v.3.5.1)[41]) under the null were excluded from analyses.

Copy number variation between tumours within strains was called using CNVkit (v.0.9.6)[42]. Non-tumour reference coverage was provided from non-tumour control WGS data (C3H $n = 11$, CAST $n = 7$) and per tumour cellularity estimates (see below) were provided.

RNA sequencing

Total RNA was extracted from P15 liver tissue ($n = 4$ biological replicates per strain) using QIAzol Lysis Reagent (Qiagen), according to manufacturer's instructions. DNase treatment and removal were performed using the TURBO DNA-free kit (Ambion, Life Technologies), according to the manufacturer's instructions. RNA concentration was measured using a NanoDrop spectrophotometer (Thermo Fisher); RNA integrity was assessed on a Total RNA Nano Chip Bioanalyzer (Agilent).

Total RNA (1 μg) was used to generate sequencing libraries using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina), according to the manufacturer's instructions. Library fragment size was determined using a 2100 Bioanalyzer (Agilent). Libraries were quantified by quantitative PCR (Kapa Biosystems). Pooled libraries were sequenced on a HiSeq4000 to produce ≥ 40 million paired-end 150-bp reads per library.

RNA-seq data processing and analysis

Transcript abundances were quantified with Kallisto (v.0.43.1)[43] (using the flag–bias) and a transcriptome index compiled from coding and non-coding cDNA sequences defined in Ensembl v91[44]. TPM estimates were generated for each annotated transcript and summed across alternate transcripts of the same gene for gene-level analysis. The TSS for each gene was annotated with Ensembl v91 and based upon the most abundantly expressed transcript. RNA-sequencing (RNA-seq) data are available at Array Express at EMBL-EBI under accession E-MTAB-8518.

Genomic annotation data

Mouse liver proximity ligation sequencing (HiC) data were downloaded from GEO (GSE65126)[45], replicates were combined, then aligned to GRCm38[46] and processed using the Juicebox (v.7.5) and Juicer scripts[47] to obtain the HiC matrix. Eigenvectors were obtained for 500kb consecutive genomic windows over each chromosome from the HiC matrix using Juicebox and subsequently oriented (to distinguish compartment A from B) using GC content per 500-kb bin. We used progressiveCactus[48] to project the 500-kb windows into the C3H reference genome and Bedtools (v.2.28.0) to merge syntenic loci between 450 and 550 kb in size, removing the second instance where we observed overlaps.

Genic annotation was obtained from Ensembl v.91[44] for the corresponding C3H and CAST reference genome assemblies (C3H_HeJ_v1, CAST_EiJ_v1). Genomic repeat elements were annotated using RepeatMasker (v.20170127; http://www.repeatmasker.org) with the default parameters and libraries for mouse annotation.

The analysable fraction of the genome

Analysis and sequence composition calculations were confined to the main chromosome assemblies of the reference genome (chromosomes 1–19 and X). Using WGS of non-tumour liver, ear and tail samples (C3H $n = 11$, CAST $n = 7$) collected and sequenced contemporaneously with tumour samples, genome sequencing coverage was calculated for 1-kb windows using multicov in Bedtools (v.2.28.0)[49]. Windows with read coverage >2 s.d. from the autosomal mean were flagged as suspect in each tumour. Read coverage over the X chromosome was doubled in these calculations to account for the expected hemizygosity in these male mice. Any 1-kb window identified as suspect in >90% of these non-tumour samples was flagged as 'abnormal read coverage' (ARC) and masked from subsequent analysis. This masked 12.7% of the C3H and 11.5% of the CAST reference genomes yielding analysable haploid genome sizes of C3H = 2,333,783,789 nucleotides (nt) and CAST = 2,331,370,397 nt.

Mutation rate calculations

Mutation rates were calculated as 192 category vectors representing every possible single-nucleotide substitution conditioned on the identity of the upstream and downstream nucleotides. Each rate being the observed count of a mutation category divided by the count of the trinucleotide context in the analysed sequence. To report a single aggregate mutation rate, the three rates for each trinucleotide context were summed to give a 64 category vector and the weighted mean of that vector reported as the mutation rate. The vector of weights being the trinucleotide sequence frequency of a reference sequence, for example the composition of the whole genome. In the case of whole-genome analysis, the same trinucleotide counts are used in (1) the individual category rates calculation and (2) the weighted mean of the rates, cancelling out. For windowed comparisons of mutation rates, the weighted mean is calculated using the genome wide composition of trinucleotides rather than the local sequence composition, providing a compositionally adjusted mutation rate estimate. For mutation rates in TCR analysis, the same compositional adjustment was carried out but using the trinucleotide composition of the aggregate genic spans of genome (minus ARC regions) for normalization.

Mutation signatures

The 96 category 'folded' mutation counts for each of the 371 C3H tumours were deconvolved into the best fitting number (K) of component signatures using sigFit (v.2.0)[50] with 1,000 iterations and K set to integers 2 to 8 inclusive. A heuristic goodness-of-fit score based on cosine similarity favoured instances where $K = 2$. The DEN1 and DEN2 signatures reported were obtained by running sigFit with 30,000 iterations for $K = 2$. Analysis of CAST tumours gave less distinct separation of signatures so the C3H derived DEN1 and DEN2 were used for both strains. To fit signatures to each tumour we used sigFit provided with the DEN signatures and additional SPONT1 and SPONT2 signatures that were derived from equivalent WGS analysis of spontaneous (non-DEN-induced) C3H tumours.

Driver mutation identification

Candidate cancer driver genes were identified by applying OncodriveFML (v.2.2.0 using the SIFT scoring scheme)[51] and OncodriveCLUSTL (v.1.1.1)[52] to mutations identified in C3H tumours. The only genes convincingly identified as significantly enriched for functionally impactful or clustered mutations were *Braf*, *Egfr* and *Hras*. *Kras* appeared as marginally significant. These four genes were identified for C3H[6]. Protein altering mutations in those genes were annotated as driver mutations in C3H and CAST tumours.

Mutational asymmetry segmentation and scoring

For each tumour a focal subset of 'informative' mutation types were defined, T→N or A→N mutations, in the case of DEN-induced tumours. The order of focal mutations along each chromosome was represented as a binary vector (for example, 0 for T→N, 1 for A→N). Vectors corresponding to each chromosome of each tumour were processed with the cpt.mean function of the R Changepoint (v.2.2.2)[53] package run with an Akaike information criterion (AIC) penalty function, maximum number of changepoints set to 12 ($Q = 12$), and implementing the PELT algorithm for optimal changepoint detection. Following segmentation, the defined segments were scored for strand asymmetry, taking into account the sequence composition of the segment. For example in tumours with T→N or A→N informative mutations the number of Ts on the forward strand is the count of Watson sites G_W and the number of T→N mutations is μ_W which together give the Watson strand rate $R_W = \mu_W/G_W$. The forward strand count of As and mutations from A likewise give the Crick strand rate $R_C = \mu_C/G_C$. From these two rates we calculate a relative difference metric, the mutational asymmetry score $S = (R_W - R_C)/(R_W + R_C)$.

The parameter S scales from 1 all Watson (for example, DEN T→N mutations) through 0 (50:50 T→N:A→N) to −1 for all Crick (for example, DEN A→N). For the categorical assignment, $S \geq 0.3$ is Watson-strand asymmetric, $S \leq -0.3$ Crick-strand asymmetric and in the range $-0.3 < S < 0.3$ symmetric, though more stringent filtering was applied where noted. Segments containing <20 informative mutations were discarded from subsequent analyses.

To test for oncogenic selection at sites with recurrent mutations, mutational asymmetry segments overlapping the focal mutation were categorised based on their asymmetry score S, as above. The test was implemented as a Fisher's exact test with the 2 × 2 contingency table comprising the counts chromosomes (two autosomes per cell) stratified by Watson-versus-Crick asymmetry and the presence of the focal mutation in the tumour. Tumours containing another known driver

gene or recurrent mutation within the focal asymmetry segment were discarded from the analysis. We estimated the minimum recurrence of a mutation necessary to reliably detect oncogenic selection through simulation. Biased segregation of chromosomes containing drivers was modelled using the observed median excess of T→N over A→N lesions (23-fold), and random segregation of non-driver containing strands (1:1 ratio). Our model predicted >33 C3H recurrences or >41 CAST recurrences would give 80% power to detect oncogenic selection if present.

Tumour cellularity estimates

We calculated tumour cellularity as a function of the non-reference read count in autosomal chromosomes $(1 - R/d) \times 2$, where R is the reference read count at a mutated site and d is the total read depth at the site. For each tumour these values were binned in percentiles and the midpoint of the most populated (modal) percentile taken as the estimated cellularity of the tumour. Given the low rate of copy number variation across the DEN induced tumours, no correction was made for copy-number distortion. Skew in the variant allele frequency (VAF) = $(1 - R/d)$) distribution was calculated using Pearson's median skewness coefficient implemented in R as (3 × (mean − median))/s.d. of the VAF distribution.

Identifying and filtering reference genome misassemblies

Since lesion segregation, mutation asymmetry patterns allow the long-range phasing of chromosome strands, they can detect discrepancies in sequence order and orientation between the sequenced genomes and the reference. We identified autosomal asymmetry segments that immediately transitioned from Watson bias ($S > 0.3$) to Crick ($S < -0.3$) or vice versa without occupying the intermediate unbiased state ($-0.3 < S < 0.3$); such discordant segments are unexpected. Allowing for ±100 kb uncertainty in the position of each exchange site we produced the discordant segment coverage metric. At sites with discordant segment coverage >1 we calculated percentage consensus for misassembly $M = ds/(ds+cs)$ where ds is the number of discordant segments over the exchange site and cs the number of concordant: where either Watson or Crick mutational asymmetry extends at least 1×10^6 nucleotides on both sides of the exchange site. The approximate genomic coordinates for a C3H strain specific inversion on chromosome 6 have been previously reported[54].

SCE-site analysis

Identified SCE sites were aggregated across tumours from each strain. Exchange sites within 1×10^6 nt of known and proposed reference genome misassembly sites were excluded from analysis. The mid-point between the flanking informative mutations was taken as the reference genome position of the exchange event, and the distance between those flanking mutations as the positional uncertainty of the estimate. To generate null expectations for mutation rate measures, the coordinate of an exchange was projected into the genome of a proxy tumour and the mutation rates and patterns measured from that proxy tumour (repeated 100 times). The permutation of tumour identifiers for the selection of proxy tumours was a shuffle without replacement that preserved the total number of exchange sites measured in each tumour.

The comparison of mutation spectra between windows was calculated as the cosine distance between the 96 category trinucleotide context mutation spectra for the whole genome and that calculated for the aggregated 5-kb window. The 96 categories were equally weighted for this comparison.

Exchange site enrichment analysis used Bedtools[49] shuffle to permute the genomic positions of exchange sites into the analysable fraction of the genome (defined above). Observed rates of annotation overlap were compared to the distribution of values from 1,000 permuted exchange sites. For genic overlaps we used Ensembl v.91[44] coordinates for genic spans; gene expression status was based on the summed expression over all annotated transcripts for the gene from P15 liver from the matched mouse strain. Expression thresholds were defined as >50th centile for active and <50th centile for inactive genes.

A higher count of informative mutations provides greater power to identify shorter mutational asymmetry segments. To fairly test for correlation between nucleotide substitution rate and SCE rate we randomly down-sampled informative mutations to 10,000 per tumour genome and recomputed the mutational asymmetry segmentation patterns from the sampled data. Tumours with <10,000 informative mutations were excluded. We then correlated the total (not down sampled) nucleotide substitution load to the count of SCE events inferred from the down-sampled data.

TCR calculations

For each protein coding gene, the maximally expressed transcript isoform was identified from P15 liver in the matched strain (TPM expression), subsequently the primary transcripts. In the case of ties, transcript selection was arbitrary. Genes were partitioned into five categories based on the expression of the primary transcript: expression level 0 (<0.0001 TPM) and four quartiles of detected expression.

Using the segmental asymmetry patterns of each tumour and the annotated coordinates (Ensembl v.91) of the selected transcripts, we identified transcripts completely contained in a single Watson or Crick asymmetric segment and located at least 200 kb from the segment boundary at both ends. We also applied strict asymmetry criteria of mutational asymmetry scores $S > 0.8$ for Watson and $S < -0.8$ for Crick asymmetry segments, though analysis with the standard asymmetry thresholds and no segment boundary margin give similar results and identical conclusions. For each transcript in each tumour we then used both the transcriptional orientation of the gene and the mutational asymmetry of the segment containing it to resolve the segregated lesions to either the template (anti-sense) or non-template (sense) strand of the gene. Transcripts contained in mutationally symmetric regions or not meeting the strict filtering criteria were excluded from analysis.

We then analysed mutation rates stratifying by gene expression level and the template/non-template strand of the lesions but aggregating between tumours within the same strain. The TSS coordinates used correspond to the annotated 5′ end of the primary transcripts.

Multiallelic variation

Aligned reads spanning genomic positions of somatic mutations were re-genotyped using Samtools mpileup (v.1.9)[55]. Genotypes supported by ≥ 2 reads with a nucleotide quality score of ≥ 20 were reported, considering sites with two alleles as biallelic, those with three or four alleles as multiallelic. The fraction of called mutations exhibiting multiallelic variation was calculated for the analysable fraction of the genome, across 10Mb consecutive windows and also for each of the mutational asymmetry segments calculated for each tumour.

A null expectation for the multiallelic rate estimate was generated per C3H tumour; genomic positions identified as mutated across the other 370 tumours were down-sampled to match the mutation count in the focal tumour. Any of these proxy mutation sites with a non-reference genotype supported by ≥ 2 reads and nucleotide quality score ≥ 20 at the focal site were referred to as 'multiallelic' for the purposes of defining a background expectation for the calling of multiallelic variation. For each tumour, this was repeated 100 times and the mean reported.

We used WES of 15 C3H tumours from prior work[6] that have subsequently been used to generate WGS data in this study as a basis for validating multiallelic calls. Multiallelic variant positions derived from WGS were genotyped in WES using Samtools mpileup, as described above. Only sites with ≥ 30× WES coverage were considered and alleles were found to be concordant if a WGS genotype was supported by ≥ 1 read in the WES data. To provide a null expectation, the analysis was repeated using WES data from a different tumour and validation rates reported for all versus all combinations of mismatched WGS–WES pairs ($n = 15^2 - 15 = 210$).

To quantify combinatorial genetic diversity for each tumour, pairs of mutations located between 3 and 150 nt apart were phased using sequencing reads that traversed both mutation sites. Distinct allelic combinations were counted after extraction with Samtools mpileup using only reads with nucleotide quality score ≥ 20 over both mutation sites.

Estimating the cell generation of transformation

Knowing the fraction of lesion segregation segments that generated multiallelic variation across a tumour genome allows the inference of the generation time post-mutagenesis of the cell from which the tumour developed, because each successive cell generation is expected to retain only 50% of the lesion containing segments. We estimate this fraction as follows. Let p denote the fraction of multiallelic segments and let q be its complement, that is, the fraction of non-multiallelic segments, for each tumour genome. Segment boundaries being SCE sites or chromosome boundaries. In order to determine p, we re-purpose the quadratic Hardy–Weinberg equation: $p+q=p^2+2pq+q^2=1$, which holds since the two possible fractions need to sum to unity. Given an asymmetric segment of interest in the diploid genome, there are three distinct scenarios: (i) both chromosomes are multiallelic (p^2), (ii) One of the chromosomes is multiallelic and the other is not ($pq+qp$) and (iii) both chromosomes are non-multiallelic (q^2). The first two scenarios are not distinguishable from the data as both appear multiallelic (m). However, in the third scenario, for a segment to be non-multiallelic (biallelic, b), both chromosomal copies have to be non-multiallelic. As described below, q^2 can be estimated directly from the data and is subsequently used to estimate $P=1-\sqrt{(q^2)}$ and hence the cell generation number of transformation post-mutagenesis.

The estimation of q^2 requires computing the ratio $q^2=b/(b+m)$. We can directly observe the counts of b as non-multiallelic segments. The number of autosomal chromosome pairs ($n=19$) and count of SCE events (x) give the total number of segments in the genome $b+m=n+x$. Exchange events are not expected to align between allelic chromosomes which will result in the partial overlap of segments between allelic copies. Although this increases the number of observed segments (b and m) relative to actual segments, assuming the independent behaviour of allelic chromosomes and that segment length is independent of multiallelic state, this partial overlap does not systematically distort the quantification of b or the estimation of q^2.

To call a non-multiallelic segment (b) we require less than 4% multiallelic sites. The threshold is based on the tri-modal frequency distribution of multiallelic rates per segment, aggregated over all 371 C3H tumours. The 4% threshold separates the lower distribution of multiallelic rates from the mid and higher distributions.

To test for the enrichment of specific driver gene mutations in early generation versus late generation transformation post-DEN treatment, we applied Fisher's exact test (fisher.test function in R) to compare the generation 1 ratio of tumours with, versus those without a focal mutation, to the same ratio for tumours inferred to have transformed in a later generation. We additionally report the same odds ratios, but requiring that the "with focal mutation" tumours had a driver mutation in only one of the driver genes: *Hras*, *Braf*, or *Egfr*.

Cell-line and human cancer mutation analysis

Somatic mutation calls were obtained from DNA maintenance and repair pathway perturbed human cells[26]. Of the 128,054 reported single nucleotide variants, 6,587 unique mutations (genomic site and specific change) were shared between two or more sister clones, so probably represent mutations present but not detected in the parental clone. All occurrences of the shared mutations were filtered out leaving 106,688 mutations for analysis, although the inclusion of these filtered mutations does not alter any conclusions drawn. Somatic mutation calls from mutagen exposed cells[5] were obtained, no additional filtering was applied to these sub-clone mutations.

Somatic mutation calls from the International Cancer Genome Consortium (ICGC)[56] were obtained as simple_somatic_mutation.open.* files from release 28 of the consortium, one file for each project. These somatic mutations have been called from a mixture of WGS and WES. Of the 18,965 patients represented (and not embargoed in the release 28 data set), 116 were excluded from analysis; these represent a distinct WES subset of the LICA-CN project that appear to show a processing artefact in the distribution of specific mutation subsets. ICGC mutations were filtered to remove insertion and deletion mutations and also filtered for redundancy so that each mutation was only reported once for each patient. Mutation signatures deconvolution was performed using the R MutationPatterns (v.1.4.2)[57] package and COSMIC signature 22 was interpreted as aristolochic acid[3].

The rl_{20} metric and runs tests

Amongst only the informative mutations (for example, T→N/A→N in DEN) three consecutive T→N without an intervening A→N is a run of three. The R function rle was used to encode the run-lengths for binary vectors of informative mutations along the genome of a focal tumour. Ranking them from the longest to the shortest run, we find the set of longest runs that encompass 20% of all informative mutations in the tumour. The run-length of the shortest of those is reported as the rl_{20} metric. The threshold percent of mutations was defined as having to be less than 50%, as on average only 50% of the autosomal genomes are expected to show mutational asymmetry patterns. On testing with randomized data, the value of 20% gave a stable null expectation (maximum observed value of a run of five when simulating 10,000 informative mutations) and still encompassed a large fraction of the informative mutations. All rl_{20} results reported were implemented so that runs were broken when crossing chromosome boundaries. To define an empirical significance threshold for genomes with fewer mutations, we simulated 1,000 random informative mutations 100,000 times, >99.995% simulations had $rl_{20} \leq 5$ and 100% $rl_{20} \leq 6$.

The Wald–Wolfowitz runs test was performed using the runs.test function of the R randtests (v.1.0)[58] library. It was applied to binary vectors of informative changes as described above, with threshold = 0.5.

The Wald–Wolfowitz runs test significance is inflated by coordinated dinucleotide changes, such as those produced by UV light exposure and also other local mutational asymmetries such as replication asymmetry[13] and kataegis events[14,59]. The rl_{20} metric appears robust to most such distortions but we find it efficiently detects kataegis events that are in an otherwise mutationally quiet background, as is often the case for breast cancer. For this reason we also indicate the total genomic span of mutations in the rl_{20} subset of mutation runs: kataegis events typically span a tiny (<5%) fraction of the whole genome.

Key resources

The key reagents and resources required to replicate our study are listed in Supplementary Table 3. For externally sourced data, where applicable, URLs that we used can be found in the Git repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

Primary data processing was performed in shell-scripted environments calling the software indicated. Except where otherwise noted, analysis processing post-variant calling was performed in a Conda environment and choreographed with Snakemake running in an LSF batch control system (Supplementary Table 3).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The WGS FASTQ files are available from the European Nucleotide Archive (ENA) under accession number PRJEB37808. RNA-seq files are

available from Array Express under experiment number E-MTAB-8518. Digitised histology images are available from Biostudies under accession S-BSST383.

Code availability

The analysis pipeline including Conda and Snakemake configuration files can be obtained without restriction from the repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

35. Thoolen, B. et al. Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol. Pathol.* **38** (Suppl), 5S–81S (2010).
36. Lilue, J. et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Picard Tools (Broad Institute, 2019); http://broadinstitute.github.io/picard
39. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
40. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
41. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2013).
42. 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).
43. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-seq quantification with kallisto. *Nat. Biotechnol.* **34**, 525–527 (2016).
44. Cunningham, F. et al. Ensembl 2019. *Nucleic Acids Res.* **47** (D1), D745–D751 (2019).
45. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
46. Church, D. M. et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
47. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
48. Armstrong, J. et al. Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. Preprint at *bioRxiv* https://doi.org/10.1101/730531 (2019).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at *bioRxiv* https://doi.org/10.1101/372896 (2018).
51. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
52. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 5396 (2019).
53. Killick, R. & Eckley, I. A. changepoint: an R package for changepoint analysis. *J. Stat. Softw.* **58**, 1–19 (2014).
54. Akeson, E. C. et al. Chromosomal inversion discovered in C3H/HeJ mice. *Genomics* **87**, 311–313 (2006).
55. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
57. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
58. Caeiro, F. & Mateus, A. randtests: testing randomness in R. (2014).
59. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
60. Singer, B. In vivo formation and persistence of modified nucleosides resulting from alkylating agents. *Environ. Health Perspect.* **62**, 41–48 (1985).

Acknowledgements We thank M. Roller and F. Markowetz for supervision; L. Mularoni and G. Ritchie for software support; the CRUK Cambridge Institute Core facilities for their valuable contribution: CRUK Biological Resources (A. Mowbray), Preclinical Genome Editing (L. Young, S. Kupczak, M. Cronshaw, P. Mackin, Y. Cheng and L. Hughes-Hallett), Genomics (J. Hadfield and F. Bowater), Bioinformatics (G. Brown, M. Eldridge and R. Bowers), Histopathology and ISH (L.-A. McDuffus, C. Brodie and J. Arnold), and Research Instrumentation (J. Gray); Edinburgh Genomics (Clinical) Facility: the EMBL-EBI technical services cluster (Z. Mears, A. Cristofori, T. Nowak, S. Nanuwa, V. Tabak and A. Checcucci); and W. Bickmore and C. Ponting for comments on the manuscript. This work was supported by: Cancer Research UK (20412 and 22398), the European Research Council (615584, 682398), the Wellcome Trust (WT108749/Z/15/Z, WT106563/Z/14/A and WT202878/B/16/Z), the European Molecular Biology Laboratory, the MRC Human Genetics Unit core funding programme grants (MC_UU_00007/11 and MC_UU_00007/16) and the ERDF/Spanish Ministry of Science, Innovation and Universities-Spanish State Research Agency/DamReMap Project (RTI2018-094095-B-I00). S.J.A. received a Wellcome Trust PhD Training Fellowship for Clinicians (WT106563/Z/14/Z) and is now funded by a National Institute for Health Research (NIHR) Clinical Lectureship. O.P. is funded by a BIST PhD fellowship supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the Barcelona Institute of Science and Technology. V.S. is supported by an EMBL Interdisciplinary Postdoc (EIPOD) fellowship under Marie Skłodowska Curie actions COFUND (664726). E.K. is supported by the EMBL International PhD Programme. C.A-P. is supported by La Caixa Foundation fellowship (ID 100010434; LCF/BQ/ES18/11670011). S.V.B. is supported by ERC Starter Grant 759967. A.E. is supported by a UKRI Innovation Fellowship (MR/RO26017/1). A.K. is a cross-disciplinary postdoctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (core grant to the MRC Institute of Genetics and Molecular Medicine). I.S. is supported by an FPI fellowship from Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from Spanish Ministry of Science, Innovation and Universities (MICINN, Government of Spain) and is supported by CERCA (Generalitat de Catalunya).

Author contributions S.J.A., F.C., C.F. and D.T.O. conceived the project and designed the experiments. S.J.A., F.C. and C.F., performed the mutagenesis experiments and sequencing experiments. E.L.-A and A.M.R. performed supporting experiments. J.S.-L. provided contract sequencing. S.J.A. performed the histopathological analyses with S.E.D. providing advice. C.J.A. and M.S.T. designed and implemented computational analysis. M.S.T. discovered lesion segregation. O.P., V.S., T.F.R., M.L., S.A., E.K. and J.L. performed supporting computational analysis. C.A-P., S.V.B., R.M.D., A.E., V.B.K., A.K., I.S. and L.T. contributed to the computational analyses. T.F.R., M.L., S.A. and A.D.Y. curated data. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. supervised the work. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. lead the Liver Cancer Evolution Consortium. S.J.A. and P.F. provided scientific and administrative organisation. S.J.A., C.A.S., N.L.-B., P.F., D.T.O. and M.S.T. funded the work. S.J.A., D.T.O. and M.S.T. wrote the manuscript. All authors had the opportunity to edit the manuscript. All authors approved the final manuscript.

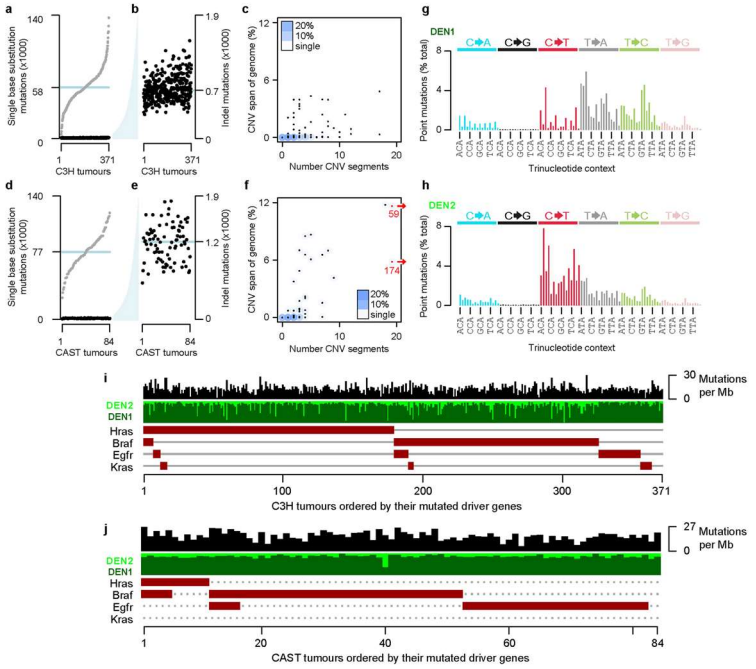
Competing interests P.F. is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd. The other authors declare no competing interests.

Additional information

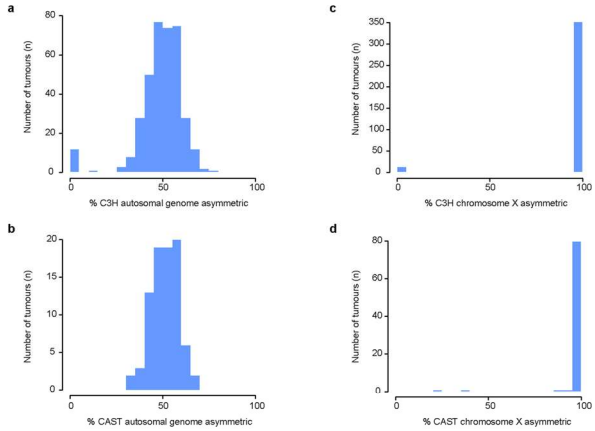
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2435-1.

Correspondence and requests for materials should be addressed to D.T.O. or M.S.T.

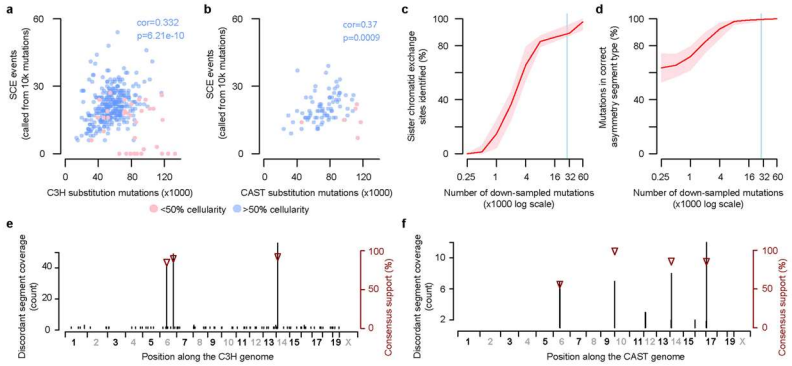
Peer review information *Nature* thanks Trevor Graham and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



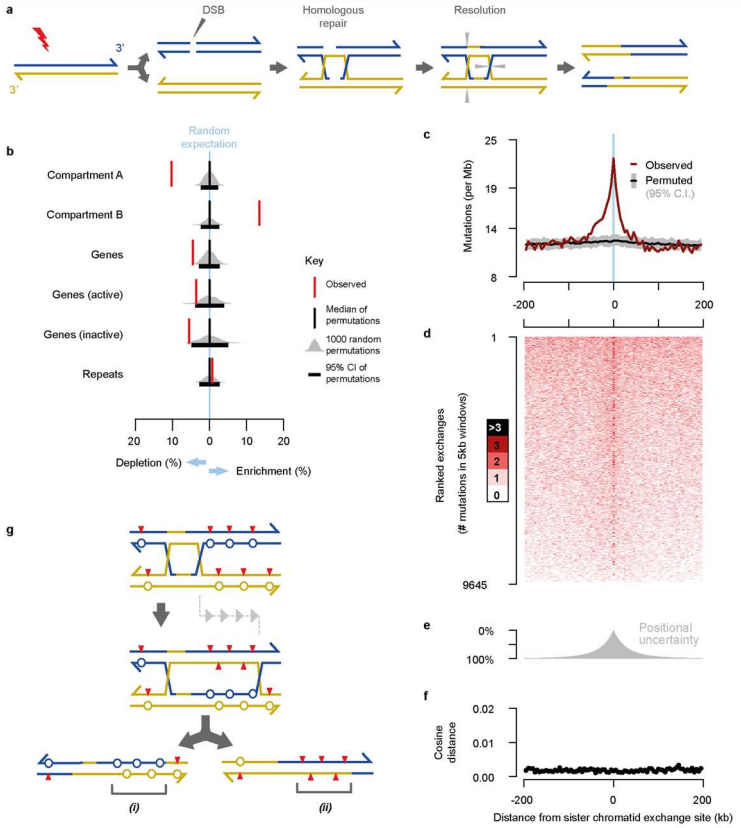
Extended Data Fig. 1 | Summary mutation metrics for C3H and CAST tumours. a, Single nucleotide substitution rates per C3H tumour, rank ordered over *x*-axis (grey points, median blue line). Insertion/deletion (indel, <11 nt) rates show as black. **b**, Y-axis from **a**, expanded to show distribution of indel rates with preserved tumour order. **c**, Number of C3H copy number variant (CNV) segments and their total span as a percent of the haploid genome. Blue shading shows intensity of overlapping points as a percent of all tumours in the plot. **d**–**f**, Corresponding plots for CAST derived tumours; **f**, two extreme *x*-axis outliers relocated (red) and *x*-axis value shown. **g**, **h**, Mutation spectra deconvolved from the aggregate spectra of 371 C3H tumours, subsequently referred to as the DEN1 and DEN2 signatures. DEN1 is dominated by T→N or their complement A→N changes thought to arise from the O^4-ethyl-deoxythymidine adduct of T[10]. DEN2 substitutions are primarily C→T or their complement G→A changes likely from O^6-ethyl-2-deoxyguanosine lesions of G[10]. **i**, Oncoplot summarizing mutation load, mutation signature composition, and driver gene mutation complement of C3H tumours. **j**, Oncoplot of CAST derived tumours as in **i**. The DEN2 signature is a minor component of most tumours but prominent in a minority (**i**, **j**).



Extended Data Fig. 2 | Mutational asymmetry across 50% of the autosomal genome and 100% of the haploid X chromosomes. a, **b**, Typically 50% of the autosomal genomic span (percent of nucleotides) in tumours is contained in segments with either Watson or Crick strand mutational asymmetry. **a**, C3H tumours, $n = 371$. **b**, CAST tumours, $n = 84$. **c**, **d**, Typically 100% of the haploid X chromosome shows Watson or Crick strand mutational asymmetry. **c**, C3H tumours ($n = 371$). **d**, CAST tumours ($n = 84$).

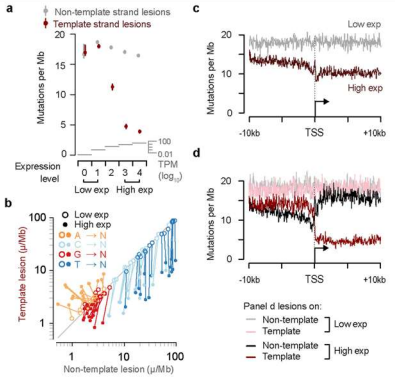


Extended Data Fig. 3 | The frequency of SCEs correlates with mutation rate, and localizing reference genome assembly errors. a, The relationship between single nucleotide substitution mutation load and detected SCE events in C3H tumours. DEN is known to produce ethyl adducts on the sugar-phosphate backbone of DNA as well as mutation-inducing modifications to the bases[10] which could lead to strand breaks[60] triggering SCE. The frequent observation and correlation between rates of SCE and point mutation supports this view. Counts of SCE (*y*-axis) are based on down-sampling to 10,000 informative mutations per tumour to ensure equal power to detect SCE in each tumour. Tumours with <50% cellularity (pink) have high mutation load and form a sub-group with few detected SCE events; these are suspected to be polyclonal tumours and were excluded from the Pearson's correlation reported ($n = 335$ independent tumour samples, implemented in a two-sided test, significance from Fisher's transform). **b**, As for **a**, but showing CAST derived tumours ($n = 84$, after cellularity exclusions $n = 77$). **c**, Evaluation of the relationship between mutation load and ability to detect SCE events. Mutations from C3H tumour 94315_N8 (shown in Fig. 2) randomly down-sampled and segmentation analysis applied. The *y*-axis shows the percentage of SCE events detected (100 replicates, mean red, 95% C.I. pink). The *x*-axis is on a log-scale: 95% of C3H and >95% of CAST tumours have mutation counts to the right of the blue vertical line. Down-sampling other tumours gave comparable results. **d**, The same down-sampling data as shown in **c** but the *y*-axis shows the percent of mutations with the correct (same as full data) mutational asymmetry assignment (mean red, 95% C.I. pink). **e**, Candidate C3H reference genome assembly errors. Genome coordinates shown on the *x*-axis. Immediate switches between Watson and Crick asymmetry are not expected on autosomes unless both copies of the chromosome have a SCE event at equivalent sites. However, inversions and translocations between the sequenced genomes and the reference assembly are expected to produce immediate asymmetry switches. The discordant segment coverage count (black *y*-axis) shows the number of informative tumours (those with either Watson or Crick strand asymmetry at the corresponding genome position) that suggest a tumour genome to reference genome discrepancy. Consensus support (brown *y*-axis) plotted as triangles shows the percentage of informative tumours that support a genomic discrepancy at the indicated position (only shown for values >50% support). The two sites on chromosome 6 in C3H correspond to a previously identified C3H strain specific inversion that is known to be incorrectly oriented in the C3H reference assembly[54]. **f**, Candidate CAST reference genome assembly errors, plotted as per **e**. The candidate misassembly on chromosome 14 in both strains occurs at an approximately orthologous position, suggesting a rearrangement shared between strains or a misassembly in the BL6 GRCm38 reference assembly against which other mouse reference genome assemblies have been scaffolded.

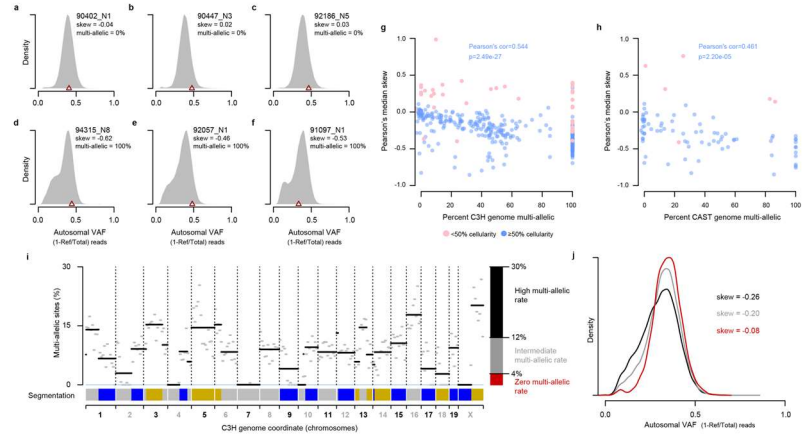


Extended Data Fig. 4 | See next page for caption.

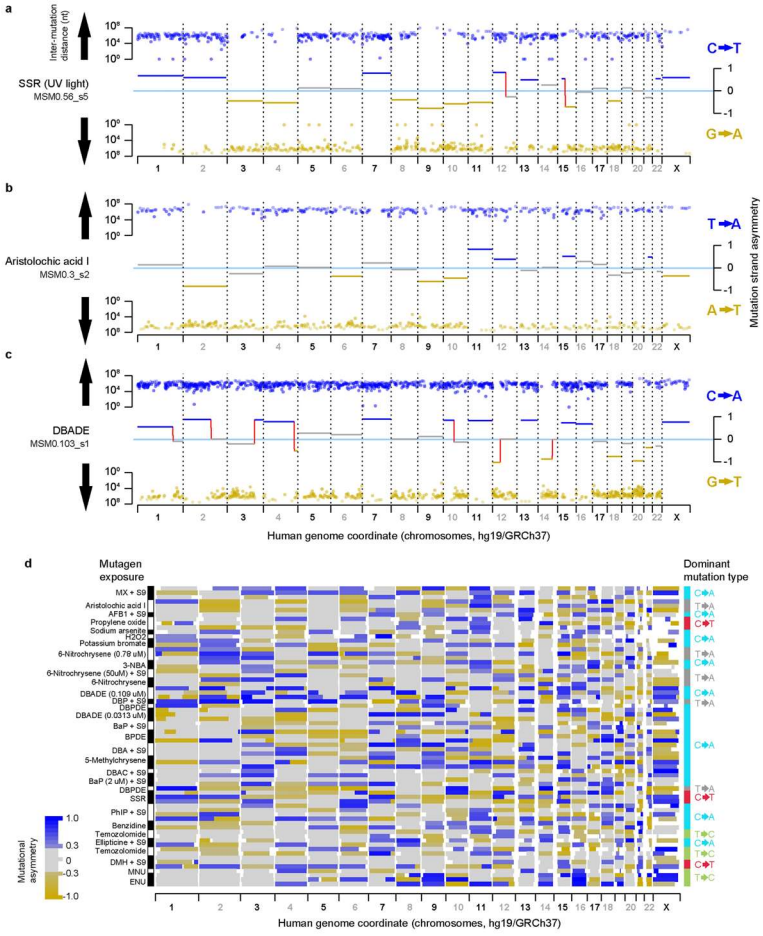
Extended Data Fig. 4 | Locally elevated mutation load is driven by SCE. **a**, Double strand breaks (DSBs) and other DNA damage can trigger homologous-recombination-mediated DNA repair between sister chromatids. The repair intermediate resolves into separate chromatids through cleavage and ligation; grey triangles denote cleavage sites for one of the possible resolutions that would result in a large-scale SCE event. Although illustrated for double-ended DNA breaks, single ended breaks from collapsed replication forks can be repaired through homologous recombination and could similarly lead to the formation of repair intermediate structures that can be resolved as SCEs. **b**, Enrichment analysis of SCE sites (red) compared with null expectations from randomly permuting locations into the analysable fraction of the genome (grey distributions), the black boxes denote 95% of 1,000 permutations. SCE events are enriched in later replicating and transcriptionally less active genomic regions (Hi-C defined compartment B), and correspondingly depleted from early replicating active regions. **c**, Aggregating across $n = 9{,}645$ SCE sites, the observed mutation rate approximately doubles at the inferred site of exchange ($x = 0$). Aggregate mutation rates (brown) were calculated in consecutive 5-kb windows. Compositionally matched null expectation was generated by permuting each exchange site into 100 proxy tumours and calculating median (black) and 95% confidence intervals (grey) while preserving the total number of projected sites per proxy tumour. **d**, The elevated mutation count is not the result of a high mutation density in a subset of exchange sites, rather it is a subtle increase in mutations across most exchange sites. Heatmap showing mutation counts calculated in consecutive 5-kb windows across each exchange site. Rows represent each exchange site, rank-ordered by total mutation count across each 400-kb interval. **e**, The distribution of positional uncertainty in exchange site location approximately mirrors the decay profile of elevated mutation frequency. **f**, Divergence of mutation rate spectra is shown as cosine distance between the analysed window and the genome wide mutation rate spectrum aggregated over all C3H tumours. Despite the elevated mutation frequency, there is no detected distortion of the mutation spectrum. **g**, A model based on homologous recombination repair intermediate, branch migration that produces heteroduplex segments of (i) mismatch:mismatch (circles) and (ii) lesion:lesion (red triangles) strands. Subsequent strand segregation would increase the mutational diversity of a descendant cell population but not the mutation count per cell (key as per Fig. 2).



Extended Data Fig. 5 | Replication of TCR with lesion strand resolution in *Mus musculus castaneus*. **a**, TCR of template strand lesions is dependent on transcription level (P15 liver, median TPM). Mutation rate estimates (circles) are the aggregate rates for expression level binned genes across CAST tumours ($n = 84$). Expression level bin 0 contains $n = 2,645$ genes, all subsequent bins contain $n = 4,323$ genes. See Methods for per-gene, per-tumour inclusion criteria. Empiric confidence intervals (99%) were calculated through bootstrap sampling ($n = 100$ replicates) of genes within the expression level bin. **b**, Comparison of mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **c**, Sequence composition normalized profiles of mutation rate around TSS loci. **d**, Stratifying the data plotted in **c** by lesion strand reveals greater detail on the observed mutation patterns, including the pronounced influence of bidirectional transcription initiation.

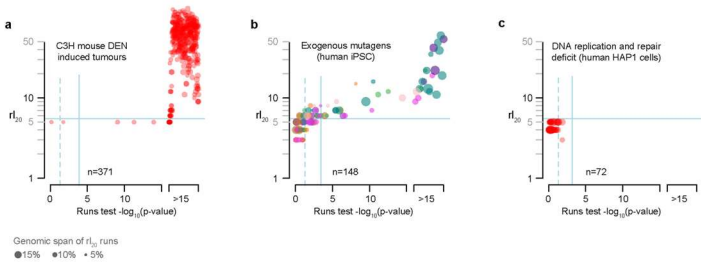


Extended Data Fig. 6 | Variant allele frequency distributions demonstrate high rates of non-mutagenic replication over segregating lesions. a–**f**, VAF distributions shown as probability density functions (total area under curve = 1) for six example tumours, calculated taking into account observed multiallelic variation. The VAF for identified driver mutations is indicated (brown triangle). Tumour identifiers are shown top right along with the percent of genomic segments (based on mutation asymmetry segmentation) that are multiallelic. Skew shows Pearson's median skewness coefficient for the VAF distributions. **a**–**c**, Tumours with no multiallelic segments and exhibit a symmetric VAF distribution showing minimal sub-clonal structure. **d**–**f**, Tumours with all segments multiallelic, illustrating the sub-clonal structure generated by segregating lesions. **g**, Tumours with a high proportion of multiallelic segments have a left-skewed VAF distribution indicating frequent non-mutagenic replication over segregating lesions. Percent of genome segments that are multiallelic (*x*-axis) plotted against VAF distribution skew for 371 C3H tumours. Tumours with low estimated cellularity indicated in pink and excluded from correlation analysis ($n = 335$ independent tumour samples in Pearson's correlation, two-sided significance from Fisher's transform). **h**, As for **g**, but showing 84 CAST tumours ($n = 77$ independent tumours included in Pearson's correlation). **i**, Mutation asymmetry summary ribbon for example C3H tumour 90797_N2; C3H genome on the *x*-axis. The percent of mutation sites with robust support for multiallelic variation (*y*-axis) calculated in 10Mb windows (grey) and for each asymmetric segment (black). Thresholds for high (black), intermediate (grey) and zero (red) rates of multiallelic sites shown on the right axis. **j**, VAF density plots for the example tumour 90797_N2 (shown in **i**) mutations in asymmetry segments stratified by the multiallelic rate thresholds defined in **i**. As with individual tumour based analysis (**a**–**h**), high multiallelic rates correspond to a leftward skew of the VAF (black, grey) whereas segments without multiallelic variation (red) show a minimally skewed distribution.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Examples of mutation patterns generated by lesion segregation from a diverse range of clinically relevant mutagens. **a**–**c**, Genome-wide mutation asymmetry plots (shown as per Fig. 2a–c) for mutagen exposed human iPSCs[5]. Cells exposed to simulated solar radiation illustrate lesion segregation for ultraviolet damage (**a**). Immediately adjacent mutations (intermutation distance 10^0) indicate CC→TT dinucleotide changes. Despite a low total mutation load (1,308 nucleotide substitutions, 842 informative T→A changes), the mutational asymmetry of lesion segregation is evident for the aristolochic acid exposed clone[5] (**b**) and the polycyclic aromatic hydrocarbon DBADE (**c**) that is found in tobacco smoke. **d**, Summary mutation asymmetry ribbons (as per Fig. 2d) for all mutagen exposed clones with $rI_{20} > 5$, which illustrates the independence of asymmetry pattern between replicate clones, almost universal asymmetry on chromosome X, and approximately 50% of the autosomal genome with asymmetry over autosomal chromosomes. The dominant mutation type is indicated for each mutagen. In those clones with low mutation rates, some sister exchange sites are likely to have been missed leading to reduced asymmetry signal (for example, on the X chromosome). Segments with <20 informative mutations are shown in white.



Extended Data Fig. 8 | Lesion segregation is evident for multiple DNA damaging agents but not for damage independent mutational processes. **a**, DEN induced C3H tumour genomes ($n = 371$) typically show significant mutational asymmetry across their genome. Wald–Wolfowitz runs test (*x*-axis) *P*-values calculated using a normal approximation (two-sided). Nominal $P = 0.05$ significance threshold indicated by dashed blue line, Bonferroni-corrected threshold shown as solid vertical blue line. *P*-values $< 1 \times 10^{-15}$ are rank-ordered. The rl_{20} metric (Fig. 5a; Methods) is shown on the *y*-axis, horizontal blue line gives emprical significance threshold of $rl_{20} > 5$. **b**, Many human iPSCs grown from single cells after exogenous mutagen exposure[5] show significant mutation asymmetry ($n = 148$ WGS, mutagen-exposed cell lines). Statistical calculations and plotting as in **a**, with adjustment of Bonferroni correction. Diverse categories of mutagen, denoted by point colour (see Fig. 5b), show asymmetry indicative of lesion segregation. **c**, Cell lines with genetically perturbed genome replication and maintenance machinery[26] and similar mutation load to those in **b** do not show significant mutation asymmetry ($n = 72$ WGS, genetically perturbed cell-lines). Statistical calculations and plotting as in **a** with adjustment of Bonferroni correction.

Extended Data Table 1 | A lesion segregation-based test for oncogenic selection

Strain	Gene	Mutation	Mutation count	Odds ratio	P-value	Known driver
C3H	Braf	6:37548568_A/T	151	2.13	5.77×10^{-6}	Yes
C3H	Hras	7:145859242_T/C	81	2.67	6.88×10^{-6}	Yes
C3H	Hras	7:145859242_T/A	65	1.02	1	Yes
C3H	Intronic Fmnl1	11:105081902_A/C	44	1.03	1	No
C3H	Intergenic	9:73125689_G/C	42	1.13	1	No
C3H	Egfr	11:14185624_T/A	34	3.87	1.23×10^{-4}	Yes
CAST	Braf	6:37451282_A/T	42	1.41	0.338	Yes

Recurrently mutated sites in C3H and CAST tumours with sufficient estimated power to detect oncogenic selection through biased strand retention analysis (required >33 C3H recurrences or >41 CAST recurrences). Odds ratio values >1 indicate the predicted correlation of driver mutation and Watson/Crick strand retention in tumours with the candidate driver mutation, but not for those without the mutation. The Fisher's exact test was performed on counts of chromosomes with Watson and Crick strand asymmetries (Methods). Each tested site was autosomal, thus total sample sizes were: $n = 2 \times 371 = 742$ for C3H, and $n = 2 \times 84 = 168$ for CAST. *P*-values (two-sided) are shown after Bonferroni correction (7 tests performed). Known driver indicates the mutation or its orthologous change has previously been implicated as a driver of hepatocellular carcinoma[6]. The CAST 6:37451282_A/T mutation is orthologous to the C3H 6:37548568_A/T mutation.