Universidad de Oviedo

Programa de Doctorado en Biomedicina
y Oncología Molecular

Development of a tool for the identification of somatic mutations and analysis of non-coding mutations in cancer

Tesis Doctoral

Ander Díaz Navarro

Junio de 2022

Universidad de Oviedo

Tesis Doctoral

Development of a tool for the identification of somatic mutations and analysis of non-coding mutations in cancer

Candidato	Directores	Tutor
Ander Díaz Navarro	Xose Antón Suárez Puente Ana Gutiérrez Fernández	José María Pérez Freije

Universidad de Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Desarrollo de una herramienta para la identificación de mutaciones somáticas y análisis de mutaciones no codificantes en cáncer	Inglés: Development of a tool for the identification of somatic mutations and analysis of non-coding mutations in cancer

2.- Autor	
Nombre: Ander Díaz Navarro	DNI/Pasaporte/NIE:
Programa de Doctorado: Biomedicina y Oncología Molecular	
Órgano responsable: Universidad de Oviedo (Centro Internacional de Postgrado)	

RESUMEN (en español)

Durante la última década, y gracias a la constitución de grandes consorcios internacionales, se han ido desentrañando las principales alteraciones moleculares presentes en los tipos de tumores más frecuentes. Conocer estas alteraciones ayuda a seguir investigando sobre los mecanismos por los que se da la desregulación celular que origina el cáncer. Sin embargo, aún queda mucho trabajo por hacer, como analizar otros tipos de tumores no tan frecuentes o conseguir que toda esta información se transforme en beneficios para los pacientes, por ejemplo, mediante un diagnóstico más preciso o el desarrollo de fármacos con menos efectos secundarios.

En este sentido, durante la presente Tesis Doctoral se ha realizado el análisis genómico más exhaustivo hasta la fecha de linfomas de célula del manto (MCL). Esto ha permitido la identificación de nuevos genes y eventos conductores de esta enfermedad. Además, también se ha observado que, mientras en otras patologías los eventos importantes en la tumorigénesis son las mutaciones puntuales, en MCL lo son las alteraciones estructurales, ya que estos tumores presentan una gran inestabilidad cromosómica. La experiencia acumulada durante la identificación de mutaciones somáticas en MCL, y otros tipos de tumores como la leucemia linfática crónica (CLL), ha llevado a desarrollar, durante esta tesis, una herramienta, llamada RFcaller, con este mismo propósito. En este sentido, RFcaller, que es más rápido e igual de preciso que los principales programas desarrollados con este fin, se basa en características en nivel de lectura junto con algoritmos de aprendizaje automático para la detección de mutaciones somáticas en muestras pareadas normal/tumor.

Por otra parte, también se ha realizado un análisis de exomas del ADN tumoral circulante presente en el líquido cefalorraquídeo (CSF) en pacientes pediátricos con meduloblastoma (MB). Esto ha confirmado que esta aproximación permite la caracterización y monitorización de este tipo de tumores. Este hallazgo es de gran importancia debido a la localización de estos tumores, que son de muy difícil acceso. De esta forma, el CSF, que se obtiene de forma rutinaria en estos pacientes, se podría utilizar para el diagnóstico y el seguimiento de la enfermedad.

Finalmente, el análisis de mutaciones en regiones no codificantes ha permitido la identificación y caracterización de *U1* como nuevo gen conductor del cáncer, así como la caracterización de una serie de mutaciones en la región 3' no codificante (3'UTR) de *NFKBIZ*. En concreto, U1 es un snRNA encargado del reconocimiento del sitio donador de splicing y que se encuentra recurrentemente mutado en MB, CLL o carcinoma hepatocelular (HCC). Tras su identificación se comprobó que las mutaciones en este gen alteraban el patrón de splicing de la célula, dando lugar a nuevas isoformas aberrantes. Por otra parte, las mutaciones detectadas en el 3'UTR de *NFKBIZ* afectan a una región altamente conservada y se ha podido demostrar que provocan un aumento en la expresión de la proteína sin alterar la estabilidad de su ARN mensajero.

Universidad de Oviedo

RESUMEN (en Inglés)

During the last decade, and thanks to the constitution of major international consortia, the main molecular alterations present in the most frequent types of tumors have been revealed. Knowledge of these alterations helps to elucidate the mechanisms by which the cellular deregulation that causes cancer occurs. However, there is still much work to do, such as analyzing other types of tumors that are not so frequent or ensuring that all this information is transformed into benefits for patients, for example, through more precise diagnosis or the development of treatments with fewer side effects.

In this regard, during the present PhD Thesis it has been possible to perform the most comprehensive genomic analysis to date of mantle cell lymphomas (MCL). This has allowed the identification of new genes and driving events of this disease. Furthermore, it was found that, although in other pathologies the main events in tumorigenesis are point mutations, in MCL they are structural alterations, since these tumors present high chromosomal instability. The experience accumulated during the identification of somatic mutations in MCL, and other types of tumors such as chronic lymphocytic leukemia (CLL), has led to the development during this PhD Thesis of a bioinformatic tool, called RFcaller, for this same purpose. In this sense, RFcaller, which is faster and as accurate as the *state-of-the-art* variant callers, is based on read-level features together with machine learning algorithms for the detection of somatic mutations in paired normal/tumor samples.

On the other hand, the whole-exome analysis of circulating tumor DNA, present in cerebrospinal fluid (CSF), in pediatric patients with medulloblastoma (MB) has confirmed that this approach is suitable for the characterization and monitoring of this type of tumor. This finding is very important due to the location of these tumors, which are very difficult to access. Thus, CSF, which is routinely obtained in these patients, could be used for diagnosis and monitoring of the disease.

Finally, the analysis of somatic mutations in non-coding regions allowed the identification and characterization of *U1* as a novel cancer driver gene, as well as a series of mutations in the 3' non-coding region (3'UTR) of *NFKBIZ*. Specifically, U1 is a snRNA responsible for splicing donor site recognition and is recurrently mutated in MB, CLL or hepatocellular carcinoma (HCC). After its identification, it was found that mutations in this gene altered the splicing pattern of the cell, giving rise to new aberrant isoforms. Moreover, the mutations detected in the 3'UTR of *NFKBIZ* affect a highly conserved region, causing an increase in protein expression without altering the stability of its messenger RNA.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN BIOMEDICINA Y ONCOLOGÍA MOLECULAR

ABBREVIATIONS

5'/3'SS	5'/3' Splice Site
AID	Activation-Induced cytidine Deaminase
AUC	Area Under the Curve
BFB	Breakage-Fusion-Bridge
BRCA	Breast Cancer
cfDNA	cell-free DNA
CLL	Chronic Lymphocytic Leukemia
CNA	Copy Number Alteration
COSMIC	Catalogue of Somatic Mutations In Cancer
CSF	Cerebrospinal Fluid
ctDNA	circulating-free DNA
DLBCL	Diffuse Large B-Cell Lymphoma
FBS	Fetal Bovine Serum
GO	Gene Ontology
GSEA	Gene-Set Enrichment Analysis
HCC	Hepatocellular Carcinoma
HRP	Horseradish Peroxidase
ICGC	International Cancer Genome Consortium
IGHV	Immunoglobulin Heavy Variable
IL1	Interleukin 1 alpha
Indel	Insertion/deletion
lncRNA	long non-coding RNA
LOH	Loss Of Heterozygosity
MAGIC	Medulloblastoma Advanced Genomics International Consortium
MB	Medulloblastoma
MCL	Mantle Cell Lymphoma
cMCL	conventional MCL
nnMCL	leukemic non-nodal MCL
miRNA	microRNA
MTC	Major Translocation Cluster
NGS	Next Generation Sequencing
PCAWG	Pan-Cancer Analysis of Whole Genomes
PladB	Pladienolide B
qPCR	quantitative PCR
RACE	Rapid Amplification of cDNA Ends
RBP	RNA Binding Protein
RNA-seq	RNA sequencing
RT-PCR	Reverse Transcription PCR
sgRNA	single-guide RNA
SHM	Somatic Hypermutation
shRNA	short hairpin RNA
snRNA	small nuclear RNA
snRNPs	small nuclear Ribonucleoprotein Particles
SNV	Single Nucleotide Variant
SSNVs	Somatic Single Nucleotide Variant
ssODN	single-stranded Oligodeoxynucleotide
SV	Structural Variant
TCGA	The Cancer Genome Atlas
TPMs	Transcripts Per Million
UTR	Untranslated Region
VAF	Variant Allele Frequency
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

CONTENTS

ABSTRACT/RESUMEN

During the last decade, and thanks to the constitution of major international consortia, the main molecular alterations present in the most frequent types of tumors have been revealed. Knowledge of these alterations helps to elucidate the mechanisms by which the cellular deregulation that causes cancer occurs. However, there is still much work to do, such as analyzing other types of tumors that are not so frequent or ensuring that all this information is transformed into benefits for patients, for example, through more precise diagnosis or the development of treatments with fewer side effects.

In this regard, during the present PhD Thesis it has been possible to perform the most comprehensive genomic analysis to date of mantle cell lymphomas (MCL). This has allowed the identification of new genes and driving events of this disease. Furthermore, it was found that, although in other pathologies the main events in tumorigenesis are point mutations, in MCL they are structural alterations, since these tumors present high chromosomal instability. The experience accumulated during the identification of somatic mutations in MCL, and other types of tumors such as chronic lymphocytic leukemia (CLL), has led to the development during this PhD Thesis of a bioinformatic tool, called RFcaller, for this same purpose. In this sense, RFcaller, which is faster and as accurate as the *state-of-the-art* variant callers, is based on read-level features together with machine learning algorithms for the detection of somatic mutations in paired normal/tumor samples.

On the other hand, the whole-exome analysis of circulating tumor DNA, present in cerebrospinal fluid (CSF), in pediatric patients with medulloblastoma (MB) has confirmed that this approach is suitable for the characterization and monitoring of this type of tumor. This finding is very important due to the location of these tumors, which are very difficult to access. Thus, CSF, which is routinely obtained in these patients, could be used for diagnosis and monitoring of the disease.

Finally, the analysis of somatic mutations in non-coding regions allowed the identification and characterization of *U1* as a novel cancer driver gene, as well as a series of mutations in the 3' non-coding region (3'UTR) of *NFKBIZ*. Specifically, U1 is a snRNA responsible for splicing donor site recognition and is recurrently mutated in MB, CLL or hepatocellular carcinoma (HCC). After its identification, it was found that mutations in this gene altered the splicing pattern of the cell, giving rise to new aberrant isoforms. Moreover, the mutations detected in the 3'UTR of *NFKBIZ* affect a highly conserved region, causing an increase in protein expression without altering the stability of its messenger RNA.

Durante la última década, y gracias a la constitución de grandes consorcios internacionales, se han ido desentrañando las principales alteraciones moleculares presentes en los tipos de tumores más frecuentes. Conocer estas alteraciones ayuda a seguir investigando sobre los mecanismos por los que se da la desregulación celular que origina el cáncer. Sin embargo, aún queda mucho trabajo por hacer, como analizar otros tipos de tumores no tan frecuentes o conseguir que toda esta información se transforme en beneficios para los pacientes, por ejemplo, mediante un diagnóstico más preciso o el desarrollo de fármacos con menos efectos secundarios.

En este sentido, durante la presente Tesis Doctoral se ha realizado el análisis genómico más exhaustivo hasta la fecha de linfomas de célula del manto (MCL). Esto ha permitido la identificación de nuevos genes y eventos conductores de esta enfermedad. Además, también se ha observado que, mientras en otras patologías los eventos importantes en la tumorigénesis son las mutaciones puntuales, en MCL lo son las alteraciones estructurales, ya que estos tumores presentan una gran inestabilidad cromosómica. La experiencia acumulada durante la identificación de mutaciones somáticas en MCL, y otros tipos de tumores como la leucemia linfática crónica (CLL), ha llevado a desarrollar, durante esta tesis, una herramienta, llamada RFcaller, con este mismo propósito. En este sentido, RFcaller, que es más rápido e igual de preciso que los principales programas desarrollados con este fin, se basa en características en nivel de lectura junto con algoritmos de aprendizaje automático para la detección de mutaciones somáticas en muestras pareadas normal/tumor.

Por otra parte, también se ha realizado un análisis de exomas del ADN tumoral circulante presente en el líquido cefalorraquídeo (CSF) en pacientes pediátricos con meduloblastoma (MB). Esto ha confirmado que esta aproximación permite la caracterización y monitorización de este tipo de tumores. Este hallazgo es de gran importancia debido a la localización de estos tumores, que son de muy difícil acceso. De esta forma, el CSF, que se obtiene de forma rutinaria en estos pacientes, se podría utilizar para el diagnóstico y el seguimiento de la enfermedad.

Finalmente, el análisis de mutaciones en regiones no codificantes ha permitido la identificación y caracterización de *U1* como nuevo gen conductor del cáncer, así como la caracterización de una serie de mutaciones en la región 3' no codificante (3'UTR) de *NFKBIZ*. En concreto, U1 es un snRNA encargado del reconocimiento del sitio donador de splicing y que se encuentra recurrentemente mutado en MB, CLL o carcinoma hepatocelular (HCC). Tras su identificación se comprobó que las mutaciones en este gen alteraban el patrón de splicing de la célula, dando lugar a nuevas isoformas aberrantes. Por otra parte, las mutaciones detectadas en el 3'UTR de *NFKBIZ* afectan a una región altamente conservada y se ha podido demostrar que provocan un aumento en la expresión de la proteína sin alterar la estabilidad de su ARN mensajero.

INTRODUCTION

Evolution is the process by which organisms change over time and is driven by natural selection and other phenomena such as genetic drift[1] or gene flow[2]. The idea of evolution was first coined by Charles Darwin in his book *On the Origin of Species*, where he referred to this process as "descent with modification"[3]. Although most people attribute this theory solely to Darwin, his contemporary, the naturalist Alfred Russel Wallace, also contributed to its development[4]. Both came to the same conclusions: i) species can change over time, ii) new species come from pre-existing ones and iii) all species share a common ancestor. They proposed natural selection, also known as "survival of the fittest", as a mechanism for evolution. In this theory, evolutionary change comes from those individuals with a higher reproductive rate because they present more favorable traits for surviving in the environment they live in, being these characteristics inherited[3].

Despite the revolutionary consequences that this hypothesis represented for our understanding of living organisms, and the overwhelming data supporting it, the mechanisms by which this gradual change is generated over time was something that escaped Darwin and Wallace knowledge. Nowadays we know that it is due to mutations in DNA. These variations, that appear naturally due to DNA replication or exposure to external agents, generate the diversity upon which evolution acts. Some of these variants might confer an advantage under certain environmental conditions, which favors transmission to offspring at an increase rate, promoting genetic selection[5]. It may also happen that mutations have a deleterious effect and although they cause a genetic disorder, they are also considered an agent of evolution[6]. However, this will depend on where the mutations occur. In this sense, if they appear in germline cells (i.e., eggs or sperm cells) they can be passed to the progeny, constituting germline mutations, and even get the chance to be fixed in the population, representing polymorphic variants. On the other hand, somatic mutations usually exceed the number of germline mutations, as they appear in any other cell of the organism and are not passed along to subsequent generations[7].

Mutations can be classified in two large categories depending on their size. If the alterations only affect a specific region they are called “point mutations”. These are the most frequent and include either substitutions or single nucleotide variants (SNVs), when a single base changes to a different base, and small insertions or deletions (indels), when there is an insertion or deletion of a few bases in the DNA sequence. On the other hand, structural variants (SVs) are less common but they usually affect larger chromosomal regions, and include inversions, deletions, duplications or translocations (Figure 1)[5,8].

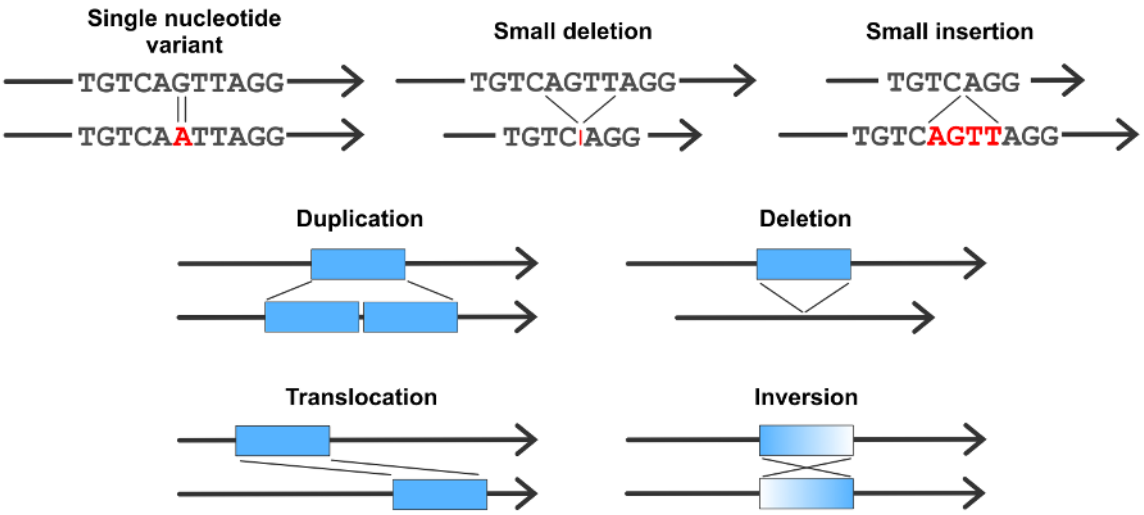

Figure 1. Main types of mutations. Single nucleotide variants (SNVs), and small deletions and insertions (indels) are point mutations. Duplications, deletions, translocations and inversion events imply large chromosomal rearrangements.

Although DNA replication is highly accurate, resulting in a very low mutation rate[9,10], the large size of the genome results in the accumulation of a considerable number of somatic mutations accumulate throughout the life of an organism[11,12]. This means that the chance of developing a disease, like cancer, increases with age and it is linked to the number of stem cell divisions[12,13]. In fact, about 66% of mutations associated with the appearance of tumors appear spontaneously during normal DNA replication[14].

Cancer: a race for evolution

Cancer is a disease in which some cells escape the mechanisms of growth control and acquire the ability to invade other tissues. During the lifetime of an organism, cells proliferate, resulting in the accumulation of mutations over time in healthy tissues[15–17]. Most of these mutations occur in non-functional regions of the genome, or even in certain genes, but without affecting cell function, and are therefore neutral from an evolutionary point of view. However, other mutations accumulate in certain genes or functional elements, resulting in the acquisition of specific features driving the process of neoplastic transformation[18]. Similar to the Darwinian theory of evolution, the tumorigenic process is driven by somatic evolution, in which cells with certain mutations will have a growth advantage over non-mutated cells and will be chosen by natural selection (Figure 2)[19–21]. These capabilities, that include the ability to ignore growth suppressors, modify the local microenvironment, evade immune destruction, spread to other organs and allow epigenetic reprogramming, are known as *Hallmarks of Cancer* and were collected by Douglas Hanahan and Robert Weinberg in 2000[22] and updated with enabling characteristics in 2011 and 2022[23,24].

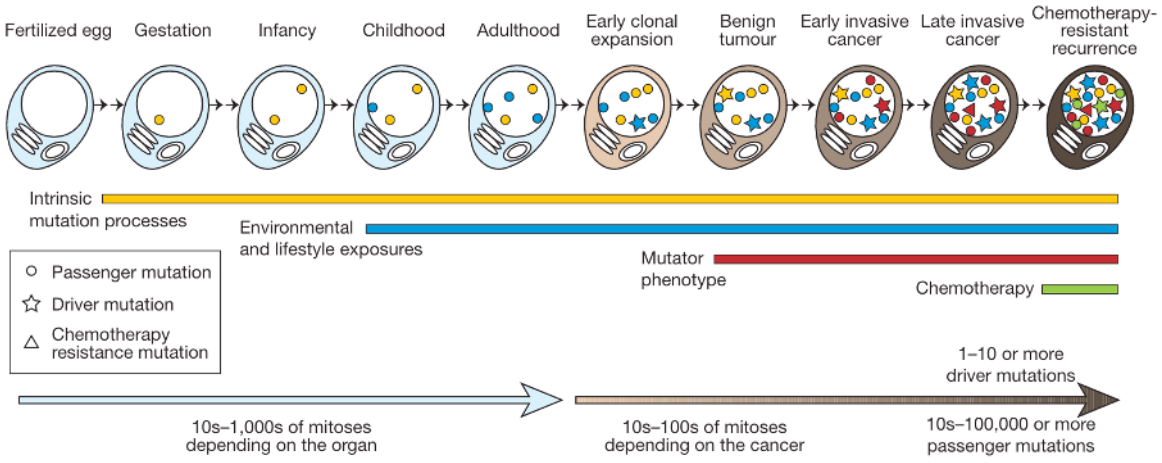

Figure 2. Representation of the acquisition of mutations in cells and the agents involved in the tumorigenic process. During mitotic divisions, mutations accumulate naturally due to DNA replication or exposure to external agents. Passenger mutations have no effect on the cell, but driver mutations cause the clonal expansion that gives rise to the tumor. In addition, mutations that confer resistance to cancer therapies can also occur, leading to relapses. From Stratton et al. 2009[19].

Depending on the ability of a mutation to confer some of these capabilities, they are classified in two major groups: driver or passenger mutations. Thus, driver mutations provide a selective growth advantage that promotes tumor progression and in general, all tumor cells will inherit these driver mutations. In contrast, passenger mutations are those that do not affect any relevant process to cancer development. Although they do not confer a growth advantage, if present in a cell that acquires a driver mutation, those non-functional mutations will be also passed to daughter tumor cells, so they will be selected in the tumor, but they are considered passengers, because the tumor is driven by the driver mutations[19]. Moreover, genes affected by driver mutations can be classified into: i) oncogenes, if the mutation causes their activation or gives them a new function or ii) tumor suppressor genes, if their inactivation favors tumorigenesis[25]. However, this classification is not so easy, since there are genes with both oncogenic and tumor suppressor functions depending on the cellular context[26]. For example, the Wilms' tumor 1 gene (*WT1*) encodes to a transcription factor with a tumor suppressor activity in kidney tumors, whereas it has an oncogenic role in other pathologies like breast cancer or different leukemias[27]. Similarly, *NOTCH1* acts as an oncogene in leukemias[28,29] but is considered a tumor suppressor in epithelial tumors[30,31]. There are also genes like *TGF-β* that exhibit this dual role depending on the stage of the tumor[32].

Cancer genomes: next-generation sequencing

More than 20 years have passed since the publication of the first draft of the human genome in 2001[33,34]. However, only recently it has been possible to complete the sequencing of a whole human genome that included even the most complex regions[35]. Throughout this period, and thanks to the development of massive sequencing technologies, the process of sequencing a genome has evolved from several years to just a few days. In combination to the considerable reduction in time, the progressive reduction in the cost of sequencing has allowed the pursue of large-scale genomic projects[36–38]. Thanks to these studies, we have been able to understand that the mutational burden along the genome is affected by epigenetic features or the nucleotide composition[39–42], despite it was long considered an stochastic process[43]. In addition, other factors have been found that can also alter the DNA sequence.

Since the first cancer-causing mutation in humans was identified, a single nucleotide change in the *HRAS* gene[44], the search for genetic alterations that might be involved in cancer initiation has constituted a priority in oncological research[45]. In this regard, this process has benefited enormously from the emergence of next-generation sequencing (NGS) approaches, either whole-genome sequencing (WGS) or whole-exome sequencing (WES), techniques that allow the study of cancer in an unbiased manner and genome-wide. These techniques allow the detection of most types of alterations present in tumors, from point mutations to large structural variants and copy number alterations (CNAs). However, the large amount of information obtained from these analyses, thousands of mutations for each patient, represented a challenge in the interpretation of the first cancer genomes, as the simple detection of somatic mutations was not enough to distinguish between driver and passenger mutations[25].

The most common strategy to determine whether a gene is a cancer driver is based on the frequency of somatic mutations that affect this gene. If the frequency is higher than what would be expected by chance considering the background mutation rate, the gene is labeled as driver[46,47]. This indicates that mutations in these genes are likely to confer a selective advantage to the cells that carry them, favoring tumor transformation[25]. Nonetheless, it is possible that a driver gene is missed if the number of analyzed tumors is too small to have enough statistical power to detect it. Driver genes can be further classified into oncogenes or tumor suppressor genes by applying the "20/20 rule". If > 20% of the mutations cause an amino acid change (missense) at a recurrent position (i.e. gain of function) it is considered an oncogene, whereas if > 20% of the mutations are inactivating (i.e. loss of function) it can be considered a tumor suppressor[48] (Figure 3).

Although the previous approach is the most commonly used for the identification of driver genes, it has some limitations that are important to keep in mind. These methods are based on the background mutation rate, however this is affected by GC content, gene density, replication time and nucleosome occupancy among other factors, so not all regions of the genome will have the same mutational ratio.

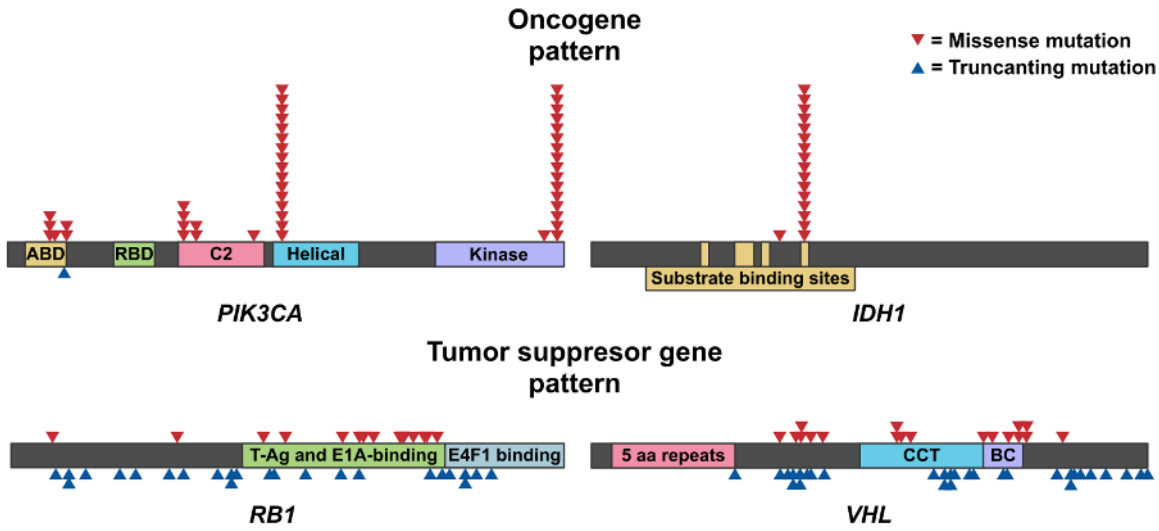

Figure 3. Distribution of mutations in two oncogenes (*PIK3CA* and *IDH1*) and two tumor suppressor genes (*RB1* and *VHL*). Oncogenes usually have activating point mutations in specific regions, whereas tumor suppressor genes show truncating mutations throughout the gene. Adapted from Vogelstein et al. 2013[48].

Furthermore, the background mutation rate is calculated considering synonymous mutations and those located in introns or untranslated regions (UTRs), since it is assumed that most of them do not have an impact on the cell[49], although it is now known that this is not entirely correct. The other aspect to consider is the large number of samples that is needed to have enough statistical power to detect rare driver genes. Most cancer genomic studies so far have found that most driver genes identified are only present at low frequency[29,36], suggesting that there are many rare driver genes to be identified.

Complementary to frequency-based approaches are function-based approaches. These strategies aim to find driver genes based on the impact of mutations on protein functionality, for example, by looking at the localization of mutations in the three-dimensional structure of proteins. In this sense, distinct mutations along the gene can be located in the same region of the protein, such as the active center of an enzyme[50]. Another function-based approach is to observe whether mutations affect highly conserved sequences in evolution, which would highlight the importance of these regions[51].

Major efforts: international consortia

The first large effort to identify the genomic alterations driving tumor development and progression was possible thanks to the efforts of two major international cancer consortia: The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)[37,38]. The reasons for the creation of these consortia were clear: to avoid duplication of efforts, to standardize analyses, to allow comparison between datasets and to favor the rapid dissemination of data, methods and results to the scientific community. In addition, this collaboration also allows the analysis of a larger number of tumor samples, enhancing the ability to detect and analyze tumor alterations.

Over the past decade, these multinational and collaborative initiatives have comprehensively characterized somatically acquired genetic events in at least fifty classes of cancer, including those with the highest incidence and mortality. In this way, each team, specialized in one type of cancer, has reported the most frequent driver alterations identified in each tumor, from point mutations to genomic rearrangements. The first publication of this consortium was published only three years after its constitution, with the results of the analysis of four chronic lymphocytic leukemia (CLL) genomes. Although the analysis was performed on a small number of samples, recurrent mutations in genes such as *NOTCH1*, *XPO1*, *KLHL6* or *MYD88* were identified, and allowed the classification of patients according to genomic drivers and immunoglobulin genes (IGHV) mutational status[52]. By extending the analysis to more than 500 patients, it was possible to show the portrait of the CLL genomic landscape, identifying novel driver genes and recurrent non-coding driver mutations[29]. In parallel, other groups shown their results for other types of tumors. For example, the analysis of 560 breast cancer (BRCA) whole-genomes allowed to identify 12 base substitution mutational signatures and six rearrangement signatures involving 93 genes[53]. Following the same approach, the analysis of 360 cases of hepatocellular carcinoma (HCC) by WES and other techniques such as DNA methylation, RNA and proteomic expression, identified three HCC subtypes, new significantly mutated genes and potential therapeutic targets[54].

These advances demonstrated not only the power of NGS approaches for tumor characterization, but also that a large number of samples was needed to detect less frequent alterations involved in the development of different types of cancer. Therefore, in order to obtain a more global view of cancer, of the mechanisms involved in its development and to analyze simultaneously all the data published to date, the pan-cancer analysis of whole genomes (PCAWG) was carried out. In this regard, the PCAWG Consortium conducted a meta-analysis of genomic features across tumor types, collecting data from more than 2,500 donors[36].

The analysis of this number of whole genomes, together with the heterogeneity of tumors studied, has not been an easy task, requiring great efforts both in time and in computational resources. The analysis of more than 2,000 tumor genomes resulted in the identification of more than 43 million somatic SNVs (SSNVs) and almost 2.5 million indels (Figure 4), which required the implementation of a new strategy to separate between passenger and driver mutations. Thus, new driver genes were found and it was confirmed that many of the tumor suppressor genes are affected by double-hit inactivation[36]. Using this same set of mutations, new mutational signatures to those previously described by the Catalogue Of Somatic Mutations In Cancer (COSMIC-v2; https:// cancer.sanger.ac.uk/cosmic/signatures_v2) were identified[53,55]. Many of these signatures are of biological origin and appear to a greater or lesser extent in all tumor types, but it has also been possible to identify mutational signatures associated with different treatments, mutations in specific genes or mutagenic agents such as UV light or tobacco, that are overrepresented in some tumor types[56].

This information, generated from these large consortia, has provided great knowledge about the altered pathways in cancer and their evolution. However, it is also allowing other groups to work on these data from another point of view, which is contributing to transfer results to the clinic, one of the ultimate goals of this collective effort. Thus, for example, from RNA sequencing (RNA-seq) expression data it has been possible to perform survival analyses for all types of cancer, identifying key genes and characteristics that support this prediction[57,58]. This could be useful when selecting genes against which to develop drugs, achieving more personalized treatments depending on the tumor type.

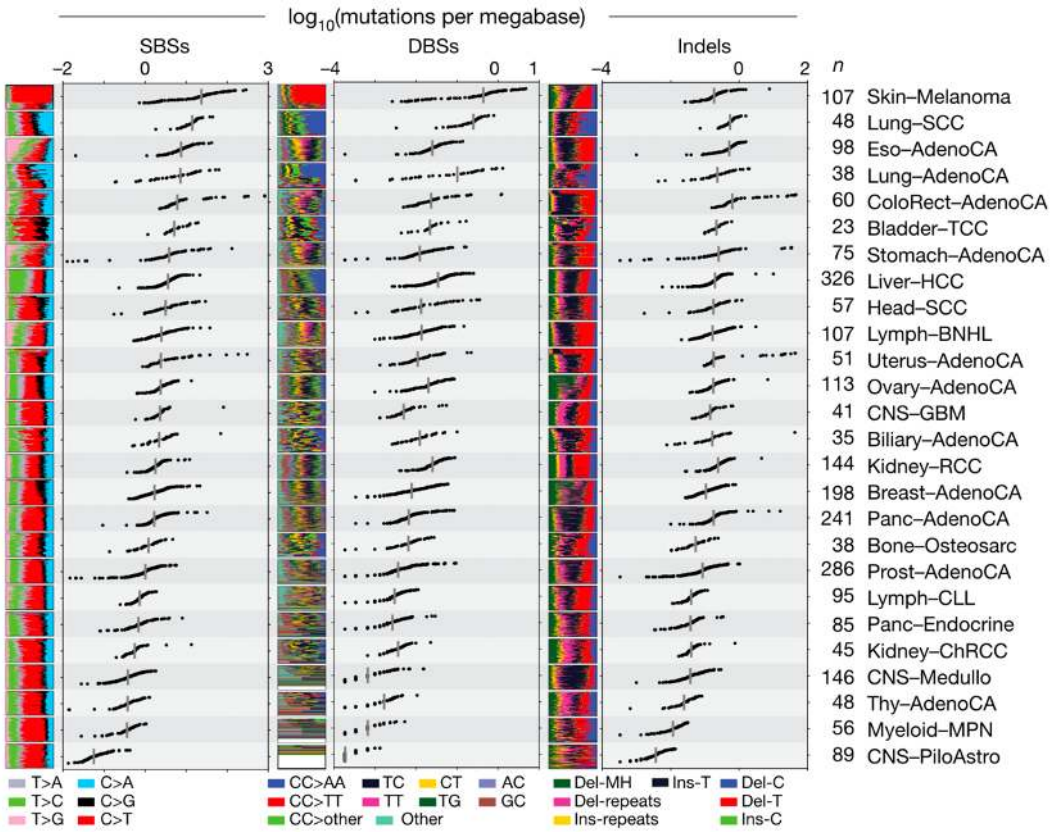

Figure 4. Mutation burdens on single-base substitutions (SBSs), doublet-base substitutions (DBS) and small indels across PCAWG tumor types. Each dot represents the number of mutations per megabase for one patient. To the left of each distribution is shown the proportion of each change and to the right the number of samples for each tumor type. From Alexandrov et al. 2020[56].

To this end, the Cancer Target Discovery and Development Network is in charge of defining the criteria used to determine which drivers have a potential for drug development against them, prioritizing those that play a key role in tumor initiation, maintenance and metastasis[59]. An example of this type of strategies is trying to block the activity of an oncogene, which will lead, in some cases, to cell death due to a phenomenon known as oncogene addiction[60]. Within this group we can find drugs such as Imatinib, for the treatment of chronic myeloid leukemia and whose target is the *BCR-ABL* fusion gene[61], or Trastuzumab, for *ERBB2* in breast cancer[62]. Another strategy is to inhibit genes whose activity is essential if the tumor has mutations in other oncogenes or tumor suppressor genes, as in the case of PARP inhibition when the tumor has mutations in *BRCA1/2*[62,63]. Finally, a promising approach, which could reduce the side effects of these treatments, is to identify protein-protein interactions that only occur in tumors due to mutations in one of the genes involved in this *de novo* interaction[60].

Non-coding mutations: new insights

The completion of these PanCancer studies provided a general view of cancer from a genomic perspective. One of the main findings is that, on average, each tumor has 4.6 driver mutations, being the most common the presence of somatic copy number alterations as well as coding mutations[36]. However, a surprising fact is that more than 5% of cases do not contain mutations in any driver gene. These findings suggest that, despite the enormous effort to understand specific alterations in genes or pathways that contribute to tumor development, some driver events might be present in different regions of the genome still unexplored.

In this regard, it must be noted that most studies performed during the first years of cancer genomics have been biased towards mutations and alterations in protein-coding regions. This is partly due to the techniques employed, such as exome sequencing[64,65], and to the high costs of WGS, that resulted in a limited number of samples analyzed by this approach, preventing the reach of enough statistical power to identify and to understand the effect of mutations in non-coding regions[66,67]. The change in perception occurred with the identification of mutations in the promoter region of the telomerase reverse transcriptase gene (*TERT*), found in more than 70% of melanoma cases, a fact that had gone unnoticed when the first melanoma genomes were analyzed[68]. This led to the reanalysis of more than 700 tumors of different types, with the finding of mutations in the *TERT* promoter in 43% of central nervous system tumors, 10% of malignant thyroid tumors and 59% of bladder cancer, resulting in an increased expression of *TERT*[69].

Following this discovery, mutations in regulatory and other non-coding regions began to be perceived as driver events in cancer. Thus, in 2014 Lee et al. carried out the first genome-wide analysis of non-coding mutations. This work demonstrated that promoter regions, enhancers and 3'/5'UTRs have a mutational frequency similar to exons, suggesting that they participate in gene regulation. On the other hand, the mutational frequency of intergenic regions is much higher, possibly because they are subject to less selective pressure. However, because hotspots of non-coding driver mutations are more infrequent and tumor type-specific[70], the analysis of only 800 samples of different types of cancer was unable to find relevant non-coding mutations[67].

After this proof of concept, which demonstrated that non-coding regions could play a role in tumorigenesis, different groups set out to analyze these regions in different types of specific tumors. Thus, it was found that in skin cancer, in addition to *TERT*, the promoter of *DPH3* is mutated, leading to increase its expression *in vitro*[71] and confers greater migratory ability in mice[72]. On the other hand, two types of mutational hotspots in non-coding regions were found in CLL, and whose mechanisms were different from those previously reported. Thus, recurrent mutations were detected in an enhancer of *PAX5*, located 330 kilobases upstream of the gene. The other hotspot was located in the 5'UTR region of *NOTCH1*[29], a gene that had already been identified as a driver for this disease[52]. Specifically, these new mutations in *NOTCH1* generate a new splicing acceptor site causing the loss of the last 53 amino acids of the protein. Thus, the effect achieved was the same as with the mutations located in the coding region, the loss of the PEST regulatory domain, increasing the stability of the protein[29]. A final example is found in long non-coding RNAs (lncRNAs) or microRNAs (miRNAs), RNA genes that do not code for protein but have regulatory functions. In this case, by analyzing samples from 300 patients with liver cancer, they found mutations in *MALAT1*[73], a lncRNA whose expression had been associated with increased metastatic capacity[74].

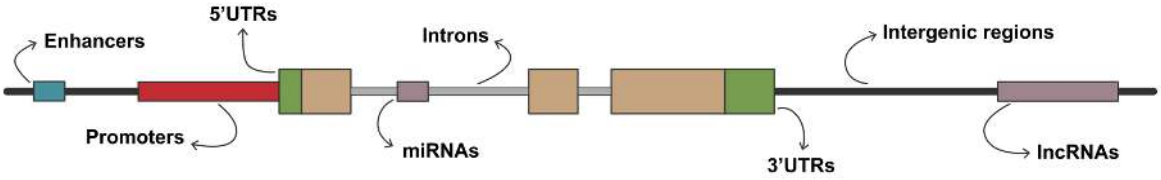

Figure 5. Non-coding regions of interests. UTRs, untranslated regions; miRNAs, microRNAs; lncRNAs, long non-coding RNAs.

All these efforts to detect new non-coding mutations and explain their mechanisms of action have allowed the development of new strategies to facilitate their identification. Thus, for the analysis of non-coding mutations, the PCAWG has focused on regulatory regions (enhancers, promoters, 3'/5'UTRs), RNA genes (long/short non-coding RNAs and microRNAs), as well as intergenic regions and structural variations that affect non-coding regions (Figure 5). This has made possible to identify new non-coding driver mutations, such as mutations in the 5'UTR of *MTG2*, which could be affecting its

expression levels[75]; in the 3'UTR of *NFKBIZ*, where it has been detected a mutational hotspot affecting a highly conserved region that could affect both the stability of the mRNA[76] and its post-transcriptional regulation[77]; mutations affecting the transcription start site and the donor splice site of the first non-coding exon of *TP53*; or structural variants involving the 10p15 arm associated with the overexpression of a family of nearby genes (*AKR1C1*, *AKR1C2* and *AKR1C3*)[70].

Alternative splicing: regulatory mechanisms

Even with all this new knowledge, it is very likely that there are non-coding mutations that are being misinterpreted. For example, when mutations are located within the coding sequence, it is very straightforward to identify the functional consequence in terms of amino acid change. In the case of evolutionary conserved regions such as promoters or even 3'/5'UTRs, a mutation affecting a highly conserved residue is likely to cause an impact in its function. However, mutations in introns, unless they occur at very specific motifs such as the splicing recognition sites, the poly-pyrimidine track or the branch point, are very difficult to predict whether they might have any effect[78]. Therefore, the problem is not only the identification of mutations, but the interpretation of the potential functional consequences. The case of introns and mutations that can affect splicing is especially complex, since it is a mechanism regulated by many elements, and not only dependent on the local context in which the mutation appears.

The mechanism of splicing, both the recognition of sequences involved and the catalysis of the reactions, is carried out by a ribonucleoprotein complex known as the spliceosome (Figure 6a). This complex is formed by small nuclear RNAs (snRNAs) and associated proteins, which together form the ribonucleoprotein particles (snRNPs). The most important components of the spliceosome are the U1 snRNP, that contains the snRNA U1 which is in charge of recognizing the 5' splice site (5'SS), the splicing factor 1 (SF1) and the U2AF that recognize the 3' splice site (3'SS), and the U2 snRNP, formed by the snRNA U2 and the splicing factor SF3B1, in charge of binding to the branch point, an adenosine located 18-35 nucleotides upstream of the 3'SS.[79].

The fact that the spliceosome is able to recognize the canonical splicing site and not another, an event known as alternative splicing, is due to cis-regulatory sequences and trans-regulatory proteins that bind to these sequences (Figure 6b). Thus, surrounding a splicing site, there are elements that promote or inhibit splicing such as exon splicing enhancers recognized by the SR family of proteins o exon splicing silencers recognized by hnRNPs. In case these motifs are in introns, they are called intron splicing enhancers or intron splicing silencer. The balance between these promoting and inhibitor signals determines the relative abundance of the different isoforms of the transcripts generated by a gene[80].

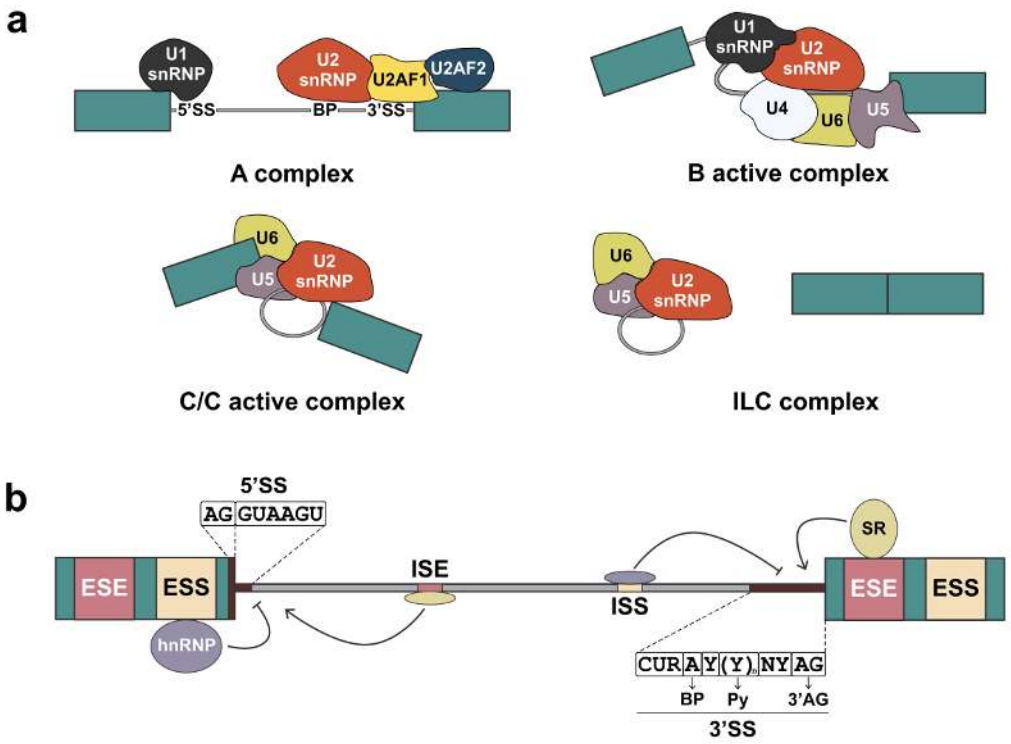

Figure 6. Splicing process and its regulatory elements. a) Summary of the splicing process. First, in the A complex, U1 snRNP binds to the 5'SS by complementarity of the U1 snRNA. This union promotes the recognition of the branch point (BP) and the 3'SS by the U2 snRNP and U2AF proteins respectively. The process continues with the recruitment of U4, U5 and U6 proteins, resulting in the B active complex. Then, the C active complex is assembled, which allows intron cleavage as it constitutes the catalytic form of the spliceosome. Finally, the ILC complex removes the intron. b) Cis-regulatory sequences like exonic splicing enhancers/silencers (ESE, ESS) are recognized by SR or hnRNP family proteins respectively to promote or inhibit the splicing. These elements can also be found in introns as intronic splicing enhancers/silencers (ISE, ISS). 5' splice site (5'SS) consists of the last two bases of the exon and the first six bases of the intron, whereas 3' splice site (3'SS) is formed by the BP, a polypyrimidine track (Py) and AG dinucleotide in the last two bases of the intron. Adapted from Will and Lührmann 2011[79] and Fu and Ares 2014[80].

Although the presence of these regulatory elements in both exons and introns is known, deciphering the effect of mutations in these regions is not easy, since we often tend to focus only on the effect they have at the peptide level. For example, a couple of missense mutations have been found in exon 12 of the *CFTR* gene, which instead of causing an amino acid change, generate an exon skipping by altering an exon splicing regulatory element[81]. Another clear example can be found in *APC*, in which missense and even synonymous mutations have also been described[82]. Or the mutation causing Hutchinson-Gilford Progeria, where a synonymous mutation in the gene encoding lamin A results in the creation of a splice donor site that removes part of the protein preventing its normal maturation[83]. However, due to the large number of mutations present in tumor genomes, synonymous mutations are usually filtered out, what would miss *bona fide* pathogenic mutations as those described above.

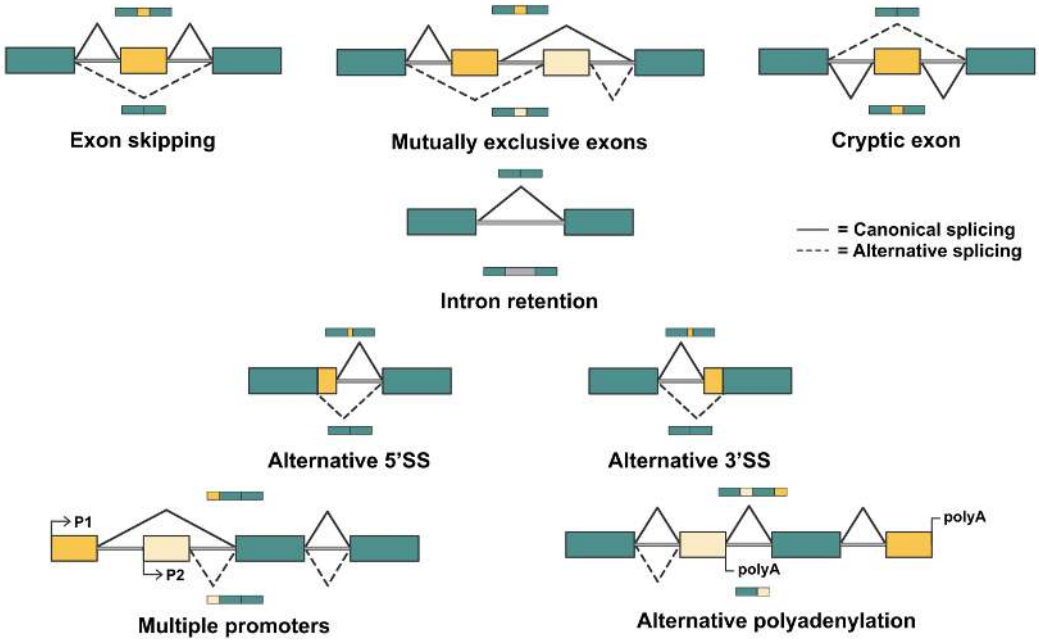

Figure 7. Most common alternative splicing events. Exon skipping. An exon is excluded from the mRNA; Mutually exclusive exons. It involves the selection of one or the other exon, they will not be found together in the mRNA. Cryptic exon. A new exon is incorporated into the transcript. Alternative 5'/3'SS. They represent exon modification events. Multiple promoters and alternative polyadenylation. mRNA isoform may change depending on promoter usage or alternative polyadenylation regulation. Canonical isoform is represented above the splicing event and alternative isoform below it. SS, splice site. Adapted from Matlin et al. 2005[88].

Programs such as Variant Effect Predictor (VEP)[84] are often used to annotate the effect of mutations. Nevertheless, mutations that alter splicing are very difficult to predict, and the easiest way to do so is to obtain WGS/WES and RNAseq data from the same patient, which is not always possible. In order to determine whether a mutation is affecting splicing, from genomic level data, new tools have been developed that allow the prediction of splicing from the pre-mRNA sequence[85]. In this way, the whole transcript is analyzed, so all the regulatory elements present can be considered. Thus, it has been feasible to identify deep-intron mutations that actually cause the use of alternative donor and acceptor sites, intron retention, exon skipping or the loss of an exon fragment (Figure 7)[86,87].

Mutation discovery: Variant callers

The rapid introduction of NGS for the study of cancer genomes required the development of analytical pipelines for the identification of somatic mutations in tumor-normal paired samples. These mutation callers provide a statistical framework to define the set of somatic mutations present in a tumor sample, which are then used to define driver mutations or mutational signatures to be considered in the clinic. Therefore, the performance of these mutation callers is of outmost importance for downstream analyses. If the sensitivity of the caller is low, many real mutations will be missed (false negatives) preventing an accurate diagnosis. In contrast, if the specificity is low, artefacts, sequencing errors or polymorphisms will be labeled as somatic mutations, reducing the utility and confidence in the data.

During the last decade, different approaches have been developed for the analysis of cancer genomes, resulting in the generation of multiple mutation callers. Most *state-of-the-art* variant callers are based in traditional statistical methods, such as Sidrón[29], CaVEMan[89], MuTect2[90], MuSE[91], Strelka2[92], Pindel[93] or SMuFin[94] among others. However, there is no consensus on the mutations detected by each caller, with a large number of private calls specific for each method. These differences are mainly due to the ability of each program to deal with the tumor heterogeneity and purity,

normal contamination, sequencing and mapping artifacts, coverage, as well as different downstream filtering steps[95]. Due to the advantages of some pipelines to detect specific *bona fide* mutations, some collaborative projects such as the PCAWG[36] or the TCGA PanCancer Atlas MC3[96], do not use a single caller but a combination of algorithms, keeping the intersection between them as the set of mutations that is more reliable. Despite the utility of this multi-pipeline approach to generate a consensus set of mutations, this strategy has a very large computational cost, demanding large servers and consuming many hours for the analysis of a single case.

However, since this technology is aimed at clinical diagnosis, the analysis requires enough sensitivity and specificity but also to be obtained in a reasonable amount of time. To increase accuracy, a final step of manual review through visual inspection is usually carried out for mutations that might be clinical informative. This manual revision increases the specificity, but at the cost of a labor-intensive process. Therefore, if the number of detected mutations is too high due to the use of an unspecific method, the effort involved in this last step would be disproportionate.

In order to simplify this process, new programs based on emerging technologies such as machine learning (Figure 8) are being developed in recent years. This field, which belongs to a branch of artificial intelligence, has two main approaches: supervised and unsupervised classification. The difference between both approaches is that for supervised training, the user, in addition to the features with which to train the program, also passes the result (whether a mutation is real or a false positive, based on validated data or an expert review). In contrast, for unsupervised training it is the program itself, based on the initial data, that tries to make the classification. Both strategies usually use the same algorithms, among which are: linear regression, decision trees, kNNs or random forest.

Some of the programs that have emerged from these strategies are designed to refine a previously extracted set of mutations using several variant callers, which does not solve the problem of computational cost and analysis time[97–99]. More recently, however, there have been new programs based on machine learning approaches[100,101]

or neural networks[102,103] that are capable of detecting mutations directly. In spite of this, most of these programs have been trained with high depth of coverage WES using *in silico*[99,100] or orthogonal validated mutations[101] and cannot be used for whole-genome analysis. The ability to train machine learning models to perform human tasks such as reviewing mutation calls represents an opportunity to refine these time-consuming tasks and generate higher quality reports which can be of great importance in the introduction of WGS in the clinical practice.

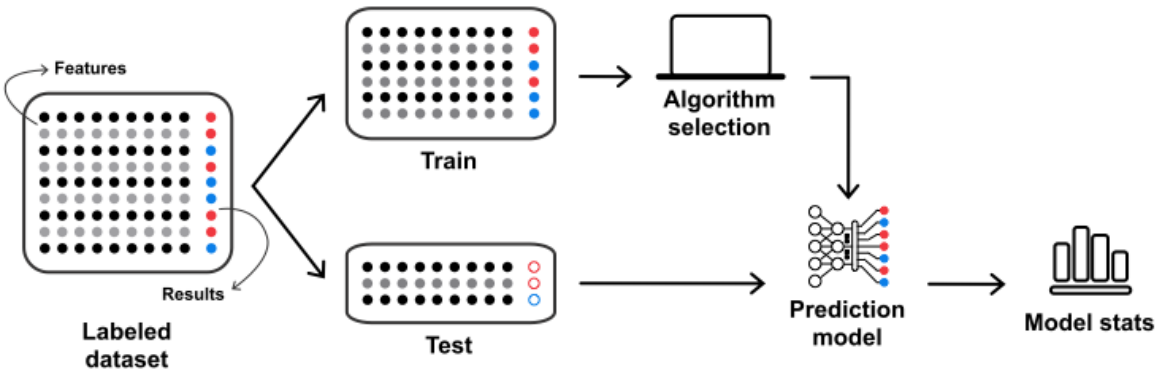

Figure 8. Workflow to train a supervised machine learning algorithm. The main thing to train this kind of algorithms is to have a good dataset with the appropriate features with which to train the algorithm and the actual value for each of the inputs. Then, the dataset is split in two to be able to train and test the accuracy of the algorithm with different inputs. Although this is sufficient for the first steps of generating a machine learning algorithm, it is necessary to run the algorithm with a second completely independent dataset to really validate its accuracy.

OBJECTIVES

The major international consortia have allowed us to have a more complete vision of the most frequent genetic alterations that are responsible for the main types of cancer. However, despite these advances, there are still cases in which no driver event has been detected that could explain the tumor origin. In this regard, there are still poorly studied regions, such as non-coding or repetitive regions, which could harbor new driver genes. In addition to the analysis of these cases in depth, it is important to focus on the study of other less frequent types of tumors, taking advantage of the tools and methods developed throughout these years. Thus, the ultimate goal of all this effort is no other than better prognosis and better personalized treatments with fewer side effects for cancer patients. In this sense, to make the study of genomes a reality in clinical practice on a routine basis, the tools used in these analyses must be accessible to small and medium-sized groups without large computational capabilities. To further address these issues, the specific objectives of this thesis were:

1. To study new mutations and structural variations in patients with mantel cell lymphoma from whole-genome analysis.
2. Development of a tool for the identification of somatic mutations from tumor-normal paired samples.
3. To determine whether ctDNA from cerebrospinal fluid can be used to characterize pediatric medulloblastoma tumors.
4. To characterize a series of non-coding mutations identified in hematological neoplasias.

EXPERIMENTAL PROCEDURES

Patient cohort

Study subjects and data collection

A written consent was obtained from all patients/legal representatives who participated in these studies through individual projects. For pediatric medulloblastoma (MB), 13 patients diagnosed and treated at the Hospital Universitario Vall d'Hebron were included in the study. For mantle cell lymphoma (MCL) cohort, 61 patients were obtained from hematopathology collection registered at the Biobank of the Hospital Clinic–Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). For the study of neoepitopes in patients with CLL, DNA from 52 patients was obtained through the Hospital Universitario Central de Asturias (HUCA) and fresh serum was extracted from 11 patients (Hospital Clinic and HUCA).

Samples used in the rest of studies were from participants recruited and anonymized by individual ICGC and TCGA projects for the PCAWG project. This dataset consisting of 2,583 donors across 37 tumor types was collected from the ICGC Data Coordination Center. The use of PCAWG data was approved by the University of Toronto Research Ethics Board under RIS Human Protocol Number 30278 and protocol title "Pan-cancer Analysis of Whole Genomes: PCAWG".

In addition to the 5,166 tumor and paired normal WGS aligned BAMs retrieved from PCAWG dataset, another 84 WGS and 10 WES paired samples from CLL patients, 228 WGS paired samples from the Medulloblastoma Advanced Genomics International Consortium (MAGIC) and 299, 387 and 225 tumor RNA-seq data from CLL, HCC and MB, respectively, were collected. The use of this extra CLL genomic and clinical data was approved by the Hospital Clinic of Barcelona Institutional Review Board under protocol number HCB/2015/0814 and protocol title "Functional and Clinical Impact of Genomic Analysis in CLL".

Molecular biology methods

DNA, RNA and protein extraction

For pediatric medulloblastoma project, tumor DNA was extracted from a fresh-frozen or paraffin-embedded section of the tumor biopsy using the QIAamp DNA mini kit (Qiagen) and the QIAamp DNA FFPE tissue kit (Qiagen), respectively. Germline DNA was extracted from peripheral blood cells using the QIAamp DNA mini blood kit (Qiagen). Peripheral blood was collected in K2EDTA containing tubes (Vacutainer) and plasma was acquired following a 1,600 × *g* centrifugation for 10 min. Both plasma and cerebrospinal fluid (CSF) samples were centrifuged at 3000 × *g* for 5 min and the supernatant were collected. Cell-free DNA (cfDNA) from plasma and CSF samples were extracted using the QIAamp Circulating Nucleid Acids kit (Qiagen). Genomic and cfDNA were quantified using the Qubit fluorometer.

DNA was extracted from CRISPR clones by lysing the cells with lysis buffer (200 mM Tris-HCl pH 7.4, 200 mM EDTA, 1% SDS) and precipitating with ammonium acetate and isopropanol.

Total RNA from cell lines was extracted with TRIzol reagent (Life Technologies), and then it was purified through alcohol precipitation and dissolved in DEPC water. RNA was assayed for quantity and quality (260/280 nm ratio) using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies) and Qubit RNA HS Assay (Life Technologies).

Proteins derived from the subcellular fractioning protocol were obtained following the TRIzol guidelines (Life Technologies). The rest of protein extractions were made through cell lysis in RIPA buffer for 5 min (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM Na_2EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM Na_3VO_4) supplemented with protease (cOmplete EDTA-free, Roche) and phosphatase inhibitor cocktails (PhosSTOP, Roche). Cell lysates were centrifuged at 13,000 rpm, 4 °C for 10 min. Protein concentration for all cell extracts was evaluated by the bicinchoninic acid technique (Pierce BCA Protein Assay Kit, Thermo Scientific).

Plasmid construction and site directed mutagenesis

To test the effect of mutations located in the *U1* gene, the pLKO.1-puro lentiviral vector (Addgene #8453) was modified by removing the internal U6 promoter (between *Nde*I and *Eco*RI), and replacing it with the *U1* locus, including 393 bases of its promoter, the *U1* sequence and 39 bases of 3'-flanking region using the following oligonucleotides (U1- EcoRI-Fwd: 5'-GTCGAGAATTCTTGGCGTACAGTCTGTTTTTG-3'; NdeI-U1-Rev: 5'-CTATCATATGTAAGGACCAGCTTCTTTGGGA-3'). The g.3A>C and g.3A>G mutations were introduced by PCR with the following oligonucleotides: U1-A_C-Fwd: 5'-GCCAGGTAAGGATGAGATCTTCGGG-3'; U1-A_C-Rev: 5'-CCCGAAGATCTCATCCTTACCTGGC-3'; U1-A_G-Fwd: 5'-GCCAGGTAAGCATGAGATCTTCGGG-3'; U1-A_G-Rev: 5'-CCCGAAGATCTCATCTTACCTGGC-3' in combination with the corresponding previous primers. The PCR products were digested with *Nde*I and *Eco*RI and cloned in the modified pLKO.1 plasmid.

For the analysis of *NFKBIZ* mutations, the psiCHECK2 vector with the *NFKBIZ*-3'UTR previously cloned downstream of the *Renilla* luciferase coding sequence was used[77]. In order to generate the mutated forms of the *NFKBIZ*-3'UTR, and using this plasmid as template, site directed mutagenesis was performed following the manufacture's recommendations (Stratagene) of the the QuikChange II XL Site-Directed Mutagenesis Kit and the oligonucleotides listed in Table 1.

For the genome editing experiments, the lentiCRISPRv2 (Addgene #52961) was used to clone the single-guide RNAs (sgRNAs), following the lentiGuide oligo cloning protocol[104,105] with minor modifications. The sgRNAs were designed to target the 3'UTR portion of *NFKBIZ* where the most recurrent mutation is located: NFKBIZ-101578250-KI_Sense: 5'-CACCGAGCAACACTCACTGTCAGTT-3' and NFKBIZ-101578250-KI_Antisense: 5'-AAACAACTGACAGTGAGTGTTGCTC-3'.

Table 1. List of primers used for mutagenesis

Name	Sequence
NFKBIZ-101578250-C_G-Fwd	GCCTGGCTAGCAACACTCAGTGTCAGTTAGG
NFKBIZ-101578250-C_G-Rev	CCTAACTGACACTGAGTGTTGCTAGCCAGGC
NFKBIZ-101578250-C_A-Fwd	AGCCTGGCTAGCAACACTCAATGTCAGTTAGGC
NFKBIZ-101578250-C_A-Rev	GCCTAACTGACATTGAGTGTTGCTAGCCAGGCT
NFKBIZ-101578250-C_CT-Fwd	CTGGCTAGCAACACTCATCTGTCAGTTAGGCAGTC
NFKBIZ-101578250-C_CT-Rev	GACTGCCTAACTGACAGATGAGTGTTGCTAGCCAG
NFKBIZ-101578254-delAGTT-Fwd	TAGCAACACTCACTGTCAGGCAGTCCTGATGTAT
NFKBIZ-101578254-delAGTT-Rev	ATACATCAGGACTGCCTGACAGTGAGTGTTGCTA
NFKBIZ-101578285-del12-Fwd	CAGTCCTGATGTATCTGTACATAGATATTGGCAAATGTAAGTTGTTTC
NFKBIZ-101578285-del12-Rev	GAAACAACTTACATTTGCCAATATCTATGTACAGATACATCAGGACTG
NFKBIZ-101578292-del16-Fwd	GCAGTCCTGATGTATCTGTACATAGACCATTTTGTAAGTTGTTTCTATGA
NFKBIZ-101578292-del16-Rev	TCATAGAAACAACTTACAAAATGGTCTATGTACAGATACATCAGGACTGC
NFKBIZ-101578304-G_A-Fwd	TACATAGACCATTTGCCTTATATTGACAAATGTAAGTTGTTTCTATGAAAC
NFKBIZ-101578304-G_A-Rev	GTTTCATAGAAACAACTTACATTTGTCAATATAAGGCAAATGGTCTATGTA

Short-hairpin RNAs (shRNAs) plasmids from the MISSION RNAi library used during knockout experiments were bought to Sigma-Aldrich. The identifiers were: TRCN0000255467 (*IGF2BP2*); TRCN0000293594 (*IGF2BP3*); TRCN0000269876 (*RPSAP52*); TRCN0000148785 (*PUM1*); TRCN0000061861 (*PUM2*); TRCN0000431553 (*MCPIP1*); TRCN0000122593 (RC3H1/2); TRCN0000122593 (RC3H1/2); TRCN0000144045 (RC3H1/2); TRCN0000432078 (RC3H1/2); TRCN0000416067 (RC3H1/2).

All plasmids were verified by digestion and/or Sanger sequencing.

Reverse transcription and quantitative real-time PCR

Quantitative PCR (qPCR) was used to study the effect of 3'UTR-variants in *NFKBIZ* over RNA stability. Reverse transcription coupled to PCR (RT-PCR), apart from qPCR, was also used for validation of mis-splicing events in U1 cell lines (JVM3, HG3, MEC1 and transfected HEK-293T).

In all cases, cDNA was synthesized with the QuantiTect Reverse Transcription kit (QIAGEN) using 1 µg of total RNA and following manufacturer's instructions. Quantitative PCR was carried out in triplicate for each sample with 4µL cDNA (1:5) using Power SYBR Green PCR Master Mix (Applied Biosystems) with the oligos mentioned in Table 2 and an Applied Biosystems 7300HT Real-Time PCR System. Relative quantification was analyzed with the $2^{-\Delta Ct}$ or $2^{-\Delta\Delta Ct}$ methods using *ACTB* as the endogenous control.

Table 2. List of primers used for RT-PCR and qPCR

Name	Sequence
Renilla-SYBR-Fwd	AAGAGCGAAGAGGGCGAGAA
Renilla-SYBR-Rev	TGCGGACAATCTGGACGAC
Firefly-SYBR-Fwd	CGTGCCAGAGTCTTTCGACA
Firefly-SYBR-Rev	ACAGGCGGTGCGATGAG
NFKBIZ-SYBR-Fwd	TGCAGTCATAGCCCACAATG
NFKBIZ-SYBR-Rev	TGCTCCCATTTGAATTAGGC
ABCD3-WT-Fwd	GCTCATCACAAACAGTGAAG
ABCD3-U1-A_C-Fwd	GGAACAGAATCTCAGTGAAG
ABCD3-Rev	CAGTTTTCGGAAGACTGAGT
MSI2-WT-Fwd	AGCGCAACCCAAGATGGTCA
MSI2-U1-A_C-Fwd	CCGCAGGAGAATCCTATGGT
MSI2-Rev	CTTGCCAAACTGCTCGAAAT
POLD1-WT-Fwd	CTGGAGATCTCACAGAGCGT
POLD1-U1-A_C-Fwd	CGGATAAAGCAGGAGAGCGT
POLD1-Rev	ACTTAGACTCCACCAGCTGC
PTCH1-WT-Fwd	AGCTGTGGGTGGAAGTTGGA
PTCH1-U1-A_G-Fwd	CACTGCCCTTCCACATTGGA
PTCH1-Rev	TATACATGGACACGGCTGGC
GLI2-WT-Fwd	TGGACGTGTCCCGTTTCTCC
GLI2-U1-A_G-Fwd	AGGAGGCGTTTGTCCCGTTT
GLI2-Rev	GCTGACAGATGCCCGTAGGA

PCRs for the mis-splicing validation of *ABCD3*, *MSI2*, *POLD1*, *PTCH1* and *GLI2* genes were performed using 4 µL cDNA (1:5) and 30 cycles and a melting temperature (Tm) of 60ºC using the oligonucleotides in Table 2.

Rapid amplification of 5' cDNA ends (5'RACE)

Rapid amplification of cDNA ends (RACE) was performed using 1 µg of total RNA from JVM3, HG3 or MEC1 cell lines infected with either pLKO.1-U1^{wt} or pLKO.1-U1$^{g.3A>C}$ and HEK-293T transfected with pLKO.1-U1^{wt}, pLKO.1-U1$^{g.3A>C}$ or pLKO.1-U1$^{g.3A>G}$ following the recommendations of the manufacturer (Sigma-Aldrich), and the following specific oligonucleotides: U1-RACE-SP1: 5'–CAGGGGAAAGCGCGAACGCAGT–3'; U1-RACE-SP2: 5'–CCCACTACCACAAATTATGC–3'). A single amplification band of the expected size (160 bp) was excised from the gel, purified through column (Macherey-Nagel) and sequenced with the internal oligonucleotide U1-RACE-SP2.

rhAMP assay

Genomic DNA from PCAWG primary tumors and CLL patients from HUCA was tested using custom rhAmp SNP assays (Integrated DNA Technology). In brief, locus and allele-specific primers were generated individually for RNU1_batch (RNU1-1, RNU1-2, RNU1-3, RNU1-4 and RNVU1-18) (CD.GT.ZVBW8769.6) and RNU1_pseudo (RNU1-27P and RNU1-28P) (CD.GT.GBJF7460.6). Assays were run in technical triplicates in 5 µL volume (DNA concentration sampled at least 10 ng), according to the manufacturer's indications, with gBlocks for wild-type, mutant and heterozygous genotypes or positive patients as controls. Reporter mix used Yakima Yellow (mutant) and FAM (wild-type) dyes as well as ROX dye for passive reference. Plates were read on the StepOnePlus (Applied Biosystems) RT-PCR machine, and genotypes called using the StepOne v2.3 software.

Sanger sequencing

To perform verification of private calls obtained from the analysis of CLLE-ES cases, five and two mutations detected only by RFcaller and PCAWG, respectively, were chosen to be verified by Sanger sequencing. These positions were chosen because they appear in known driver genes for CLL and because tumor and/or normal DNA was available. The list of primers and melting temperatures are listed in Table 3.

Table 3. List of primers used for Sanger validation of CLL mutations

Name	Sequence*	Tm (ºC)	Amplicon (bp)	Extension time (seconds)	ID
ADAMTS4-Fwd	TACCCAGTTCATGAGCAGCA	56	303	25	1_161166093 DO655 C>A
ADAMTS4-Rev	GGAGTTGTGTGACATGGTGC				
CREB1-Fwd	CTGCCTCTGGAGACGTACAA	58	288	25	2_208442348 DO6558 C>T
CREB1-Rev	GCAAGATCCATTAATTCTGCTGG				
NFKBIE-Fwd	GGACCTCAAAAGTGGGCTGA	58	288	25	6_44232738 DO7172 TGTAA>T
NFKBIE-Rev	TCACCTACACCCTGTCCTTG				
MED12-Fwd	CTGCCCTTTCACCTTGTTCC	58	290	25	X_70339253 DO6558 G>A
MED12-Rev	CCCTATAAGTCTTCCCAACCCA				
ITPKB-Fwd	ACAAAAGTCTCTGCCAGTGG	56	290	25	1_226827324 DO52712 CT>C
ITPKB-Rev	CTGGGTGGGGTGTTCTCTT				
SETD2-Fwd	TTCTTACTGGCTGCAAGGGCTGA	54	214	25	3_47088090 DO6934 G>A
SETD2-Rev	CAACTTGGAAGTCAGTCTGT				
IKBKB-Fwd	ACCCTCAGCTTTCTCCTTCC	56	255	25	8_42163889 DO7084 A>C
IKBKB-Rev	TGTGACCTCATGCATCTCCA				

**Primers used for sequencing are underline.*

Luciferase assays

Luciferase assays were performed to check the post-transcriptional effects of the mutations located in the 3'UTR of *NFKBIZ*. HeLa and HEK-293T cells were transfected (triplicates) with the psiCHECK2 constructions, 48 h later luminescence determination was performed using Dual-Glo Luciferase Assay System (Promega) following the manufacturer's instructions. Measures were carried out in a Varioskan Flash plate reader.

Firefly luciferase activity was used as endogenous control for normalization. Relative luciferase activity was calculated as the ratio of luminescence from the experimental reporter (Renilla) to that of the control reporter (Firefly). Each transfection experiment was repeated three times.

Western Blot Analysis

Protein extracts were separated in SDS-PAGE gels and electrotransferred to nitrocellulose or polyvinylidene fluoride (PVDF) membranes (GE Healthcare Life Sciences). Then, membranes were blocked with 5% non-fat dry milk or Bovine Serum Albumin (BSA) in TBS-T buffer (20 mM Tris pH 7.4, 150 mM NaCl, 0.05% Tween 20). Following 4 °C overnight incubation with the primary antibody (Table 4). After washing with TBS-T, membranes were incubated with secondary antibodies (Table 4) for 1 h at room temperature. To detect the horseradish peroxidase-conjugated (HRP) species-specific secondary antibodies the Luminata Forte HRP Substrate (Millipore) was used on a LAS-3000 (FUJIFILM), while for IRDye secondary antibodies with fluorescence protein bands were visualized and recorded using LI-COR Odyssey Imaging System (LI-COR). For the study of neoepitopes, 100 µL of fresh serum from patients was used as primary antibody and the western blot was revealed with anti-human-IgG (HRP) as secondary antibody.

Table 4. List of antibodies used in this work

Antibody	Species and type	Used dilution	Reference
Anti-β-Actin	Mouse monoclonal	1:5000	Sigma (A5441)
Anti-Cleaved-Caspase-3	Rabbit polyclonal	1:1000	Cell Signaling (9661T)
Anti-GAPDH	Rabbit monoclonal	1:1000	Santa Cruz (sc-32233)
Anti-Histone-H3	Rabbit polyclonal	1:1000	Cell Signaling (9715)
Anti-Human-IgG	Goat polyclonal	1:20000	ThermoFisher (31413)
Anti-IkBβ	Rabbit polyclonal	1:1000	Cell Signaling (9244)
Anti-NFkB-p52	Mouse monoclonal	0.5 µg/mL	Millipore (05-361)
Anti-PARP	Rabbit polyclonal	1:1000	Cell Signaling (9542)
Anti-α-Tubulin	Mouse monoclonal	0.2 µg/mL	Sigma (T6074)
Anti-U1-snRNP70	Mouse monoclonal	1:1000	Santa Cruz (sc-390899)
IRDye 680RD	Goat anti mouse	1:10000	Li-COR (926-68070)
IRDye 680RD	Goat anti rabbit	1:10000	Li-COR (926-68071)
IRDye 680RD	Goat anti rat	1:10000	Li-COR (926-68076)
IRDye 800CW	Goat anti mouse	1:10000	Li-COR (926-32210)
IRDye 800CW	Goat anti rabbit	1:10000	Li-COR (926-32211)
HRP	Goat anti mouse	1:10000	Jackson ImmunoResearch (115-035-062)
HRP	Goat anti rabbit	1:10000	Cell Signaling (70745)
HRP	Goat anti rat	1:5000	Amersham (NA935V)

Subcellular fractioning

In order to study the localization of lncRNAs in JVM3 cell lines harboring *U1*^{wt} or *U1*$^{g.3A>C}$, subcellular fractionation was carried out to extract RNA and proteins from each of the fractions (cytoplasm, nucleoplasm and chromatin). Only the following buffers, with minor modifications, described by Mayer A. and Churchman S (2017)[106] in their protocol were prepared: cytoplasmic lysis buffer, glycerol buffer, nuclei lysis buffer, nuclei wash buffer and sucrose buffer. Briefly, cells (1x10^7) were washed and collected by centrifugation. Then, they were resuspended in the cytoplasmic lysis buffer and centrifuged in the sucrose buffer to separate the cytoplasm fraction from the nuclei. After centrifugation, the supernatant (cytoplasm) was frozen in TRIzol and the pellet (nuclei) was resuspended in the nuclei lysis buffer to centrifugate it again. Finally, both the supernatant (nucleoplasm) and the pellet (chromatin) were frozen in TRIzol to continue with RNA and protein extractions.

Cell biology methods

Cell culture

Chronic B cell leukemia cell lines: JVM3 (no. ACC 18), HG3 (no. ACC 765) and MEC1 (no. ACC 497) were obtained from DSMZ. JVM3 and HG3 were grown in RPMI 1640, 10% fetal bovine serum (FBS), 1% penicillin-streptomycin-L-glutamine (PSG) and 1% antibiotic-antimycotic (AA), whereas MEC1 was grown in Iscove's modified Dulbecco's medium (IMDM), 10% FBS, 1% PSG and 1% AA (all from Gibco). On the other hand, HEK-293T and HeLa were grown in Dulbecco's modified Eagle's medium containing 10% FBS and 1% PSG. In the case of interleukin 1 alpha (IL1) stimulation (eBioscience), cells were cultured overnight without FBS the day before luciferase assay or RNA extraction protocols. IL1 was added four hours before experiments at a concentration of 2 ng/mL. The authenticity of the cell lines was tested with the AmpFLSTR Identifiler Plus PCR Amplification Kit (Applied Biosystems). In addition, cell lines are routinely tested for mycoplasma.

Transient transfections and viral infections

Transient transfections were performed using Lipofectamine and Plus reagents (Invitrogen), following manufacturer's instructions and with cells at 60-70% confluence. The subsequent transcriptomic and proteomic experiments were performed 48 h after transfection.

For lentiviral infection, HEK-293T cells ($5x10^6$ cells) were cultured in 10-cm plates and transfected using Lipofectamine Plus (Invitrogen) with 2 µg of the specific lentiviral vector (pLKO.1-$U1^{wt}$, pLKO.1-$U1^{g.3A>C}$), together with 1 µg of psPAX2 packaging plasmid and 1 µg of pMD2.G envelope plasmid. Twelve hours after transfection, the medium was replaced with complete medium, and 24 h later 10 mL of supernatant were filtered (0.45 µm), and 4 mL was used to infect cell lines (JVM3, HG3 and MEC1 for U1 and JVM3 and HG3 for NFKBIZ experiments) in the presence of 8 µg/mL polybrene. The infection was repeated 24 h later, and after 24 h cells were plated in complete medium for one day. Then, cells infected with pLKO.1-$U1^{wt}$ or pLKO.1-$U1^{g.3A>C}$ were selected with 1.2 µg/mL of puromycin for four days.

CRISPR-Cas9 experiments

For the generation of *NFKBIZ*-3'UTR mutated cells, the lentiCRISPRv2-NFKBIZ plasmid was used together with two megamers (Integrated DNA Technology) designed to insert two of the detected mutations in the 3'UTR hotspot of *NFKBIZ*: ssODN-101578250-C_G: `5'-CTATGTACAGATACATCAGGACTGCGTAACTGACACTGAGTGTTGCTAGCCAGGCTCCAAGCTAATGGAGC-3'` and ssODN-101578254-delAGTT: `5'-CAAATGGTCTATGTACAGATACATCAGGACTGCGTGACAGTGAGTGTTGCTAGCCAGGCTCCAAGCTAAT-3'`. HeLa and HEK-293T were transfected with 1 µg of the ssODN-101578254-delAGTT and ssODN-101578250-C_G megamers, respectively. Although the transfection was transitory, 24 h later, HeLa and HEK-293T were selected with 1 µg/mL and 2.5 µg/mL of puromycin for 48 h, respectively. Then, cells were let to recover for 24 h and finally plated in four 96-well plates with a density of 0.5 cell/well to be sure that two clones did not fall into the same well. As the clones grew, they were genotyped (NFKBIZ-Seq-Hotspot-Fwd: `5'-TGTTCCTGTTAGTTGAGGCTGA-3'`; NFKBIZ-Seq-

Hotspot-Rev: 5'-TGTGCCACAAATCAAAGTCC-3'; Cycles: 25; Tm: 60 °C; Extension: 50 seconds) to check which mutations were present in each one. Because complex mutations can be inserted into each allele, the TIDE (Tracking of Indels by DEcomposition) tool was used to assist in the interpretation of electropherograms[107].

Short Hairpin RNA knockdown

pLKO.1 plasmids derived from the MISSION RNAi library and containing the desired short hairpin RNAs (shRNAs) were transfected in HeLa cells to silence *IGF2BP2*, *IGF2BP3*, *RPSAP52*, *PUM1*, *PUM2* and *MCPIP1* genes, as they might be interacting with the *NFKBIZ*-3'UTR.

Proliferation assays

To quantify cell proliferation, a Cell Titer 96 Non-Radioactive cell proliferation kit was used following manufacturer's instructions (Promega). Briefly, cells were seeded into 96-well plates at a density of 2,000 cells per well and plates were incubated at 37 °C, 5% CO_2 for four consecutive days. Cell proliferation was monitored by measuring the conversion of a tetrazolium salt into formazan in metabolically active cells. At the desired time points (0 h, 24 h, 48 h, 72 h, 96 h and 120 h), 15 µL of dye solution were added into each well (n = 5) and cells were incubated at 37 °C for 2 h. Then, 100 µL of solubilization mixture were added into each well to stop the reaction. The formazan absorbance was measured at 570 nm, and 650 nm to substrate the background, with a Power Wave XS Microplate reader (Biotek). Then, each point was normalized with time 0 h and mean was represented.

Proteasome and splicing inhibitors on cell proliferation

Due to the aberrant splicing at the transcriptomic level found in cell lines (JVM3 and HG3) harboring the *U1*$^{g.3A>C}$ mutation, it was decided to study the effect that inhibition of the spliceosome or proteasome could have on cell viability using *U1*^{wt} as control. To block the spliceosome activity the Pladienolide B (PladB, R&D Systems) inhibitor was used at concentrations between 0 nM and 50 nM. On the other hand, Bortezomib (Selleckchem) was employed for proteasome inhibition at concentrations

from 0 nM to 9 nM. To test the effect of these drugs on cell viability, proliferation assays and western blots against PAPR and cleaved caspase 3 and TrypanBlue counts (triplicates) were carried out at different concentrations and times (0 h, 24 h and 48 h).

Bioinformatical and statistical analyses

Libraries preparation

Whole-exome DNA sequencing was performed using 2.5 µg of matching tumor and germline genomic DNA samples of the 12 patients with pediatric MB. Library preparation and enrichment was performed using SureSelect Human All Exon V5 (Agilent Technologies) and sequenced using Illumina paired-read sequencing platform HiSeq2500 (2×100 bp; coverage: >75x control, >100x tumor; 50.4 Mb). For one patient (MB13), WES was performed using the SureSelect XT HS Human Focused Exome 17.7 Mb. Given the limited amount of cfDNA available from the 13 patients, CSF-derived cfDNA from four patients underwent DNA sequencing (6-200 ng). For one patient (MB8), SureSelect Human All Exon V5 (Agilent Technologies) was used for the CSF-derived cfDNA obtained at relapse (51.6 Mb). For the remaining 3 patients, longitudinal CSF samples were obtained before surgical intervention of the tumor biopsy and during progression or follow-up. Matching tumor and germline (200 ng), and follow-up CSF samples (6-200 ng) were analyzed using the SureSelect XT HS Human Focused Exome 17.7 Mb (custom constitutional panel, Agilent Technologies) for library enrichment and sequenced using Illumina platform NextSeq with a read length of 150 bp. For MCL samples, WGS was performed using the TruSeq DNA PCR-free protocol and sequenced in an Illumina HiSeq X Ten (2x150 bp; coverage 30x).

In the case of RNA from U1 cell lines, two technical replicates for each of the three cell lines (JVM3, HG3 and MEC1) and two conditions (pLKO.1-U1^{wt} or pLKO.1-U1$^{g.3A>C}$) were prepared as stranded total RNA-seq libraries and then sequenced with the Illumina HiSeq 4000 system (2×76 bp) with >40 million paired-end reads per sample. For RNA from HEK-293T, transfected with either pLKO.1-U1^{wt} or pLKO.1-U1$^{g.3A>G}$ (two replicates), the mRNA library construction was performed based on oligo dT-based

mRNA isolation using NEBNext Poly(A) mRNA Magnetic Isolation Module and sequenced on NextSeq 550 (2x100 bp; > 40 M). To study the localization of lncRNAs from subcellular fractioning samples (chromatin and total), the ERCC RNA Spike-In Control Mixes (Ambion by Life Technologies) was added to total RNA as endogenous control following the manufacturer's guides. Then, TruSeq Stranded Total RNA with Ribo-Zero Gold libraries were prepared and sequenced on NovaSeq6000 (2x150 bp; > 60 M).

RNA-seq libraries from *NFKBIZ* CRISPR clones (HEK-293T) were prepared for two technical replicates and two conditions (wild-type or g.101578250C>G) from total RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina). Each library was sequencing with the Illumina HiSeq2000 (2×76 bp) with >40 million paired-end reads per sample.

Read alignment

All genomic sequencing data were aligned using BWA-MEM (v0.7.17)[108] to the GRCh37 (hs37d5) version of the reference genome except for MCL cases, which aligned to the GRCh37 (v13) that does not contain the patches of the alternative chromosomes and the decoy. PCR duplicates were marked using biobambam (v0.0.148)[109] for medulloblastoma WGS samples, MarkDuplicates (http://broadinstitute.github.io/picard/) for PCAWG data or Samblaster (v0.1.25)[110] for MCL cases. For pediatric medulloblastoma WES samples, due to the low amount of starting DNA, a high percentage of PCR duplicates were found so they were removed using Samblaster (v0.1.25).

RNA-seq raw reads were aligned to the GRCh37 (hs37d5) genome using the two-pass STAR (v2.7.3a)[111] method, to maximize the sensitivity of novel junction discovery, and GENCODE v19 as the reference gene annotation. The quality control process was done with FastQC (v0.11.7) and multiQC (v1.5)[112] and transcript integrity number was calculated with RSeQC (v2.6.4)[113]. For RNA-seqs containing the ERCC RNA Spike-In, the sequences of control templates were added to the GRCh37 (hs37d5) genome before alignment.

Variant calling

Somatic mutation calling for pediatric medulloblastoma WES derived was performed using Sidrón[29] and annotated using the Variant Effect Predictor[84]. Germline mutations were extracted using bcftools[114] and filtered using common SNPs from dbSNP v153. Filtering was performed based on the following criteria. For SSNVs, only mutations affecting the protein coding sequence were considered; a depth of coverage ≥20 was required to compare the tumor with the circulating tumor DNA (ctDNA) variant allele frequency (VAF); if VAF > 5% in any of the samples from the same case, it was considered as a mutation, and its presence in the other samples was studied. To determine whether a sample contained a mutation that was detected in another sample from the same case, a VAF greater or equal to 1% was established as threshold. For germline mutations, only coding regions with a mapping quality ≥50 and heterozygous variants ($0.35 \leq VAF \geq 0.75$) were analyzed. Moreover, mutations with a VAF greater or equal to 0.001 in the control cohort of the gnomAD database were removed to ensure that polymorphism were not present in the final set of germline mutations.

In the case of the study of MCL genomes, somatic single-nucleotide variants were called using Sidrón, short insertions/deletions combining Sidrón and Pindel[93], and germline variants using HaplotypeCaller[115].

As *U1* is a repetitive gene, traditional callers are not able to detect mutations, so a specific calling method was developed to detect them. First, samples without enough coverage were flagged as genotype-undetermined and left to manually review. The coverage was determined by the median read depth at the 5' splice-site recognition sequence of seven *U1* genes. For 2,434 donors with enough coverage (≥15 median coverage in at least five *U1* genes), all reads mapped by BWA MEM to *U1* genes and pseudogenes as well as their flanking 1-kb regions were extracted with samtools[116] and saved as miniBAMs. These miniBAMs were then converted into paired FASTQ files and re-aligned with Bowtie2 (v2.3.4.1)[117] to GRCh37 in multiple mapping report mode (-k). Non-default parameters for Bowtie2 were "–score-min L,-0.3,-0.3–no-mixed–no-discordant -k 100–very-sensitive".

Then, for each pair of multiple mapped reads, only alignments with minimal total edit distance (sum of edit distance in two mates) were kept. Reads mapped to *U1* pseudogenes or other genomic regions were discarded. Next, for each re-aligned BAM, we counted the number of variant reads and the read depth (number of reference reads + number of variant reads) for each position, and for forward and reverse strand separately. To account for multiple mapping, an extra procedure only for the read depth counting were performed: that is, when a read had k equally good alignments, it is only counted as 1/k read. Then a beta-binomial error model trained on a project-specific panel of normal samples was used to call mutations, which was implemented with a modified version of EBCall. Finally, IGV was used to manually curate all mutation calls and filtered out mutations that were supported by reads with multiple mismatches in the same gene, or that had three or more variant reads in the paired normal sample according to BWA MEM or Bowtie2 alignments. To further minimize the false-negative rate for the g.3A>C mutation, tumors that were called as wild type but that had two or more variant reads at the third base of any *U1* genes were also assigned to the undetermined group.

Copy number and structural variants

For exome data, copy number variants were detected using the program exome2cnv[118] and regions which included <10 probes were removed from the analysis. Then, normal samples were compared to each other to identify any region that could be altered only by the region's own variability. Finally, a summary with the events was done to obtain the percentage of loss/gain of each chromosome. Regions with a $\log_2$Ratio greater or equal to 0.5 were considered gains, and lower or equal to −0.5 were considered deletions. For a more stringent analysis $-1 \leq \log_2\text{Ratio} \geq 1$ was used as a threshold to report CNAs. Copy number alterations in whole-genome data were extracted using Battenberg[119] and structural variants were analyzed using SMuFin[94] and LUMPY[120].

Identification of driver genes

For the identification of driver genes after the analysis of 61 WGS of patients with MCL, all detected mutations and structural alterations focused on a specific gene were taken into account. The results of the previous analysis of another 21 cases with WES were also incorporated[121]. After integration of these data, we followed a frequency-based strategy to determine which genes could be considered as drivers of the disease. This type of strategy is characterized by comparing the mutational background with respect to the number of mutations detected in each gene, to check whether the gene has more mutations than expected by chance. In this way, the mutational background was first calculated based on the number of non-coding mutations present in a 0.5 Mb window at each side of the gene, together with the length of the gene and the coverage and callability of the coding region in each patient. Then, we considered that a gene was mutated in a patient if it had a missense or nonsense mutation, a non-coding mutation altering its functionality or was affected by a structural alteration (gain or loss). Finally, to test whether a gene was mutated in more samples than expected by chance, we calculated the probability that each gene carried a non-synonymous mutation, following the statistical model described by Puente et al. (2015)[29]. The list of statistically significant genes was manually reviewed to eliminate possible false positives, such as genes that had a higher number of mutations than expected by chance due to other indirect mechanisms such as somatic hypermutation (SHM).

Analysis of mutational signatures

For the analysis of mutational signatures two bases upstream and downstream of the mutations were selected to reconstruct the context and generate a 1,536 mutational matrix with which extract the signatures for each sample. The sigminer[122] R wrapper (nrun= 300 and refit = TRUE) was selected to run the SigProfilerExtractor framework[55], which uses non-negative matrix factorization iterations to identify the matrix of mutational signatures and the matrix of the activities of these signatures. This analysis was performed for three independent datasets: mutations detected only by RFcaller or PCAWG, respectively, and variants detected by both pipelines.

Gene expression and splicing analysis

After STAR alignment of RNA-seq data, gene-level expression was counted by htseq-count (v0.12.4)[123] and transcript level expression was estimated by Kallisto (v0.44.0)[124]. The limma[125] package with the limma-trend method was used for differential expression analysis. Genes that do not have a worthwhile number of reads in any sample were filtered and only those with a $q < 0.1$ or adjusted p-value < 0.01 and absolute $\log_2$-transformed fold change > 1 were considered as significantly differentially expressed.

For intron-centric differential splicing analysis, the LeafCutter[126] package was used to quantify intron usage and identify differentially spliced intron clusters between two conditions. Splice junction files (SJ.out.tab) generated by STAR were used as input for LeafCutter. Only splice junctions supported by uniquely mapped reads and with at least 6-bp maximum overhang were used. An intron was considered as significantly differentially spliced when $q < 0.1$ and absolute $\log_2$(effective size) > 1.

For the study of *U1* mutations, only RNA-seq data that met the following criteria was considered: first, FASTQ files passed at least three main FastQC flags (overrepresented sequences, per base N content, per base sequence quality, per sequence GC content and per sequence quality scores); second, more than 50% reads were uniquely mapped and the total number of reads mapped by STAR was greater than 1 million; third, the total number of fragments counted by htseq-count was greater than 5 million; and fourth, the transcript integrity number was greater than 50.

RNA-seq data for subcellular fractions were analyzed manually. Htseq-count results were normalized by ERCC spike-in expression and cell line batch effect was removed using the *removeBatchEffect* function from limma. Transcripts per million (TPMs) were calculated and genes with lower expression than 3 TPMs were filtered.

All false discovery rate controls were conducted with the Benjamini–Hochberg procedure and false-discovery rate of 10% ($q < 0.1$) was selected as the significant threshold. P-values were adjusted by Bonferroni method.

Gene-set overrepresentation and enrichment analysis

Gene lists that were significantly overrepresented in differentially expressed and spliced genes were identified from Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes and Reactome databases using g:Profiler[127], GOseq[128] and clusterProfiler[129]. Gene-set enrichment analysis (GSEA) was also conducted for differentially expressed genes using pre-ranked gene lists ordered by $-\log_{10}$(p-value) x (sign of fold change) by camera and clusterProfiler packages. For GSEA, we focused on C1 (hallmarks), C2 (curated) and C5 (GO) gene sets in the Molecular Signatures Database (MSigDB v7.0).

RFcaller algorithm training

For the development of machine learning algorithms, two different set of mutations were used, a training set and a testing set. To build them, all possible somatic mutations from four WGS MCL samples sequenced at 30x coverage (M032 and M439 for training; M065 and M431 for testing) were extracted with bcftools. For the initial training, previously published mutations were defined as true positive mutations. With each iteration, all discordant calls were manually reviewed by three experts, through visual inspection, and the database was updated accordingly. This procedure resulted in the identification of novel *bona fide* mutations that would constitute false negatives in the initial set, as well as the rejection of certain mutations, such as artifacts or germline mutations present in the original dataset, that would represent false positives, respectively. After several rounds of training the algorithms and curating the set of mutations, all discordant variants had already been examined, which allowed us to obtain a reliable dataset for training and testing the final version of the algorithms.

To train the algorithms, a training set containing 66,096 SSNVs and 931 indels was used for which read-level features were previously extracted (Table 5). These data were used as input by TPOT (v0.11.1)[130], with the default configuration of the *TPOTRegressor* function, to find the best pipeline to train the regression algorithms. As a result, an extremely randomized tree "Extra-Tree" Regressor for SSNVs and a Random Forest Regressor for indels were built. In both cases, a transformation of the data was carried out before the regression using the *StackingEstimator* function.

Table 5. List of read-level features required to run the algorithms

Key	Definition	Algorithm
Q30N_cov	Coverage in normal sample when the mapping quality is high (Q≥30)	SSNV; indel
Q30T_cov	Coverage in tumor sample when the mapping quality is high (Q≥30)	SSNV; indel
Q30N_mut_reads	Number of mutated reads in normal sample with high mapping quality (Q≥30)	SSNV; indel
Q30T_mut_reads	Number of mutated reads in tumor sample with high mapping quality (Q≥30)	SSNV; indel
N_mut_reads	Number of mutated reads in normal sample without filtering reads (Q≥0)	indel
T_mut_reads	Number of mutated reads in tumor sample without filtering reads (Q≥0)	indel
N_normal_mapQ	Average mapping quality for wild-type reads in normal sample without filtering reads (Q≥0)	SSNV; indel
T_normal_mapQ	Average mapping quality for wild-type reads in tumor sample without filtering reads (Q≥0)	SSNV; indel
N_mut_mapQ	Average mapping quality for mutated reads in normal sample without filtering reads (Q≥0)	SSNV; indel
T_mut_mapQ	Average mapping quality for mutated reads in tumor sample without filtering reads (Q≥0)	SSNV; indel
Normal_Error_Ratio	Percentage of mismatched nucleotides around the position of the mutation in normal sample	SSNV; indel
Tumor_Error_Ratio	Percentage of mismatched nucleotides around the position of the mutation in tumor sample	SSNV; indel
Normal_cigar	Number of reads in normal sample containing a different cigar other than M	SSNV; indel
Tumor_cigar	Number of reads in tumor sample containing a different cigar other than M	SSNV; indel
Dimers	Number of repeated dinucleotides around the position of the mutation	SSNV; indel
GC_percentaje	Percentage of GC	SSNV; indel
Interval_size	The distance between the leftmost and rightmost mutation in the reading	SSNV; indel
Mean_position	Mean position of the mutation along the mutated reads	SSNV; indel
N_repeat_indel	Number of times the indel is repeated around the position of the mutation	indel
Normal_insertion_count	Count of insertions around the position of the mutation in normal sample	indel
Normal_insertion_lenght	Length of the largest insertion around the position of the mutation in normal sample	indel
Normal_deletion_count	Count of deletions around the position of the mutation in normal sample	indel
Normal_deletion_lenght	Length of the largest deletion around the position of the mutation in normal sample	indel
Tumor_insertion_count	Count of insertions around the position of the mutation in tumor sample	indel
Tumor_insertion_lenght	Length of the largest insertion around the position of the mutation in tumor sample	indel
Tumor_deletion_count	Count of deletions around the position of the mutation in tumor sample	indel
Tumor_deletion_lenght	Length of the largest deletion around the position of the mutation in tumor sample	indel

Once we had the algorithms, the test dataset, with 63,948 SSNVs and 2,506 indels, was used to select the best cutoffs for both pipelines. With this purpose, the result from RFcaller was filtered to get the "QUAL" field for those mutations that passed all filters. This parameter is calculated considering the initial quality from bcftools and the regression value for SSNV and indels, and only the regression value for homopolymer indels (polyindels):

$$QUAL_{SNV} = bcftools\ qual * regression\ value^2$$

$$QUAL_{indel} = bcftools\ qual^{regression\ value}$$

$$QUAL_{polyindel} = regression\ value$$

Then, ROC curves and area under the curve (AUC) were calculated using OptimalCutpoints[131] with the *MaxEfficiency* method. False/True Positive/Negative ratios were calculated using the formulas described in the ROCR R package[132].

Computational cost

To compare the performance of RFcaller with other *state-of-the-art* tools, the docker containers corresponding to the four callers used by PCAWG for the detection of SSNVs were downloaded (https://dockstore.org/organizations/PCAWG/collections/PCAWG). After minor fixes of broken links in the Sanger and DKFZ tools, all of them were run with the default parameters for one random donor. In case the tools allowed to choose the number of threads and RAM to be used, 20 threads and 200 Gb of memory were specified. In addition, because RFcaller allows multiple samples to be run simultaneously, four cases were run in parallel using the default parameters to calculate the computational cost.

Comparative analysis of RFcaller and PCAWG mutations

To validate that the trained models are applicable for liquid and solid tumors and to compare the results to those obtained by the PCAWG pipeline, RFcaller was run for the CLLE-ES and BRCA-EU studies. PCAWG BAM files were downloaded from the "collaboratory" repository using the score-client program. RFcaller was run with its default parameters for all samples and the obtained results were combined into a single

VCF file for each study. Then, a custom panel of normals was used to annotate variants in complex regions. The set of mutations detected by the PCAWG pipeline were extracted from the controlled consensus callsets for SSNV/Indel. To analyze coding and non-coding mutations, the VEP tool was launched for both datasets using the following options: --offline --format vcf --dir_cache homo_sapiens –-symbol --force_overwrite --total_length --numbers --ccds --canonical --biotype --pick --vcf --assembly GRCh37.

To be able to compare both set of mutations in the most accurate manner: i) dinucleotides and trinucleotides from RFcaller were split as this feature is not available for PCAWG, ii) RFcaller mutations located in alternative chromosomes and PCAWG's variants that appear in our custom dbSNP were removed and iii) only mutations that passed all filters were studied. For this comparison, a mutation was considered as subclonal when its variant allele frequency was lower than 0.15.

For the purpose of calculating the precision and recall of both pipelines in each study, 1% or at least 50 discordant mutations from each section were manually reviewed by a panel of experts. Thus, a total of five blocks were checked: mutations detected only by RFcaller and mutations detected between one and four of the callers used by the PCAWG, as the ratio of false positives may be different between them. The results obtained were then extrapolated to the whole set of mutations to calculate the parameters needed to define precision and recall for both pipelines. These measures were calculated with the *prediction* and *performance* functions of the R package ROCR.

Additionally, deep sequencing data generated by previous studies[133,134] for some CLLE-ES cases were used to analyze possible subclonal mutations in driver genes. In order to compare both results, only mutations in CLL driver genes and donors analyzed by both WGS and deep sequencing were selected. In addition, mutations detected by deep sequencing were removed from the analysis if they were germline or there was insufficient coverage or reads supporting the mutation by WGS.

To also test the performance of RFcaller on exome sequencing data, five CLLE-ES cases previously analyzed by WGS and for which exome data were available, were selected. RFcaller was run with default parameters and LIKELY_GERMINAL variants were

removed. Only mutations within the targeted regions of the exome (Agilent – SureSelect Human All Exons V4) were taken into account. Finally, for those mutations not detected by both methods, total coverage and number of mutated reads were extracted in order to determine the cause for loss.

Additional information

The bioinformatics analyses mentioned above have resulted in a large amount of raw data, which due to their length could not be inserted in this thesis. In this regard, we refer for example to the lists of mutations, CNAs or SVs detected in the different projects, the differential expression and splicing analyses, or the results of RFcaller training and its comparison with PCAWG. However, all these data are already published and can be found as supplementary material within the list of articles that constitute this thesis.

RESULTS

Genomic analysis of MCL whole genomes

Mantle cell lymphoma is a mature B-cell neoplasm genetically characterized by the translocation t(11;14)(q13;q32), leading to *CCND1* overexpression[135–137]. Although this tumor has a very heterogeneous behavior, two molecular subtypes with different clinical and biological features can be identified[138–140]. A poor prognosis conventional MCL (cMCL) characterized by the expression of SOX11 and a large number of genomic alterations and leukemic non-nodal MCL (nnMCL), negative for SOX11 and genetically stable which follows an indolent behavior.

Although the mutational profile of MCL has been reported previously[121,141,142], these studies were carried out with a small number of cases and by whole-exome or targeting sequencing, which did not allow to explore the genome-wide mutational and structural alterations of both MCL subtypes.

Genome-wide mutational and structural alterations

To extend the knowledge of mutational and structural landscape in MCL, the group of Dr. Elías Campo (IDIBAPS) performed WGS in 44 cMCL and 17 nnMCL cases. We used our previously developed Sidrón pipeline to perform somatic mutation calling for SSNVs and small indels in these cases. This resulted in the detection of a median of 3,593 somatic mutations per case (1.2 mutations/Mb), including 33 coding mutations per tumor. Additionally, MCL tumors carried a median of 9 SV and 9 CNA. The mutational burden was similar in both MCL subtypes, but cMCL carried higher number of SV (median, 13 vs 3) and CNA (median, 12 vs 1) than nnMCL (Figure 9). A complex genomic landscape, defined by the presence of ≥15 SV, ≥15 CNA and/or complex alterations (chromothripsis, chromoplexia, kataegis or breakage-fusion-bridge) was observed in 23 (52%) cMCL and 3 (18%) nnMCL.

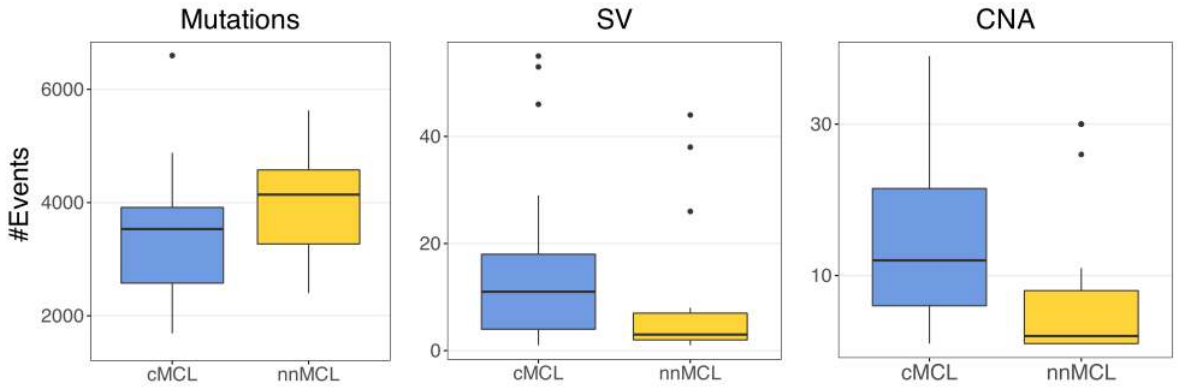

Figure 9. Number of mutations and structural alterarions identified by WGS between MCL subtypes. Box plots of the total number of mutations (median: 3593 (total); 3553 (cMCL); 4112 (nnMCL)), structural variants (SV) (median: 9 (total); 13 (cMCL); 3 (nnMCL)) and copy number alterations (CNA) (median: 7 (total); 12 (cMCL); 1 (nnMCL)) in cMCL (*blue*) and nnMCL (*yellow*).

Chromothripsis events were clonal in all but one case, and recurrently targeting *RB1* in 4 (9%) cMCL and *TERT* in 2 (12%) nnMCL. Chromoplexia affected 14 different chromosomes with *TERT* the only cancer gene affected in one case. Chromothripsis and chromoplexia occurred in both MCL subtypes, but breakage-fusion-bridge (BFB) cycles, a novel and frequent finding in MCL, was only observed in cMCL (20%). This is a mechanism of genomic instability that occurs during mitosis due to the loss of the telomere in one of the chromatids. Upon duplication during mitosis, both sister chromatids, as they lack telomeres, join together to form a single dicentric chromosome. When both centromeres separate, if the breakage does not happen at exactly the same point of fusion, structural rearrangements occur. Because the resulting chromosomes lack telomeres, duplications, deletions, inversions or translocations continue to occur in each cycle[143]. Thus, BFB cycles generated recurrent high-level amplification of *BMI1* (4 cases) and *MIR17HG* (2 cases) and were associated with worse clinical outcome. On the other hand, although chromosomal translocations and inversions were relatively frequent in MCL, they were not recurrent and very few were associated with known cancer genes. In this regard, only two cMCL had SVs that truncated *PAX5,* and just one cMCL had a balanced 2p inversion that fused *MYCN* with IGK enhancer, leading to high overexpression of the gene.

The CNA detected by WGS confirmed the specific MCL profile previously characterized by frequent losses of 1p22-p13, 6q, 9p21/*CDKN2A*, 9q22-q31, 11q22-q23/*ATM*, 13q14/*RB1*, 13q33q34, and 17p/*TP53*, and gains of 3q25-q29 and 7p. We also identified novel recurrent losses at 10q21.1 and 15q14-q21.1 and significant differences in the distribution of specific alterations in cMCL (losses of 1p22-p13, 6q, 9q22-q31, 11q22-q23/*ATM*, 13q33-q34, and gains of 3q25-q29 and 7p) and nnMCL (loss of 17p/*TP53*) (Figure 10).

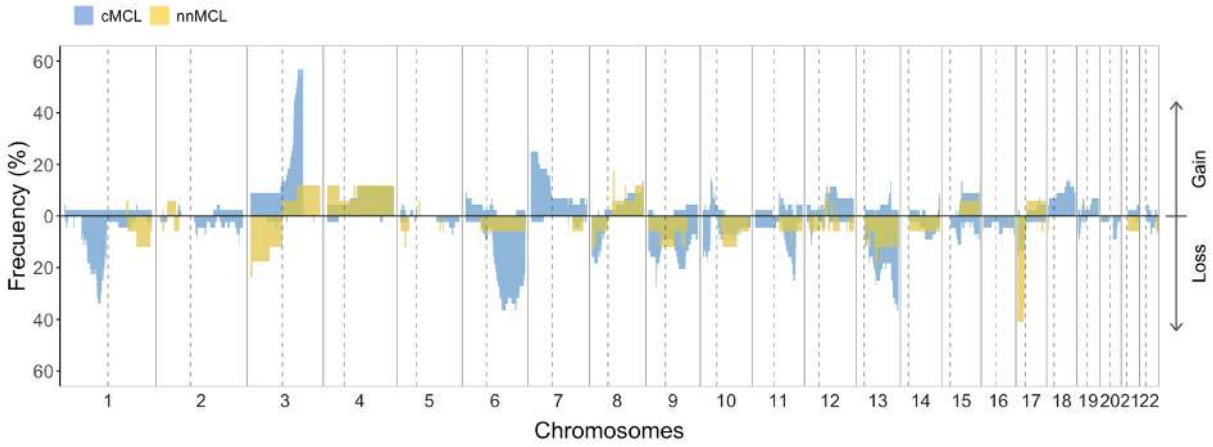

Figure 10. Global profile of CNA in cMCL and nnMCL. Only regions with at least 6 altered cases were included in the comparison. cMCL is colored in blue and nnMCL in yellow.

Since the translocation between chromosomes 11 and 14, which results in *CCND1* overexpression, was detected in practically all patients with WGS (60/61), and as this technique provides sufficient information to be able to detect breakpoints, we decided to investigate this event in more detail. Thus, we detected nineteen translocations (14 cMCL, 5 nnMCL) occurring at a small region of just 89 bp within the previously recognized major translocation cluster (MTC). The remaining breakpoints were similarly scattered at both sides of the MTC in cMCL and nnMCL (Figure 11). Moreover, most 5' and MTC breaks occurred near CpG sites and activation-induced cytidine deaminase (AID) motifs, whereas 3' breaks were only found near AID motifs. This provides a strong and direct evidence that translocations between *IgH* and *CCND1* are initiated by AID in both subgroups.

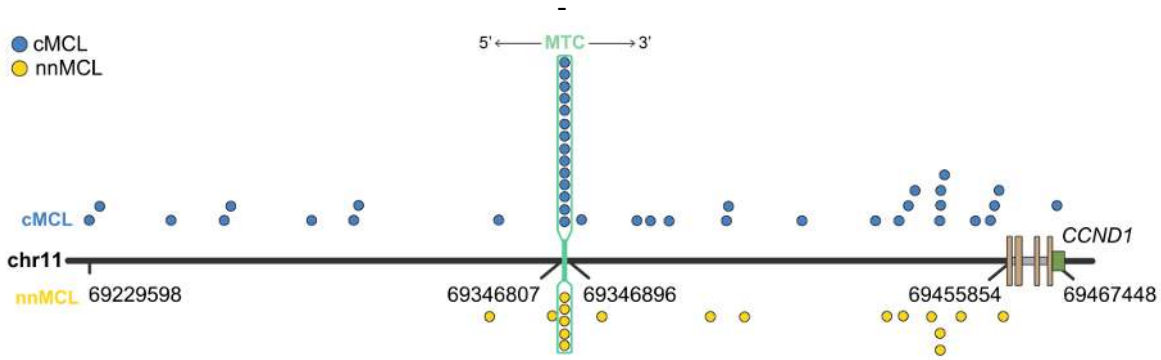

Figure 11. *CCND1* breakpoints. Distribution of the *CCND1* breakpoints in cMCL (*blue-top*) and nnMCL (*yellow-bottom*). The major translocation cluster (MTC) corresponding to 89 bp region is highlighted in green. Both cMCL and nnMCL have the same breakpoint distribution: 21% vs 12% (5'), 33% vs 29% (MTC) and 44% vs 59% (3').

Drivers in MCL

To discover genomic alterations involved in MCL lymphomagenesis we integrated all mutations, CNAs and SVs detected in the 61 WGS sequenced in this study, together with 21 nonoverlapping and previously analyzed WES cases (74% cMCL and 26% nnMCL)[7]. In this sense, for the discovery of new driver genes a frequency-based strategy was followed. Thus, due to the different mutational burden of different genomic loci, we calculated the mutational background for each locus using the number of non-coding mutations in a 0.5 Mb window centered in the gene. To compute the number of expected mutations in the coding region, we considered the length of each gene, as well as the coverage and the callability of the coding region along all analyzed samples. The number of affected cases for each gene was calculated considering missense and nonsense mutations, non-coding mutations affecting the functionality of the gene as well as other associated events, such as amplifications or deletions. This procedure allowed us to identify 26 genes significantly altered in the whole cohort.

Among them we confirmed some of the previously described driver genes in MCL, such as *ATM* (48%), *CCND1* (44%) with exon1/intron1 somatic mutations (26%) and/or 39 UTR activating alterations (21%), *TP53* (26%), *KMT2D* (23%), *RB1* (23%), *BIRC3* (22%), *CDKN2A* (21%), *SP140* (13%), *NSD2* (12%), *BMI1* (11%), *MIR17HG* (10%), and *UBR5* (6%). However, this analysis also revealed 7 novel MCL driver genes altered by missense or truncating mutations and deletions, including *CDKN1B* (12%), *SAMHD1*

(10%), *BCOR* (9%), *SYNE1* (6%), *HNRNPH1* (6%), *SMARCB1* (4%), and *DAZAP1* (4%). Furthermore, 4 genes mutated at lower frequency, but carrying known driver alterations (*NOTCH1*, *NOTCH2*, *TLR2*, and *PAX5*) were also detected.

In this regard, it must be noted that despite the increase in the number of cases analyzed over previous studies, only a few novel driver genes were discovered. Moreover, this was achieved mainly to the identification of SVs and CNAs, which play a major role in this pathology, while point mutations played a more limited role in the identification of novel driver genes. Additionally, with respect to CNAs, 13 chromosomal regions without a defined target gene were recurrently altered, most of them corresponding to deletions. Overall, 81 of 82 (99%) MCL cases had at least 1 driver alteration in addition to the t(11;14) (median, 6). Collectively, 8 main pathways were frequently altered in MCL including proliferation, cell survival, DNA damage response, telomere maintenance, chromatin remodeling, B-cell receptor/Toll-like receptor/NF-kB signaling, NOTCH signaling, and RNA regulation (Figure 12).

Finally, most MCL drivers were found to be preferentially altered in cMCL cases, with the unique exception of somatic hypermutation in *CCND1* mainly found in nnMCL. Of note, *ATM* alterations (64%); deletions of 1p, 10p, and 19p; and gain of 7p were exclusively seen in cMCL, whereas *TP53* and *TERT* alterations were the only drivers slightly enriched in nnMCL, with all 5 cases with *TERT* alterations carrying concomitant *TP53* aberrations. Altogether, cMCL cases had a significant higher number of driver alterations than nnMCL (median, 7 vs 2), what might explain the different evolution course of both subtypes of MCL.

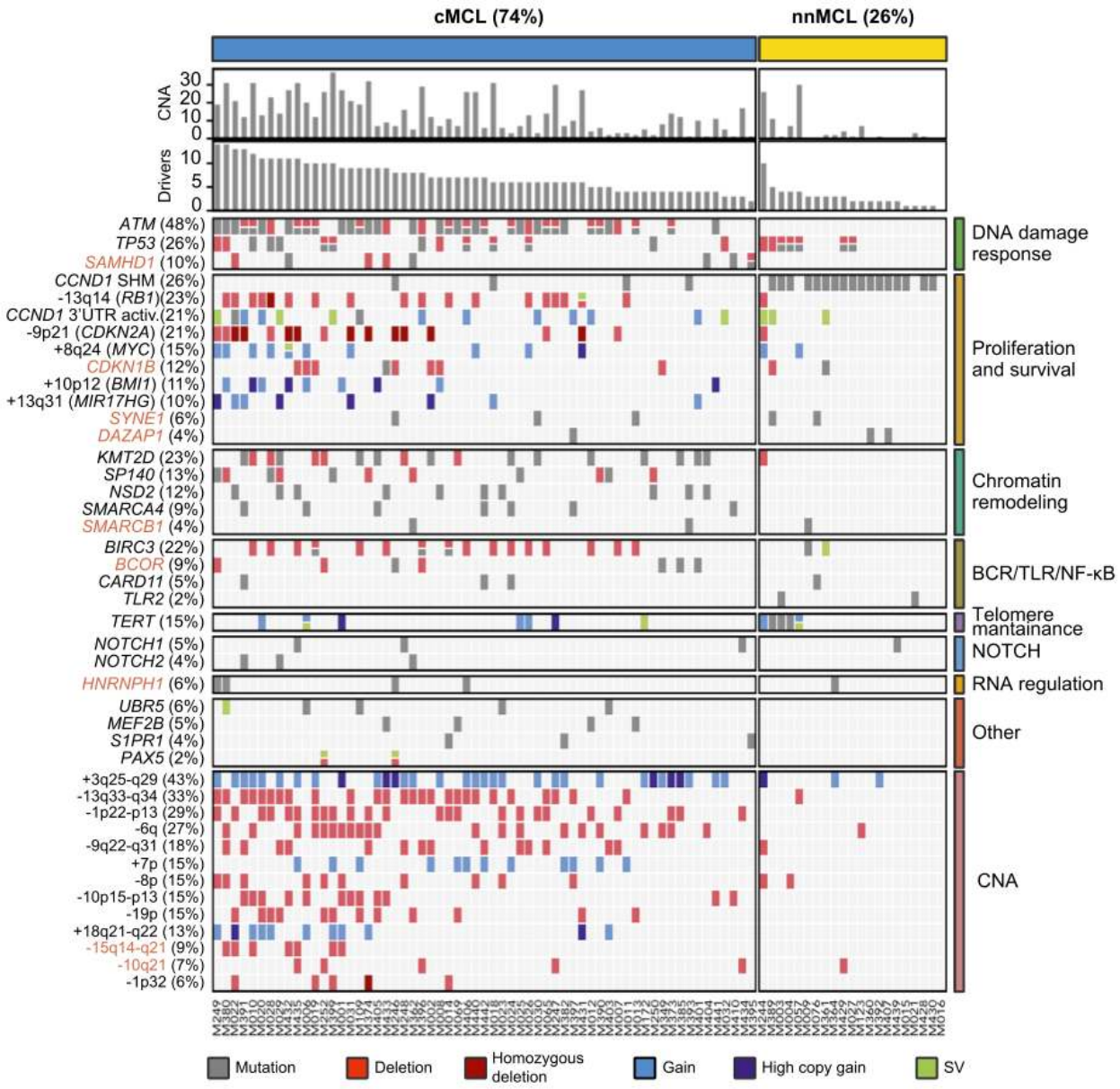

Figure 12. Oncoprint representation of the 43 driver alterations identified in MCL. Drivers are depicted in rows and group according to pathways in which they are involved. Cases are displayed in columns. Novel driver alterations identified in this study are highlighted in dark orange. From Nadeu et al. 2020.

Development of a machine learning based tool for the detection of somatic mutations

The identification of cancer driver genes, as well as secondary analysis such as mutational burden and mutational signatures depends on the performance of somatic mutation callers. In the case of the PCAWG analysis, the most comprehensive analysis of tumor genomes to date, we noticed that this collective effort used up to 5 variant callers for the identification of somatic mutations. Due to the importance of this step, and the large computing resources required for this redundant analysis, we thought that it was necessary to develop an easy-to-use tool with fewer requirements so that more modest groups and even clinical centers could also carry out this type of analyses in a reasonable amount of time. Hereby, our previous experience during the analysis and review of MCL whole genomes allowed us to generate a database with manually curated mutations that could be used to train a machine learning algorithm, a promising technology that offered many possibilities. We have thus developed RFcaller, an algorithm based on read-level features that uses machine learning for somatic mutation detection in paired normal/tumor samples.

RFcaller workflow

An overview of the RFcaller's workflow is provided in Figure 13. The pipeline takes as input the BAM files from the normal-tumor paired samples and starts performing a basic variant calling using bcftools (v1.10.2) with the -P option set to 0.1 to enable calling of low frequency variants. Then, indels are normalized, and common SNPs (dbSNP v153), and variants within five base pairs of an indel, are removed. To increase the speed of the pipeline, low quality calls are filtered out (<15 for SSNVs and <40 for indels). Remaining mutations are divided into three different files to be processed independently: SSNVs, short-indels (<7bp) and long-indels (≥7bp).

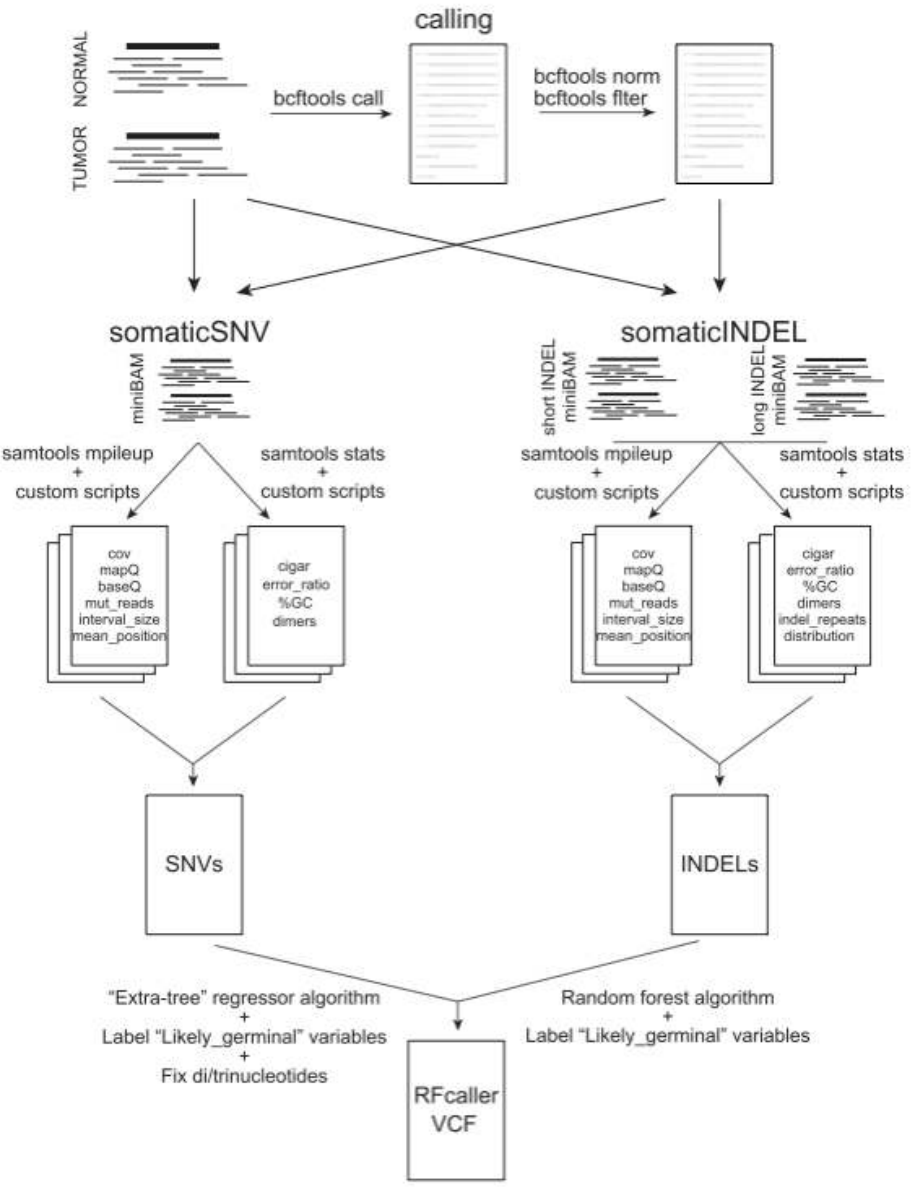

Figure 13. Flowchart of the RFcaller pipeline.

SSNVs and indels have a specific pipeline where read-level features are extracted for those mutations that meet basic requirements that can be customized, such as having a minimum coverage (≥7), a maximum number of mutated reads in normal (≤3 for SSNVs and ≤2 for indels) or a minimum number of mutated reads in tumor (≥3 for SSNVs and ≥4 for indels). These filters have been chosen because positions that fail to meet these requirements cannot be confidently classified as *bona fide* mutations from the available data. Once all features have been extracted, a CSV file is generated to be used by the algorithm. The result is a VCF file with the mutations that have passed the threshold for the "QUAL" field.

To classify mutations that might be germinal but have passed the previous filters, a 95% confidence interval is applied to calculate the number of mutant reads expected in the normal sample, considering: the VAF of the mutation in tumor sample, the expected contamination of tumor in normal sample defined by the user and the normal coverage. Thus, if the number of mutated reads in normal is greater than the expected, the position is labeled as "LIKELY_GERMINAL".

Finally, the RFcaller pipeline for SSNVs searches for dinucleotide or trinucleotide mutations within the results. With this step, if two mutations are found together in cis, they are merged into a single mutation to be accurate when predicting its functional effect, a step that is usually missed by most commonly used somatic callers, resulting in incorrect predictions of the potential functional effect.

RFcaller training

For the initial training step, previous results from the genomic analysis of two mantle cell lymphomas[144] were used to annotate the set of mutations, and RFcaller was trained with this initial dataset. The obtained results were compared with those used for training, and all discordant positions were manually reviewed to improve the accuracy of the dataset. These steps were repeated until all discrepancies were classified by an expert panel. After that, 2,208 and 2,901 calls were reviewed for training and testing, respectively, resulting in a high quality set of mutations to train and test the final versions of the algorithm (Table 6).

Table 6. Number of total and manually reviewed mutations used for training and testing RFcaller

	Training set				Test set			
	SSNV		Indel		SSNV		Indel	
	TP	TN	TP	TN	TP	TN	TP	TN
Manually reviewed	915	730	321	242	924	959	528	490
Total	8,362	57,734	504	427	6,909	57,039	696	1,810

TP: Number of True Positive mutations; TN: Number of True Negative mutations

In order to select the best cutoff for the pipeline, SSNVs, indels and homopolymer indels were considered independently as they represent mutations whose detection is influenced by different features. The distinction between both types of indels (isolated or within a homopolymer trait) was introduced due to the bias of the initial calling performed by bcftools against indels within homopolymeric tracts, giving very low scores to mutations that otherwise appeared to be real. Furthermore, different formulas were considered to calculate the "QUAL" threshold used by RFcaller (Table 7).

Table 7. Metrics given by OptimalCutpoints for each formula and mutation type

	Formula	OptimalCutpoint*	AUC	IC95%	TP (TPR)	FP (FPR)	TN (TNR)	FN (FNR)	F1
SSNVs	RF	0.3179	0.999	0.998 0.999	6681 0.989	55 0.023	2334 0.977	75 0.011	0.9903
	BCF*RF	14.4152	0.998	0.997 0.999	6714 0.994	78 0.033	2311 0.967	42 0.006	0.9911
	BCF*(RF**2)	<u>10.726</u>	0.999	0.998 0.999	6685 0.99	31 0.013	2358 0.987	71 0.011	0.9924
	BCF**RF	4.2335	0.999	0.998 0.999	6683 0.989	39 0.016	2350 0.984	73 0.011	0.9916
Indels	RF	0.6232	0.94	0.926 0.953	507 0.888	72 0.099	652 0.901	64 0.112	0.8817
	BCF*RF	140.9368	0.848	0.828 0.869	391 0.685	92 0.127	632 0.873	180 0.315	0.7419
	BCF*(RF**2)	76	0.909	0.893 0.926	458 0.802	88 0.122	636 0.879	113 0.198	0.82
	BCF**RF	<u>32.1418</u>	0.936	0.922 0.95	504 0.883	52 0.072	672 0.928	67 0.117	0.8944
Homopolymer indels	RF	<u>0.7723</u>	0.98	0.954 1	95 0.99	4 0.08	46 0.92	1 0.01	0.9743
	BCF*RF	3.0406	0.814	0.735 0.893	95 0.99	23 0.46	27 0.54	1 0.01	0.8878
	BCF*(RF**2)	3.3084	0.935	0.89 0.981	92 0.958	9 0.18	41 0.82	4 0.042	0.934
	BCF**RF	3.0395	0.947	0.906 0.987	94 0.979	8 0.16	42 0.84	2 0.021	0.9494

**Selected cutoffs are underline.*

AUC: Area Under the Curve; IC95%: Interval Confidence 95%; TP: Number of True Positive mutations; FP: Number of False Positive mutations; TN: Number of True Negative mutations; FN: Number of False Negative mutations; TPR: True Positive Rate; FPR: False Positive Rate; TNR: True Negative Rate; FNR: False Negative Rate; F1: F1-score

Although the RFcaller score provided high accuracy, we observed that by combining the regression obtained by RFcaller with the score given by bcftools, the accuracy was improved over each one independently, suggesting both scores complement each other. We did not observe major differences between formulas for SSNVs and indels according to the AUC, so we selected the formulas with the highest F1 score. Thus, the cutoffs were 10.726 for SSNVs, 32.1418 for indels and 0.7723 for homopolymer indels (Figure 14), which achieved 1.3%, 7.18% and 8% of false positive mutations, respectively. We observed that many of the false positives belonged to complex regions like microsatellites or GC-rich sites, appearing also in normal samples from other donors. Therefore, we used a panel of normals to filter these calls and improve the accuracy of the pipeline.

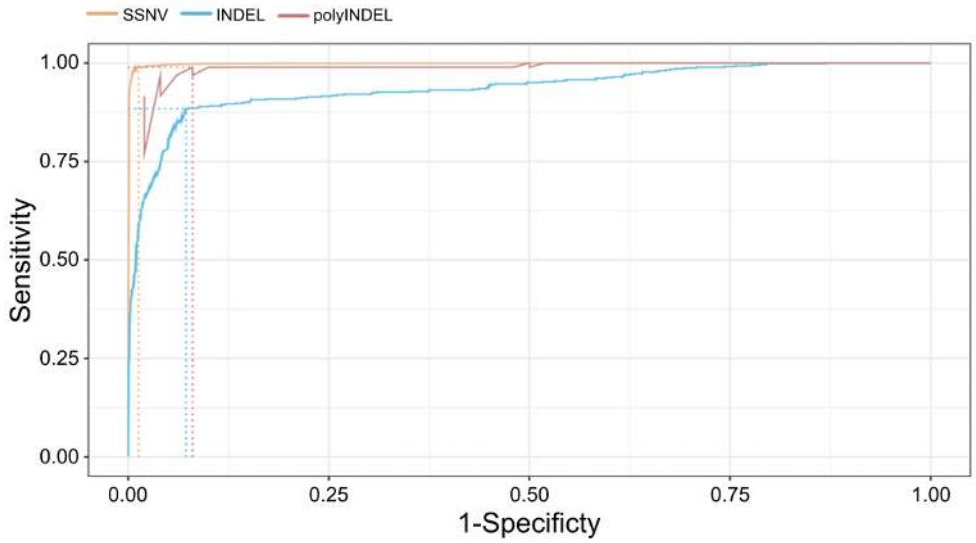

Figure 14. ROC curves for each mutation type with RFcaller results for the test data set using the formulas with the best F1 score. Cutoffs were obtained with the MaxEfficiency criterion. polyINDEL, Homopolymer indel.

In terms of the number of variables selected, only 16 and 27 read-level features were considered for SSNVs and indels respectively, which helped us to avoid overlapping features that can be counterproductive and lead to overfitting. Another important aspect we considered during the selection of these features was the difficulty by which they can be extracted, resulting in a fast pipeline for medium-size servers. Thus, the analysis of four WGS tumor-normal paired samples using 20 threads consumes only ~5 GiB of RAM and takes ~3 hours/case, while using only 10 processors the analysis is extended up to ~4.5 hours/case (Figure 15d).

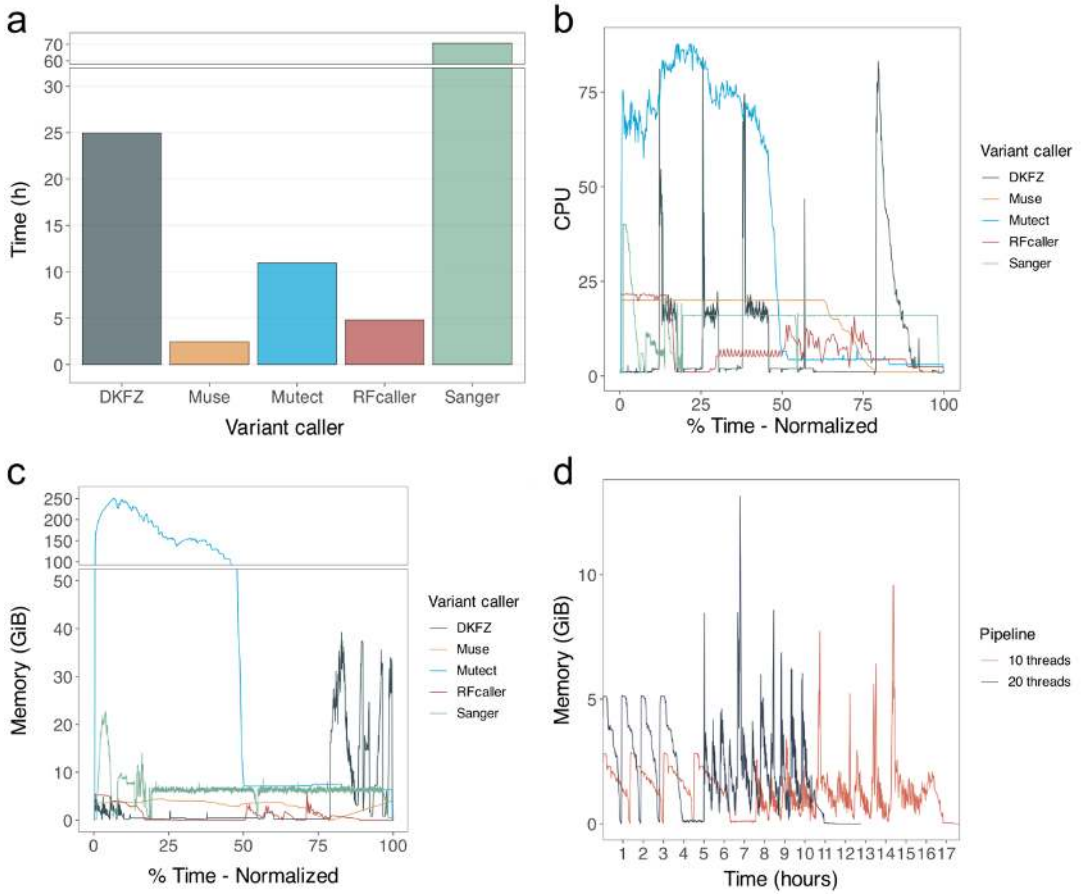

Figure 15 Representation of the computational cost of running RFcaller and the other callers used by the PCAWG. a) the time it takes for each caller to complete a case. B, c) CPU and memory (RSS) usage for the duration of the pipelines (RSS memory was limited to 200Gb when the caller allowed it). D) memory (RSS) usage of RFcaller pipeline when it is run with 10 or 20 threads for four independent cases simultaneously.

When RFcaller was compared with the callers used by PCAWG for the detection of SSNVs, only the muse-variant-caller (~2.5 h) was faster than RFcaller (~4.8 h) (Figure 15), while sanger-variant-caller was the slowest, taking more than 70 h for a single case. In terms of memory consumption (RSS), mutect-variant-caller is the most demanding, consuming between 100 GiB and 250 GiB during half the time it is running (~5 h). In this case, RFcaller and muse-variant-caller consume the least memory with an average of 5 GiB. It is important to note that although we have used the SSNVs specific callers, all of them, except MuSE, also detect indels, which would imply that RFcaller is the fastest and least resource consuming tool for the simultaneous calling of SSNVs and indels.

Validation of RFcaller pipeline: PCAWG analysis

To test RFcaller against a validated set of cancer WGS cases we used data from the PCAWG study belonging to two different projects (CLLE-ES and BRCA-EU), representative of liquid and solid tumors, with a total of 89 and 75 cases, respectively. RFcaller results were compared to those mutations labeled as "PASS" by the PCAWG mutation calling pipeline. Due to the inherent differences between SSNVs and indels, we performed each analysis independently.

Somatic Single Nucleotide Variants

After merging RFcaller and PCAWG "PASS" mutations, we observed that ~70% of SSNVs were detected by both pipelines in both studies. However, and even though the number of shared mutations was almost the same, for samples from the CLLE-ES project 11% of mutations were detected only by the PCAWG pipeline vs. 16.3% mutations specifically detected by RFcaller. For BRCA-EU-derived mutations, only 4.4% mutations were RFcaller-specific, vs. 25.4% for PCAWG pipeline (Figure 16a). A detailed analysis of those differentially called mutations revealed that the mean VAF for SSNVs detected by both pipelines was 0.41 and 0.27 for CLLE-ES and BRCA-EU, respectively (Figure 16c). However, those detected by the PCAWG pipeline but not RFcaller had a mean VAF of 0.16 and 0.10 for CLLE-ES and BRCA-EU, respectively (Figure 16c), suggesting that they constitute subclonal mutations. In fact, only 29% and 50% of them could be detected by more than two callers in the PCAWG pipeline for CLLE-ES and BRCA-EU, respectively (Figure 16e). Furthermore, those SSNVs detected by RFcaller but not the PCAWG pipeline had a mean VAF of 0.46 for CLLE-ES and 0.28 for BRCA-EU, similar to those detected by both pipelines, suggesting that they constitute clonal mutations detected by RFcaller. Some of them showed minor tumor in normal contamination (1-3 mutant reads), common in hematological tumors, resulting in most callers missing these true positive somatic mutations, while RFcaller is able to retain most of them.

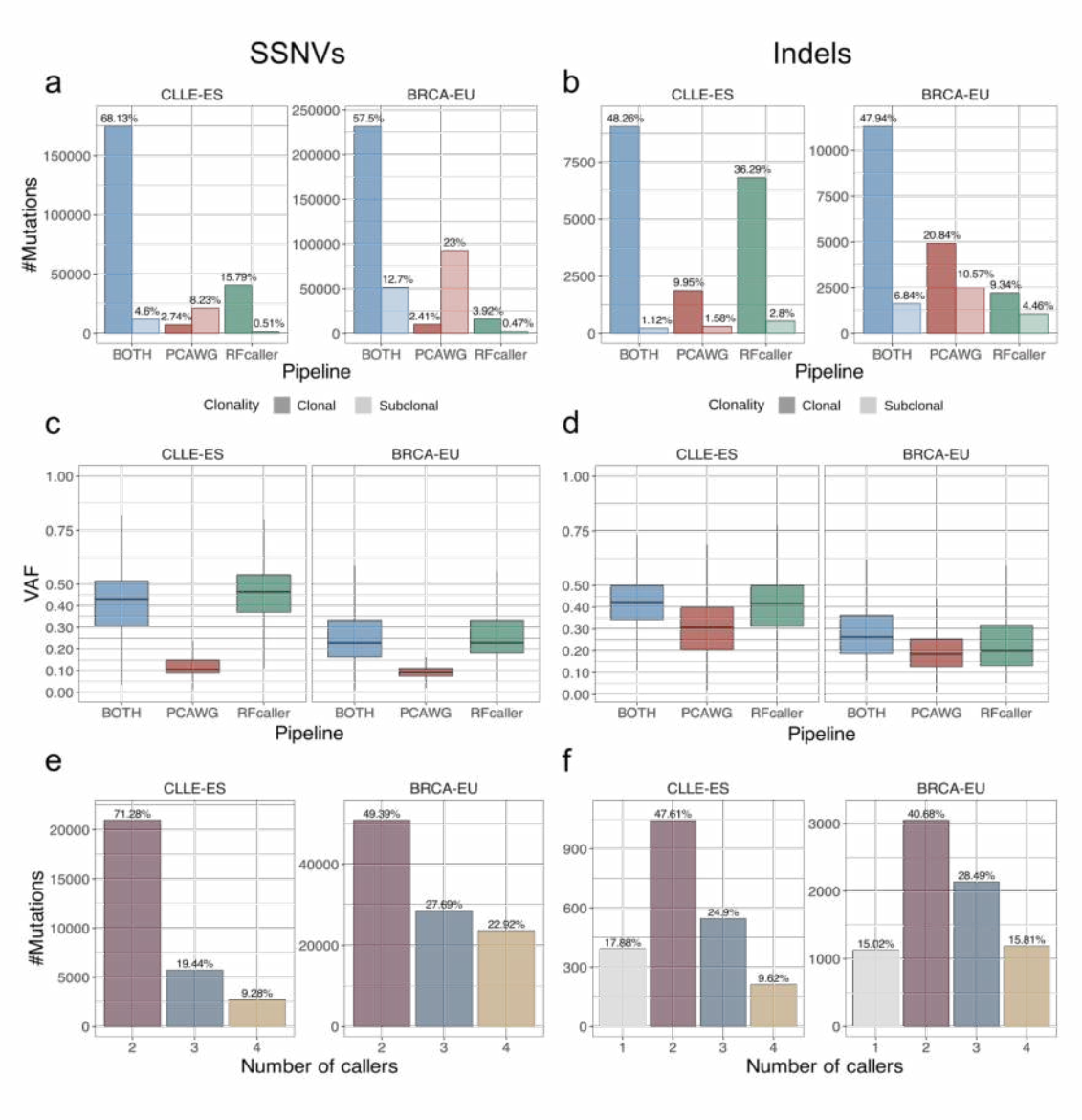

Figure 16. Summary of mutations detected by PCAWG and/or RFcaller pipelines for SSNVs and indels. a,b) Classification of mutations according to the pipeline that can detect them. Mutations are divided in clonal (VAF≥0.15) and subclonal (VAF<0.15) mutations. c, d) Distribution of the variant allele frequency of the mutations identified by both pipelines, or specifically by RFcaller or PCAWG pipeline. e, f) Number of callers detecting each of the PCAWG-private mutations.

To explore the set of discordant mutations between both pipelines, we randomly selected 1-2% of the pipeline-private calls (n = 776 for CLLE-ES and n = 1,233 for BRCA-EU) to be manually reviewed by a panel of experts. As expected, PCAWG-specific variants detected by four callers are more precise than those identified by two tools (Table 8). Surprisingly, the difference in precision for RFcaller-private mutations between studies was very high, 98.5% for CLLE-ES and 74.5% for BRCA-EU, probably

reflecting the fact that RFcaller was trained using a hematological tumor. However, despite the apparently higher number of false positives, RFcaller-private calls only represent 18.3% and 5.9% of the total number of SSNVs detected by RFcaller in the CLLE-ES and BRCA-EU projects, respectively. Considering the observed number of false positive calls within these sets, the real precision of RFcaller calls for SSNVs is 99.7% and 98.5% for CLLE-ES and BRCA-EU, respectively, while the precision of the PCAWG pipeline is 97.3% for both studies (Figure 17).

Table 8. Total number of false positive private SSNVs extrapolated after manual revision

	CLLE-ES				BRCA-EU			
	SSNVs	TP	FP	Precision	SSNVs	TP	FP	Precision
2 callers	20,956	16,097	4,859	76.81%	50,823	42,804	8,019	84.22%
3 callers	5,715	5,001	714	87.50%	28,490	26,469	2,021	92.91%
4 callers	2,729	2,519	210	92.31%	23,580	23,080	500	97.88%
RFcaller	41,772	41,153	619	98.52%	17,699	13,189	4,510	74.52%

TP: Number of True Positive SSNVs; FP: Number of False Positive SSNVs

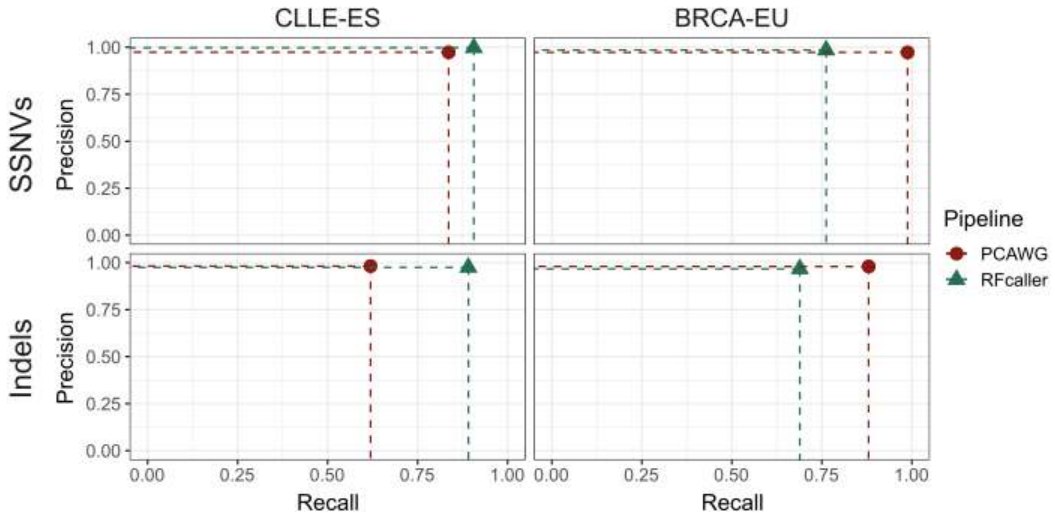

Figure 17. Accuracy of RFcaller and PCAWG pipelines for SSNVs and indels against CLLE-ES and BRCA-EU datasets. RFcaller shows a higher recall in both SSNVs and indels for CLLE-ES, whereas in BRCA-EU the PCAWG manages to detect a higher number of mutations. The precision of the two pipelines is similar in all conditions.

To further explore these private mutations, we extracted the mutational signatures independently for the set of mutations detected by both pipelines, as well as for those specific for each caller (Figure 18). We could see that in CLLE-ES study, both RFcaller-private SSNVs and those common to both pipelines contained the same signatures (SBS1, 5, 8 and 9), while PCAWG-private SSNVs shared 3 signatures (SBS1, 5

and 8), missed one (SBS9) and contained two signatures not detected in the common set (SBS23 and 51), although affecting a limited number of samples. While on the BRCA-EU study, both pipelines missed some signatures present in the common set (2 PCAWG and 4 RFcaller), and both detected one and two signatures, respectively, not present in the common. Together, these results suggest that the private mutations detected by RFcaller constitute *bona fide* calls, with a similar profile to those detected by both pipelines.

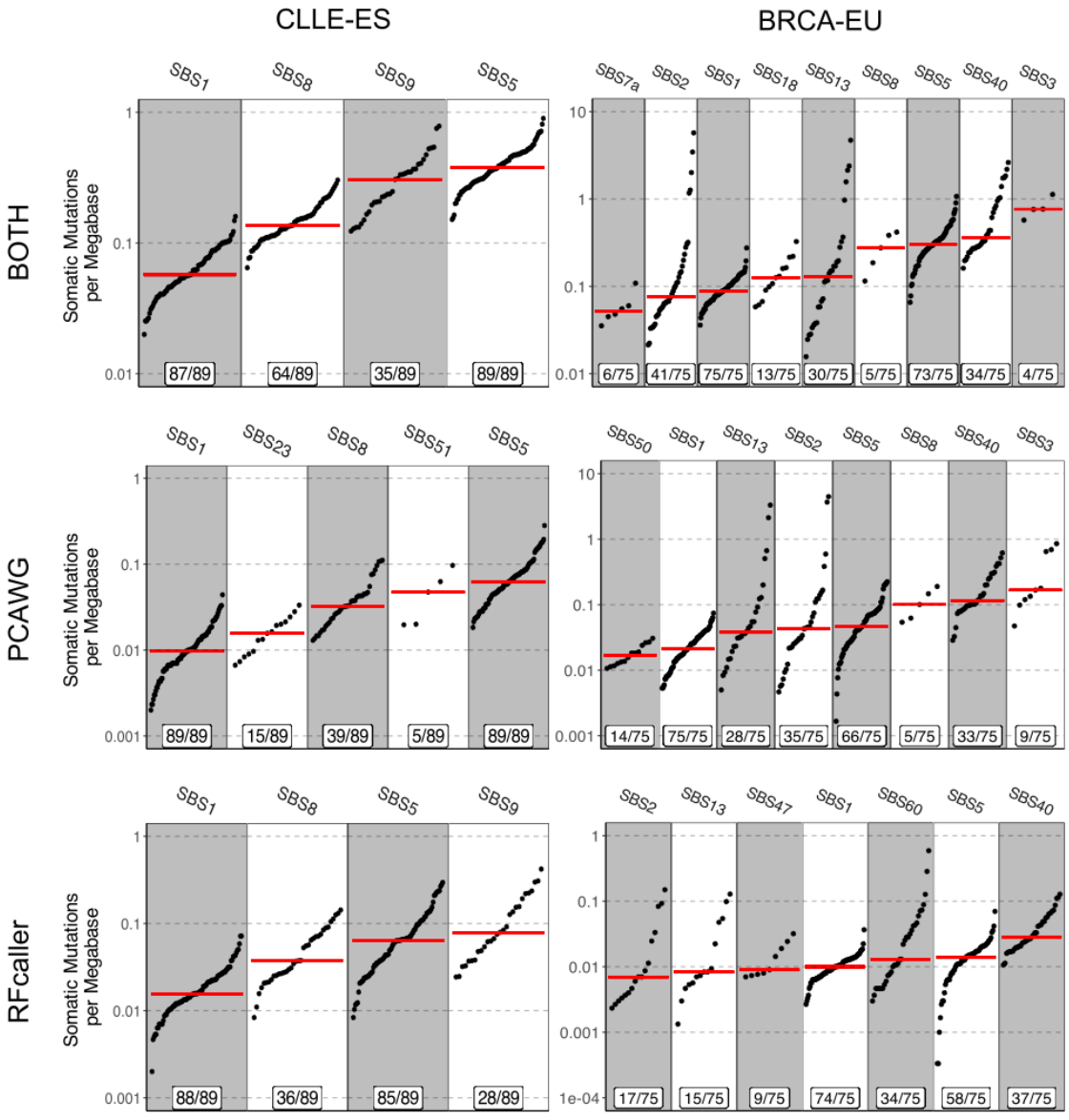

Figure 18. Mutational signatures extracted for CLLE-ES and BRCA-EU studies using the set of mutations detected by both pipelines and RFcaller and PCAWG-private SSNVs independently. At the bottom of each lane the number of patients presenting that signature with respect to the total is indicated. The median number of mutations per megabase for each signature is shown in red.

Small insertions/deletions

The analysis of small indels revealed that there were more differences between pipelines than those seen for SSNVs. In this regard, only ~50% of indels were detected by both RFcaller and PCAWG pipelines, however for CLLE-ES RFcaller-private calls represented 39.1% of the total number of indels whereas only 11.5% of them were PCAWG-specific. In contrast, in BRCA-EU, RFcaller and PCAWG-private mutations accounted for 13.8% and 31.4% respectively (Figure 16b). Moreover, among them, less than 45% of PCAWG-private indels were detected by more than two callers (Figure 16f), reflecting the difficulty to identify somatic indels in tumor samples.

To further explore pipeline-private indels, we selected at least 50 indels from each group for expert review (n = 283 for CLLE-ES and n = 429 for BRCA-EU). We observed that the precision within PCAWG-private indels was very high, varying between 70% and 99% depending on the number of individual callers supporting the call (Table 9). In contrast, the precision observed for RFcaller was 89%, despite the fact that the total number of indels detected by this pipeline was much higher. Similar to SSNVs, the observed VAF was slightly higher in CLLE-ES compared to BRCA-EU (0.42 vs 0.29), probably reflecting higher tumor purity. Nonetheless, we did not observed differences in VAF between pipeline-private indels (Figure 16d), suggesting that pipeline-specific mutations were not due to clonality, as they were for SSNVs, but to other factors such as alignment issues, size of the indel, the presence of microsatellites or if they were within homopolymer tracks.

Table 9. Total number of false positive indels extrapolated after manual revision

	CLLE-ES				BRCA-EU			
	Indels	**TP**	**FP**	**Precision**	**Indels**	**TP**	**FP**	**Precision**
1 caller	392	276	116	70.37%	1,125	1,008	117	89.57%
2 callers	1,044	1,002	42	96.00%	3,046	2,792	254	91.67%
3 callers	546	495	51	90.74%	2,133	2,097	36	98.31%
4 callers	211	207	4	98.18%	1,184	1,174	10	99.14%
RFcaller	7,335	6,916	419	94.29%	3,257	2,721	536	83.54%

TP: Number of True Positive indels; FP: Number of False Positive indels

Despite the higher precision obtained by the PCAWG pipeline for indel calling, this might be at the expense of a larger number of false negative calls in otherwise clonal and *bona fide* somatic indels, as shown by the number of true positive calls detected by RFcaller (Figure 17). In the case of mutational signatures detected for indels (Figure 19), private calls from both pipelines contained most of the mutational signatures present in common mutations, while in the case of RFcaller-private indels, 3 signatures not present in the common set were detected for both CLLE-ES and BRCA-EU studies.

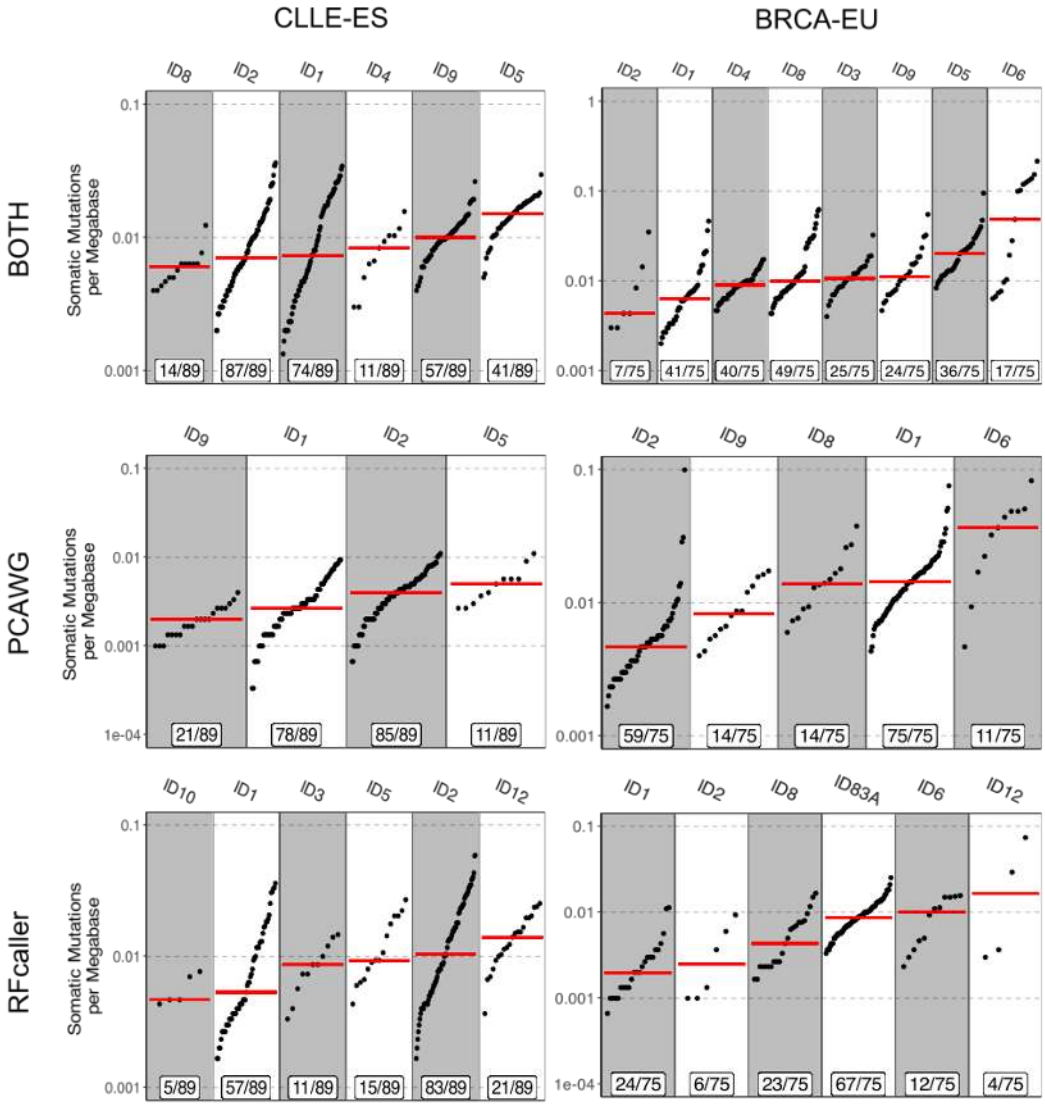

Figure 19. Mutational signatures extracted for CLLE-ES and BRCA-EU studies using the set of mutations detected by both pipelines and RFcaller and PCAWG-private indels independently. At the bottom of each lane the number of patients presenting that signature with respect to the total is indicated. The median number of mutations per megabase for each signature is shown in red.

Exome analysis

RFcaller was trained with WGS data, but as the features used for the prediction are at read level, this pipeline could also be used for exome analysis. In order to test the ability of RFcaller to detect mutations by WES, exomes from five cases previously analyzed by WGS were run with default parameters. Results were compared with those obtained by RFcaller and PCAWG in the WGS analysis after filtering for mutations within target regions in WES. Thus, 63% (n = 110) of mutations detected by WES were also detected by WGS. Additionally, we were able to identify 47 novel mutations for which there was neither coverage nor any mutated read in WGS (Figure 20a). When we made the comparison in the opposite direction, we found that 55% (n = 136) of the mutations detected by WGS did not appear by WES. However, 93% (n = 126) of these missing mutations had no coverage or any mutated read in the exome or were clearly germinal (Figure 20b). Only 10 mutations detected by WGS had enough coverage in WES and were not detected, constituting false negatives (RFcaller exome recall = 94%). Similarly, considering the 17 mutations that were labeled as germinal by WGS but detected by WES as false positives, RFcaller achieves a precision of 90%.

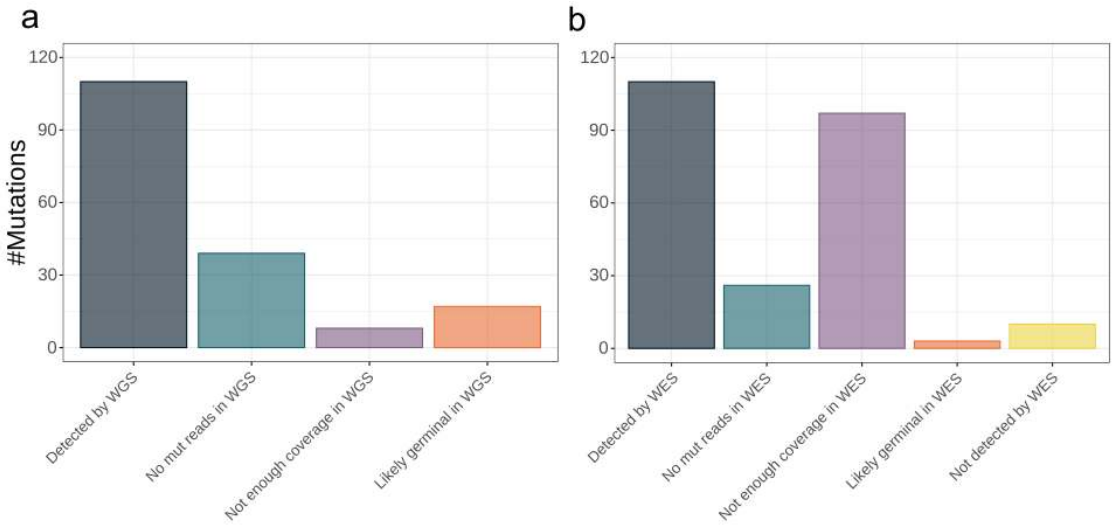

Figure 20. Comparison of mutations detected by analysis of WGS and WES in selected donors. Comparison is limited to exomic regions. a) Mutations detected by WES and analysis of their status in WGS. b) Mutations detected by WGS and analysis of their status in WES samples.

Detection and verification of mutations in driver genes

From the above data we can conclude that RFcaller has a similar accuracy to detect SSNVs, and an increased sensitivity to detect indels at the cost of a smaller specificity. To explore if these differences might allow the detection of previously missed mutations with potential clinical impact, we analyzed somatic mutations on the set of driver genes previously described in these two tumor types (Knisbacher, B et al. 2022 (in press))[53]. This analysis resulted in the identification of 155 coding mutations in driver genes in the CLLE-ES project and 162 in the BRCA-EU study. Out of those calls, 83% of them were shared between both pipelines, while 53 (17%) in 35 driver genes, were pipeline-specific.

Those pipeline-specific mutations were manually reviewed, resulting in the identification of 19 clonal mutations detected by RFcaller (12 SSNVs and 7 indels) vs. 4 clonal SSNVs detected by the PCAWG pipeline in CLLE-ES. For the BRCA-EU project, 8 clonal mutations were detected by RFcaller (5 SSNVs and 3 indels) vs. 4 clonal detected by PCAWG pipeline (3 SSNVs and 1 indel).

For seven private calls detected in the CLLE-ES study (5 by RFcaller, and 2 by the PCAWG pipeline), tumor and normal DNA was available for verification by Sanger sequencing, except two cases in which only tumor DNA was available. This analysis resulted in the verification of all RFcaller-private calls (Figure 21), as well as one of the PCAWG-private SSNVs. The last call could not be verified because it was a subclonal mutation with a very low VAF (8.7%), that falls below the detection limit of this technique.

To further perform an orthogonal validation of these pipelines, we took advantage of a previous study in which 26 CLL driver genes had been analyzed by deep-sequencing in some of the CLL cases used by PCAWG[133,134]. A total of 77 mutations, excluding germline calls, were detected in 28 cases, for which enough coverage was available in WGS to make a call. Due to the high depth of sequencing, VAF was very variable (range 0.0029 to 0.9665), therefore, mutations were classified as clonal if VAF≥0.15 (n = 44, median 0.43), and subclonal if VAF<0.15 (n = 33, median 0.03). As

expected, most subclonal mutations could not be detected from WGS data, as each pipeline was only able to detect 6/33 subclonal mutations (18%). By contrast, most clonal mutations detected by deep sequencing could be identified by RFcaller (39/44, 89%), while the performance of the PCAWG pipeline was slightly lower (31/44, 70%).

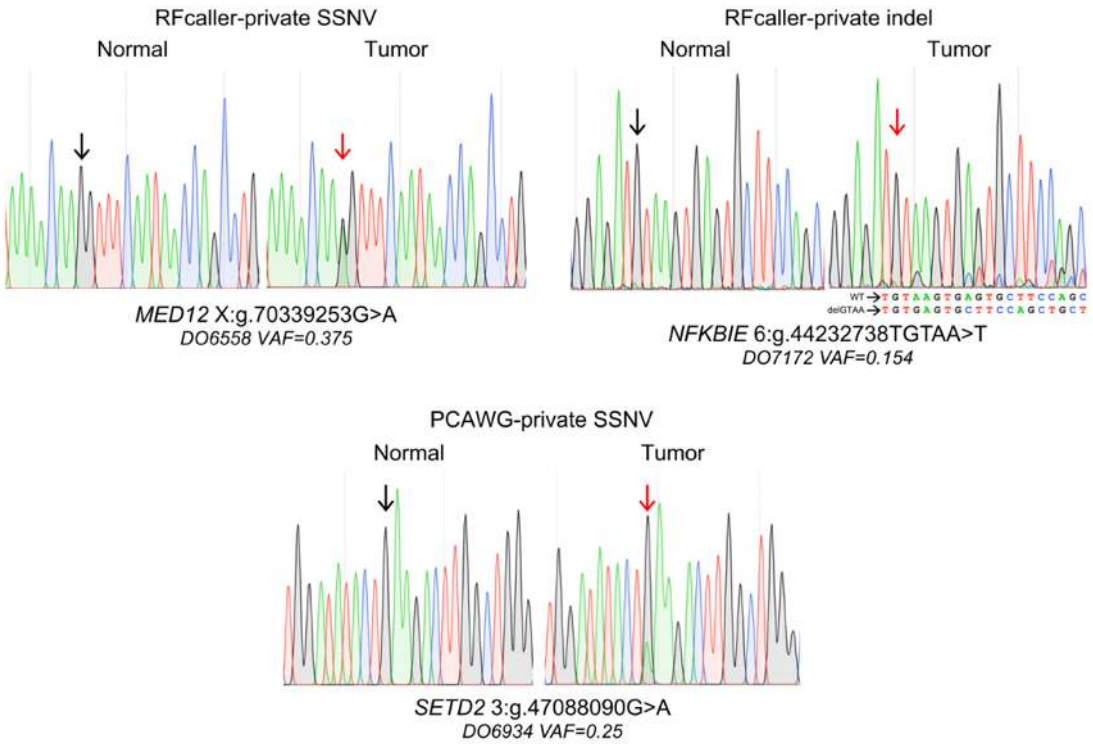

Figure 21. Electropherograms corresponding to Sanger verification of private mutations detected by RFcaller and PCAWG in CLL driver genes. Black arrows correspond to wild-type positions, whereas red arrows indicate the mutations. For the RFcaller-private indel, wild-type and mutated sequences appear below the electropherogram.

The mutations specifically detected by RFcaller affected *NOTCH1* (3), *ATM* (2), *TP53*, *RPS15*, *MGA* and *DDX3X*, some of which have been associated with poor prognosis and whose presence might impact clinical decisions. The PCAWG pipeline was able to identify a mutation in *ATM* that was not detected by RFcaller due to a very low VAF (0.065). Together, these results support the utility of RFcaller to identify novel clonal driver mutations of potential clinical value.

Circulating tumor DNA from the cerebrospinal fluid allows the characterization and monitoring of medulloblastoma

The importance of somatic mutation callers in the clinical practice depends on the increasing relevance of molecular classification of tumors for prognosis and clinical decisions. In this regard, medulloblastoma constitutes the most prevalent malignant brain tumor in childhood and can be subdivided into different subtypes with very diverse prognosis[145,146]. Accordingly, a risk-classification scheme has been established to stratify patients in four groups with a 5-year progression-free survival of 91% (favorable), 81% (standard), 42% (high risk) and 28% (very high risk)[145,147]. To achieve this characterization, both clinical and molecular features are needed.

As first line, surgical specimens are used for diagnosis and tumor characterization. However, MB tumors evolve with time highlighting the need for longitudinal samples for disease monitoring. In this regard, a relatively noninvasive method, such as liquid biopsies and ctDNA, is required to assist the clinical management of MB patients. However, even though the detection of ctDNA in CSF, which is obtained routinely for cytologic analysis, was previously reported in adult brain malignancies[148–151], it has not yet been explored in the pediatric setting.

Classification of MB patients

In order to assess whether the analysis of CSF ctDNA could be a tool for the diagnosis and monitoring of MB patients, the group of Dr. Joan Seoane at VHIO collected tumor, blood and CSF samples from 13 patients, and whole-exome sequencing was performed in those samples. My involvement in this project was the analysis of somatic and germline mutations, as well as copy number alterations, derived from the exome sequencing of tissue biopsies and normal DNA, as well as from CSF ctDNA. To carry out the initial classification of patients (Table 10, Figure 22a) into previously defined MB subgroups and subtypes[145,146], common MB germline and somatic driver mutations (median 0.36 mutations/Mb), focal copy number alterations, arm-level and whole chromosome gain/loss were identified from WES from tumor-normal pairs (Figure 22b).

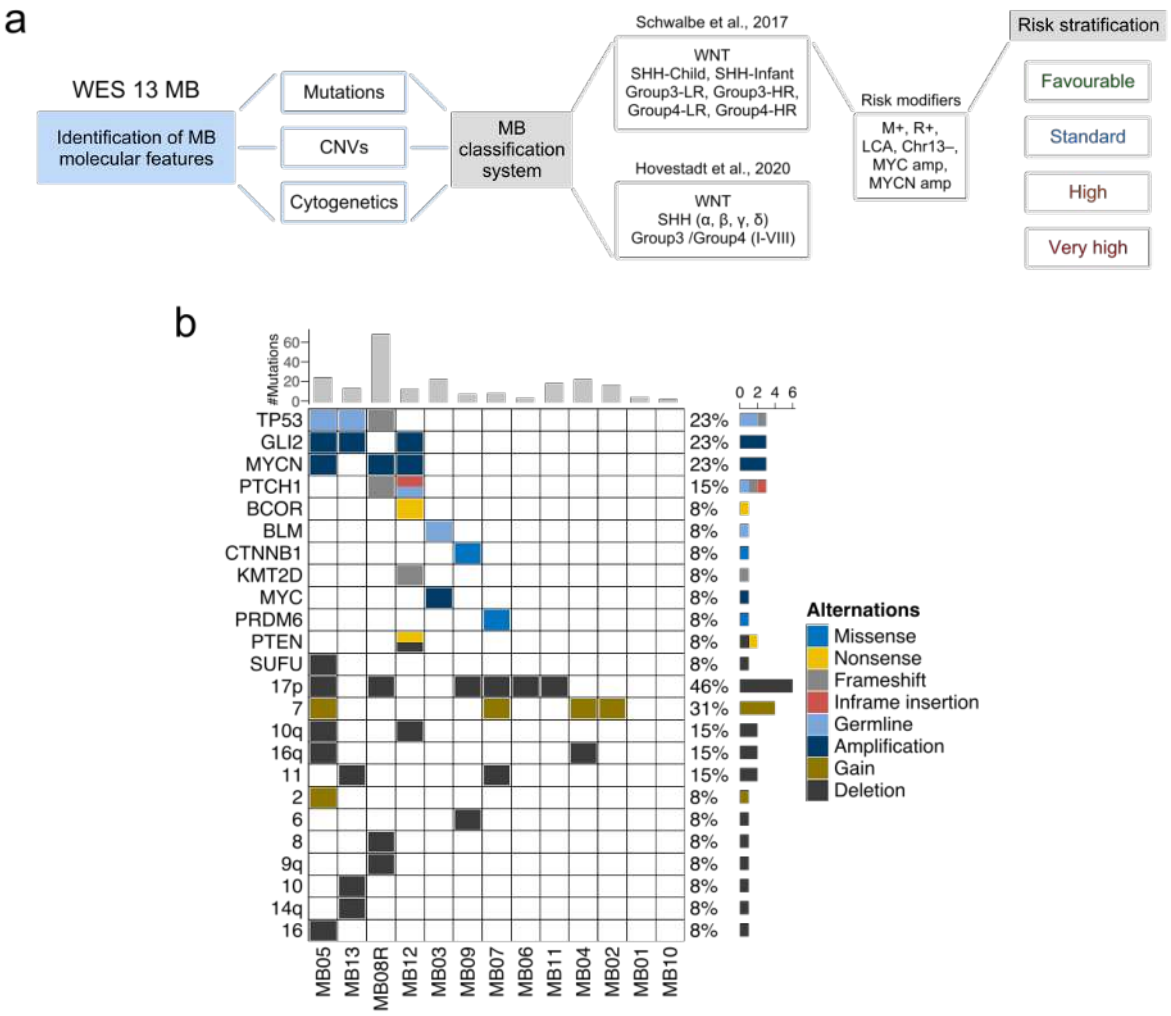

Figure 22. Project outline and characterization of a cohort of 13 pediatric patients with MB. a) Summary of the MB common molecular alterations investigated for the classification into subgroup and association of risk stratification. b) Oncoprint with molecular features, including driver events in MB, identified from the WES characterization of 13 MB patients. The number of mutations identified for each patient is indicated in the top and the proportion of patients with each alteration on the right. Distinct alterations are indicated in the color legend.

Table 10. Clinicopathological characteristics, molecular features, MB-subgroup classification and risk stratification (based on Schwalbe et al., 2017 5-year progression-free survival (PFS)) for the 13 MB patients

	Demographics		Clinical features								Molecular features					MB subgroup classification model		
	Age	Sex	H*	Tumor location	Vol (cm3)	Pathology	M status	R status	Treatment protocol	PFS (years)	Germline mutations	Mut/ Mb	Mutations	CNAs	Cytogenetics	Hovestadt et al., 2020	Schwalbe et al., 2017	Risk stratification
MB1	5	M	EVD ETV	Vermis	63.7	LCA	M-	R+	SIOPE high risk	5.46	----	0.12	----	----	----	----	----	----
MB2	13	F	EVD ETV	Vermis	19.8	LCA	M-	R-	SIOPE high risk	5.31	----	0.36	----	----	7+	G3/4 III-VII	G3/4-LR	Favourable
MB3	14	F	EVD ETV	Vermis	27.7	LCA	M-	R+	SIOPE high risk	5.23	BLM	0.48	----	MYC	----	G3/4 II,III,V	G3-HR	High
MB4	0	F	EVD ETV	Vermis	23.3	LCA	M-	R-	SIOPE high risk	4.96	----	0.48	----	----	7+ 16q-	G3/4 V	G4-HR	Favourable
MB5	13	M	EVD	Cerebellar hemisphere	53.6	LCA	M-	R-	SIOPE high risk	2.49	TP53	0.52	----	MYCN SUFU GLI2	17p- 16q- 2+ 7+ 10q- 16-	SHH-α	SHH-child	Very high
MB6	6	M	EVD ETV	Vermis	33.8	C	M-	R+	SIOPE high risk	2.29	----	0.10	----	----	17p-	SHH-α	SHH-child	Very high
MB7	6	M	EVD ETV	Vermis	65.0	C	M-	R-	PNET 5	3.27	----	0.20	----	----	7+ 17p- 11-	G3/4 IV,VI	G3/4-LR	Favourable
MB8	3 6	F	EVD ETV	Vermis	41.4 4.4	D/N LCA	M-	R-	SIOP/UKCC SG CXT+RT	3.06 0.44	----	1.21	PTCH1 TP53	MYCN	17p- 8- 9q-	SHH-α	SHH-child	Very high
MB9	6	F	EVD ETV	Vermis	28.2	LCA	M-	R-	SIOPE high risk	2.88	----	0.18	CTNNB1	----	17p- 6-	WNT	WNT	Favourable
MB10	1	F	EVD ETV	Vermis	64.4	D/N	M-	R-	HIT-MED bloc SKK	2.51	----	0.08	----	----	----	----	----	----
MB11	6	M	EVD ETV	Vermis	35.3	C	M-	R-	PNET 5	2.48	----	0.40	----	----	17p-	SHH-α	SHH-child	Favourable
MB12	3	M	EVD ETV	Middle cerebellar peduncle	21.0	C	M+	NA	COG 99703	3.09	PTCH1	0.28	PTCH1 BCOR KMT2D	MYCN PTEN GLI2	10q-	SHH-α,β,γ	SHH-child	Very high
MB13	13	F	EVD ETV	Cerebellar hemisphere	56.3	LCA	M-	R-	SIOPE high risk	1.32	TP53	0.85	----	GLI2	14q- 11- 10-	SHH-α	SHH-child	Very high

PFS indicated as years until the date of progression (and if not applicable, death). The frequency of somatic mutations in coding regions per Mb (Mut/Mb) was calculated based on a 50.4 Mb (MB1-MB12) and 17.7 Mb (MB13) WES panel. CNV deletion (log2 value < −1, blue), amplification (log2 value >1, red). For the cytogenetic analysis, the threshold for the percentage of arm-level or whole chromosome gain/loss was determined by the mean and the percentiles of the events from the cohort, greater than 75% percentile (>65.5%) in black and greater than the mean (>44.5%) in grey. The definition of the symbols and acronyms are as follows. Hydrocephalus (H): external ventricular drain (EVD), endoscopic third ventriculocisternostomy (ETV). Pathology: large cell anaplastic (LCA), classic (C), desmoplastic/nodular (D/N). Metastasis (M) status: positive (+), negative (–). Residual (R) disease status: >1.5 cm2 (+), <1.5 cm2 (–). Treatment protocol: chemotherapy (CXT), radiotherapy (RT). High risk (HR), Low-risk (LR).

WES of CSF ctDNA reflects the tumor genomic alterations allowing MB-subgroup and risk stratification

Together with the analysis of primary tumors, whole-exome sequencing analysis of CSF ctDNA was performed in four patients with very high risk (MB5, MB6, MB8 and MB13) to determine whether WES of ctDNA in the CSF could reliably identify and reflect the genomic alterations of the tumor.

After WES analysis, we found that 98.9% (88/89) of the mutations detected in the primary tumor sample with VAF > 5% could also be detected in the matching CSF ctDNA. Importantly, there was a significant correlation between tumor VAFs and the ones obtained from the CSF ctDNA (Figure 23). This observation indicates that the CSF ctDNA recapitulates the intratumor heterogeneity present in the primary tumor providing information about the subclonal genomic architecture of MB tumors. Using our previously developed pipeline for the analysis of copy number alterations from WES (exome2cnv), we identified CNAs from primary tumor and CSF ctDNA. This analysis revealed that the overall CNA landscape in the CSF samples resembled the one found in the matching tumors (MB6, MB8 and MB13). Despite the limited amount of ctDNA from sample MB5-CSF, key MB CNAs detected in the primary tumor could also be detected in the CSF (Figure 24).

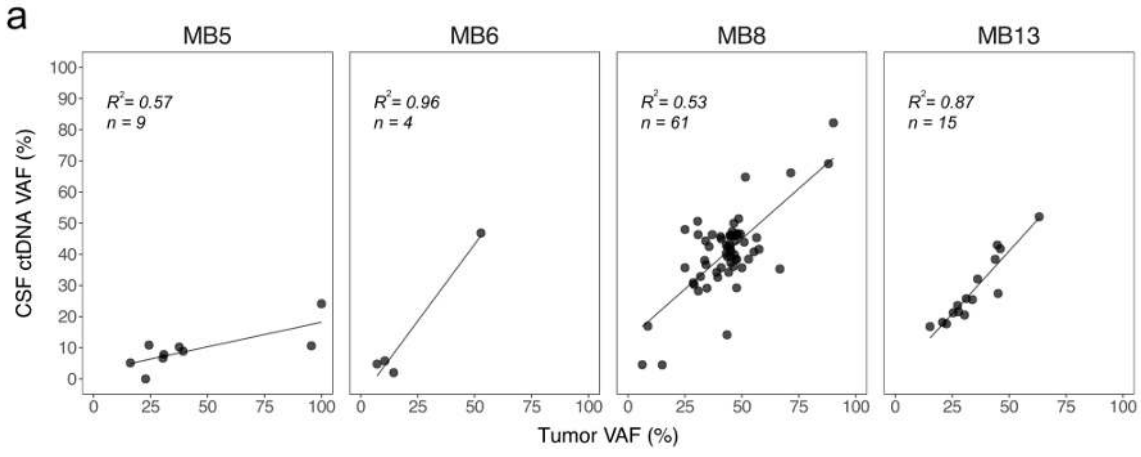

Figure 23. CSF ctDNA allows the detection of the main mutations present in MB tumors. a) Correlation of CSF ctDNA VAF with tumor VAF for each mutation identified in the tumor sample (VAF > 5%). Linear regression and Goodness-of-fit R^2 indicated in each figure. n, represents the number of mutations detected in the primary tumor sample with VAF > 5%.

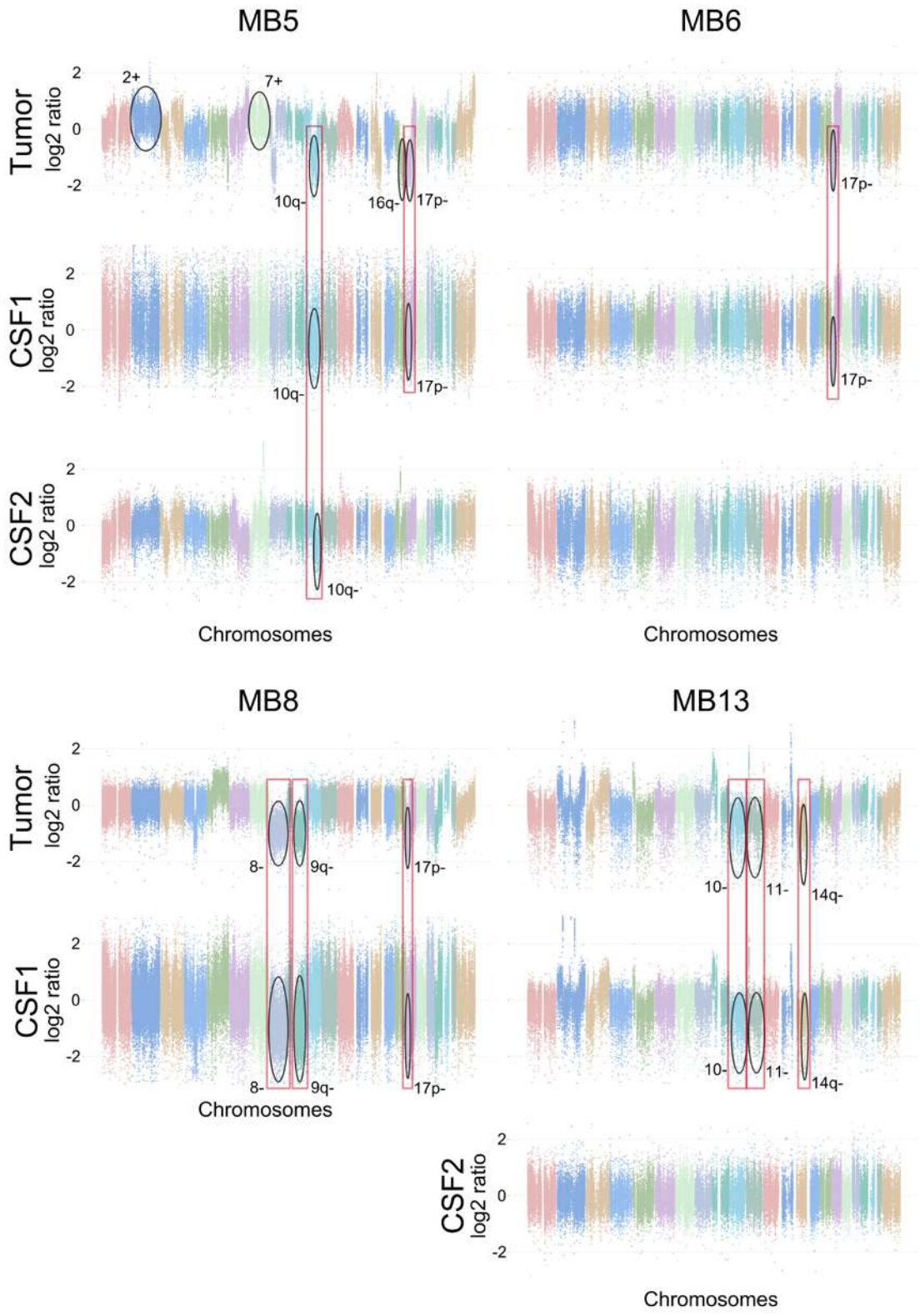

Figure 24. CNA detected in tumor and CSF for four MB patients with longitudinal samples. The marked alterations are those used for molecular characterization and classification into the different MB subtypes. Each dot shows log2 ratios of tumor-normal probe intensities. The chromosomes, from 1 to X, are represented each one in a different color. CSF1 samples were collected at the same time that the tumor, while CSF2 and MB8-CSF samples were extracted for patient monitoring months later.

To determine whether the molecular information obtained from CSF ctDNA could be enough to perform a clinical classification of the primary tumor, we classified patients into a MB-subgroup and risk-stratified. In this regard, many of MB common molecular alterations used for the classification could be detected in CSF, such as *PTCH1* and *TP53* mutations; *MYCN* and *GLI2* amplifications; *SUFU* deletion and 17p loss, which facilitated classification by subgroups. These results matched the classification obtained through the WES analysis of primary tumors, demonstrating that the CSF ctDNA analysis could be used as a diagnostic tool to classify and risk-stratify patients.

CSF ctDNA allows MB tumor monitoring

We next sought to address whether WES from CSF ctDNA could help follow tumor evolution of MB patients and provide information concerning response to treatment or disease progression. For this aim, collected follow-up samples of CSF were analyzed and compared to samples at diagnosis.

In this sense, MB6 and MB13 patients underwent surgical resection, but residual disease was detected by magnetic resonance imaging. After chemo and radiotherapy, MB6 showed a small nodule whereas for MB13 there was no evidence of residual disease during follow-up imaging. The analysis of longitudinal CSF samples for these two patients was able to identify some of the mutations previously detected in tumor and in early CSF samples, but no novel mutations were detected. Even though the frequency at which these mutations appeared in the CSF was very low (VAF < 1.5%), the fact that a small amount of ctDNA was observed in the CSF sample obtained at the end of the treatment indicated that residual disease was still present. Thus, both cases, MB6 and MB13, evidenced that CSF ctDNA could facilitate the identification of minimal residual disease in MB patients.

In contrast to the previous cases outlined, patients MB5 and MB8 progressed and died from their disease following relapse. After surgery, postoperative imaging of patient MB8 identified residual disease and treatment was initiated, but without achieving the expected results. Sequencing analysis of tumor and CSF samples at relapse revealed common alterations including predicted drivers and CNAs. In addition to the tumor-specific mutations that were also detected in the ctDNA, CSF-private mutations were identified, including a new driver mutation in *KMT2A*.

Similarly, patient MB5 underwent surgical resection, although residual disease was also identified on postoperative imaging. Treatment was initiated and thirty months after diagnosis, the patient relapsed. Following tumor resection and chemotherapy, an acute tumor progression was observed, and the patient died 9 months after relapse (Figure 25a). Despite the limited amount of CSF cfDNA, mutations and most relevant genomic alterations identified in the tumor were also detected in the CSF ctDNA, including *MYCN* and *GLI2* amplification and partial loss of 17p. Surprisingly, the analysis of CSF ctDNA at relapse revealed a genomic transformation, with a distinct molecular profile not shared with the primary nor with the first relapsed tumor (Figure 24 and Figure **25**b).

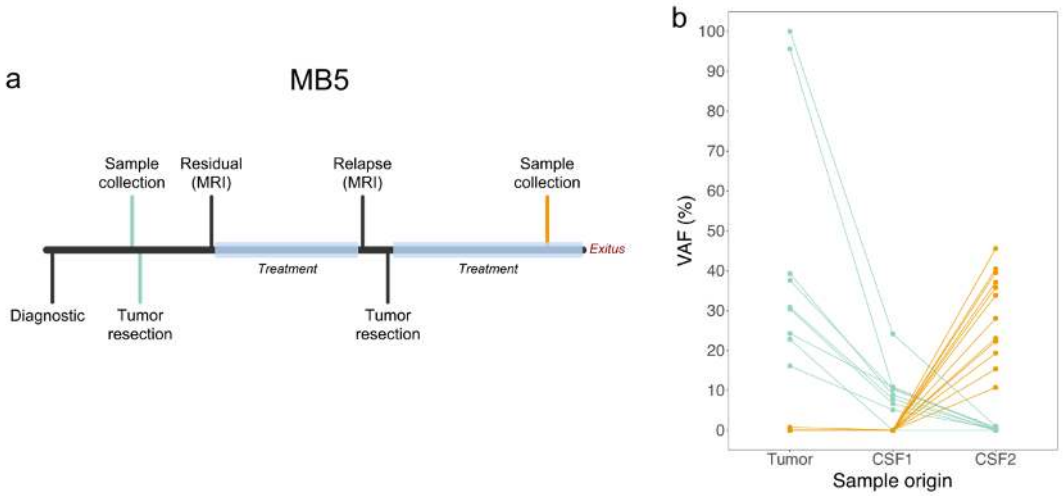

Figure 25. CSF ctDNA reveals a new primary tumor in MB5 longitudinal sample. a) Longitudinal monitoring of MB5 patient that relapsed and died from the disease. b) VAFs (%) identified from the WES of tumor DNA, CSF1 and CSF2 cfDNA samples. Most of the mutations detected in CSF2 had not been previously detected in the tumor or in CSF1. The color refers to whether the samples (a) or the mutations (b) belonged to the first (green) or to the second independent tumor (orange). MRI stands for magnetic resonance imaging.

The almost complete disappearance of all somatic mutations detected in the primary tumor and the appearance of a new set of somatic mutations at relapse strongly supported the idea of the presence of two independent tumors, one at diagnosis and a different one at relapse. Due to the low probability of developing MB, we sought to determine whether this patient might have a predisposing condition for MB development. Thus, analysis of germline variants by WES revealed that the patient carried a *TP53* p.R282W mutation, indicating that he suffered from Li-Fraumeni

syndrome. This mutation had been previously described in Li-Fraumeni[152], and as expected, both tumor samples from this patient had loss of heterozygosity (LOH) of the normal allele. Importantly, our data highlights that the analysis of CSF ctDNA can be used to monitor genomic tumor evolution. In this case, CSF ctDNA was able to identify two completely different tumors at diagnosis and relapse. This result prompted us to analyze germline variants from the 13 MB patients analyzed (Table 11), resulting in the identification of another Li-Fraumeni case (MB13 *TP53* p.S127F). Although in ClinVar (VCV000182928) the clinical significance for this mutation is "Conflicting interpretations of pathogenicity", it has been identified in another Li-Fraumeni patient and the last three submissions have labeled it as "Pathogenic". Moreover, this mutation appears 90 times in COSMIC as somatic mutation, and we confirmed LOH of the normal allele in the tumor, strongly suggesting that it is a *bona fide* loss of function mutation.

Table 11. Germline mutations in cancer predisposing syndrome genes detected in MB patients

Case	Chr	Pos*	Ref	Alt	VAF	Consequence**	Gene	gnomAD***	ClinVar
MB12	9	98268688	C	T	0.69	Splice donor c.196+1C>T	*PTCH1*	-	VCV000428849 *Pathogenic*
MB7	13	32911754	C	T	0.46	Missense p.P1088S	*BRCA2*	5.00E-05	VCV000037829 *Conflicting interpretations of pathogenicity*
MB3	13	32929431	G	A	0.42	Splice region c.7435+6G>A	*BRCA2*	2.40E-04	VCV000038096 *Conflicting interpretations of pathogenicity*
MB8	13	32944624	C	T	0.48	Missense p.S2806L	*BRCA2*	3.40E-04	VCV000142874 *Conflicting interpretations of pathogenicity*
MB7	15	91312388	C	G	0.46	Missense p.S778C	*BLM*	7.00E-05	VCV000127484 *Conflicting interpretations of pathogenicity*
MB3	15	91346807	C	T	0.59	Stop gained p.R1139*	*BLM*	-	VCV000156484 *Pathogenic*
MB9	15	91354640	T	G	0.51	Splice region c.4076+4T>G	*BLM*	3.45E-03	VCV000136520 *Conflicting interpretations of pathogenicity*
MB5	17	7577094	G	A	0.59	Missense p.R282W	*TP53*	-	VCV000012364 *Pathogenic*
MB13	17	7578550	G	A	0.53	Missense p.S127F	*TP53*	-	VCV000182928 *Conflicting interpretations of pathogenicity*
MB13	17	41243700	T	C	0.42	Missense p.H1283R	*BRCA1*	1.00E-05	VCV000055028 *Conflicting interpretations of pathogenicity*

**The mutation manually added after being eliminated for appearing with a frequency in the population higher than 0.001 is underline.*

***Consequences were predicted with the program "Variant Effect Predictor".*

****Frequency of the mutation detected for general population in gnomAD database.*

Furthermore, we found three cases with variants in the RecQ Bloom-syndrome associated helicase *BLM*. Case MB3 had a p.R1139* premature stop codon (15:g.91346807C>T) that appears in ClinVar (VCV000156484) as "Pathogenic" in patients with Bloom syndrome. Case MB7 had a p.S778C (15:g.91312388C>G). In ClinVar its accession number is VCV000127484 and is labeled as "Conflicting interpretations of pathogenicity". It has been detected in a patient with hereditary cancer-predisposing syndrome, and 3 other individuals for which condition was not indicated. There are no mutations in COSMIC, so the clinical significance is unknown. There was an extra case (MB9) with another mutation in *BLM* c.4076+4T>G (15:g.91354640T>G) affecting the 4^{th} base of intron 21-22, and with the label "Conflicting interpretations of pathogenicity" in ClinVar (VCV000136520). It has been detected in a patient with Bloom syndrome and two patients with hereditary cancer predisposing syndrome, but also in 7 patients for which condition was not specified. This variant also appears in gnomAD with a VAF of 0.00345 (reason why it did not pass our filters). However, we think that having two more cases with mutations in this gene and knowing that it is also implied in a cancer predisposition syndrome (Bloom syndrome) it may be relevant in the susceptibility of this patient. As this mutation is near a splice donor site, it might have an impact on splicing. Thus, we performed an analysis with the Human Splicing Finder tool, and we obtained that the mutation creates a new donor site as it generates the consensus donor sequence (AGGT), suggesting that it might have an impact on splicing.

In addition to germline variants in *TP53* and *BLM*, we identified three cases with germline variants in *BRCA2*, a well-known gene for hereditary predisposition to ovarian cancer. Case MB7 had a p.P1088S variant (13:g.32911754C>T) which is present in ClinVar (VCV000037829), labeled as "Conflicting interpretations of pathogenicity". It has been detected in 7 patients with familial breast-ovarian cancer or hereditary cancer predisposing syndrome, and in 5 individuals with no clinical information. This same variant is also present in gnomAD at a frequency of $4x10^{-5}$, suggesting that is not a common variant, but its status as pathogenic is not clear. MB3 presented a mutation c.7435+6G>A affecting the 6^{th} base of intron 14 (13:g.32929431G>A). This variant is

present in ClinVar (VCV000038096) as “Conflicting interpretations of pathogenicity”, and it has been detected in 9 patients with familial breast-ovarian cancer or hereditary cancer predisposing syndrome, and in 7 individuals with no clinical information. In gnomAD it is present at $2x10^{-4}$, indicating that it might constitute a rare variant. Finally, MB8 had a p.S2806L (13:g.32944624C>T). Again, although this variant appears in ClinVar (VCV000142874), it is classified as “Conflicting interpretations of pathogenicity”. It has been detected in 4 patients with familial breast-ovarian cancer or hereditary cancer predisposing syndrome, and in 3 individuals with no clinical information. This variant is present in gnomAD at a frequency of $2.5x10^{-4}$, suggesting that it might be a rare variant with clinical implications. Although the three *BRCA2* variants have been labeled as “Conflicting interpretations of pathogenicity”, the finding of 3 variants in just 12 patients is remarkable and is much higher that it would have been expected by chance, suggesting that they might be implicated in the etiology of the disease.

In addition, two variants were found in *BRCA1*, MB13 p.H1283R (17:g.41243700T>C), and in *PTCH1*, MB12 c.196+1C>T affecting the first base of intron 2 (9:g.98268688C>T), which disrupts the splicing donor site. Both are present in ClinVar (VCV000055028 and VCV000428849, respectively) and classified as “Conflicting interpretations of pathogenicity” in the case of *BRCA1* or “Pathogenic” for *PTCH1*. Although *BRCA1* variant was found in two patients with hereditary predisposing syndrome and three patients with breast-ovarian cancer, it is not clear whether it might be important in the etiology of this tumor. On the other hand, this same variant in *PCTH1* was also detected in a patient with hereditary cancer predisposing syndrome, and in another individual for which clinical condition was not reported. In summary, although many of these germline mutations might be considered rare variants, they all appear to have clinical implications, suggesting that they may be associated with an increased predisposition to develop MB.

Mutations in the U1 spliceosomal RNA

The studies carried out by the major international cancer consortia such as PCAWG and TCGA have provided a detailed view of the molecular alterations involved in cancer development. Hundreds of driver genes have been identified, most of them located in coding regions[36]. However, despite the efforts made, there is still a bias in the detection of mutations in certain regions due to the characteristics of NGS. Thus, repetitive regions, where the alignment quality is very low due to the inability to align short reads to any individual sequence, are excluded from these analyses. Therefore, so far it has not been possible to detect mutations in these regions.

U1 snRNA is recurrently mutated in multiple cancers

The small nuclear RNA U1 is a non-coding component of the spliceosome, responsible for the recognition of the 5' splice site by base pairing (Figure 6). There are 7 identical copies of this gene in our genome and more than 130 annotated pseudogenes with minor changes, in addition the promoter regions of all of them are also very similar[153,154]. The repetitiveness of these sequences makes the classical variant callers exclude them from their analysis, suggesting that there are no mutations affecting this gene. However, targeted and exhaustive analysis of this gene in 2,538 whole-genome sequenced donors across 37 tumor types from the PCAWG has identified three mutation hotspots, at positions 3, 9 and 28, in more than 5% of patients in at least one tumor type.

The most common frequent mutation affected base 3 of *U1*. This base forms part of the sequence that recognizes the 5'SS, and was mutated in more than 50 cases across five tumor types. In particular, the g.3A>G mutation was found in 26 out of 135 (19.3%) medulloblastoma cases and only 1 case of pancreatic ductal adenocarcinoma, concluding that this is extremely specific to medulloblastoma. In addition, expanding the analysis to the entire ICGC dataset and 114 samples from MAGIC and using an allele-specific PCR (rhAMP) for the detection of the g.3A>G mutation, we found that this mutation was largely restricted to cases of SHH medulloblastoma subtype in adulthood (SHHδ, present in 97% of cases) and adolescence (SHHα, present in 25% of cases), and

absent from those in infancy (Figure 26). On the other hand, the g.3A>C mutation was found in 8/78 cases of chronic lymphocytic leukemia (10.3%), 16/189 cases of hepatocellular carcinoma (8.5%) and 2/107 (1.9%) cases of B cell non-Hodgkin lymphoma.

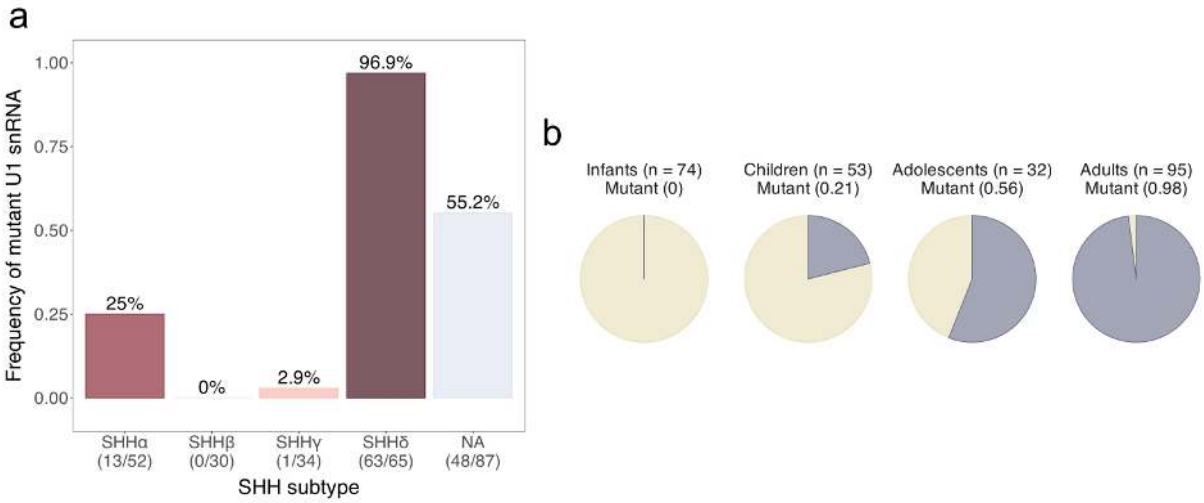

Figure 26. Clinical and cytogenetic features of SHH medulloblastomas with mutant *U1* snRNA. a) Frequency of *U1* snRNA mutations across subtypes of SHH medulloblastoma. NA, not available (samples for which the subtype is unknown). b) Frequency of *U1* snRNA mutation by age group (n = 74 for infants, n = 53 for children, n = 32 for adolescents, n = 95 for adults).

Looking at the interaction of these mutations with those previously detected in these pathologies, we found that $U1^{g.3A>G}$ co-occurred significantly with mutations in the *TERT* promoter (Fisher's q = $3.18x10^{-8}$) and *DDX3X* (Fisher's q = $2.02x10^{-8}$) for SHH medulloblastoma cases. In the case of $U1^{g.3A>C}$, in CLL it co-occurred with *NFKBIE* (Fisher's q = 0.0077), whereas in HCC it co-occurred with *APOB* (Mantel-Haenszel test q = 0.018) and *TERT* (Fisher's q = 0.016), although the latter only in one liver project (LIHC-US). Although *SF3B1* is one of the most recurrently mutated genes in CLL[29,155], also causing global splicing changes when mutated[156], no mutations were found in both genes in the same sample. However, just by chance we would have expected to find at least two cases with both genes altered. These data suggest that the mutations may be mutually exclusive, although a larger data set is needed to have sufficient statistical power to confirm this fact.

Expression and splicing deregulation

The U1 snRNA is responsible for the recognition of the 5'SS sequence in the transcribed RNA. U1 anneals from nucleotides 3 to 10 with the 5'SS of the mRNA, marking this region for the subsequent splicing steps (Figure 27). Although recognition occurs by Watson-Crick base complementarity, this pairing does not have to be perfect, leading to the presence of strong or weak splicing donor sites depending on how many and in which positions mismatches occur. Thus, it has been possible to generate a consensus 5'SS with the probability of each nucleotide to appear in each position, with some bases, such as the first two of the intron (GU nucleotides at positions 1 and 2), being more important than the rest. Therefore, because the mutation found at position 3 is part of the 5'SS recognition sequence, we hypothesized that these mutations could shift the U preference towards a C or G at the sixth position of the 5'SS (hereafter C/G6-5'SS).

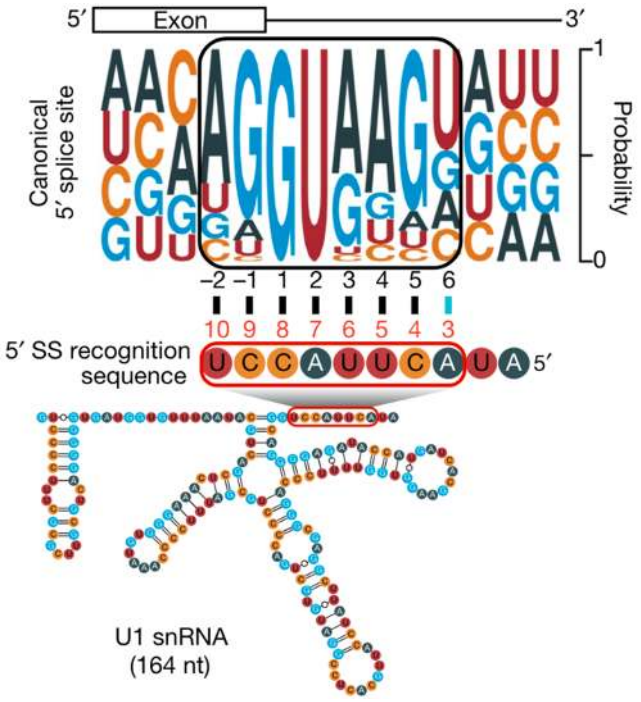

Figure 27. The RNA–RNA interaction between U1 and the 5′ splice site. Bases 3 to 10 of U1 (red box and numbering) can base-pair with the 5′ splice site (black box and numbering). The base-pairing affected by the mutation located in base 3 is in blue.

To check if there were global changes in the splicing pattern, an intron-centric differential splicing analysis with LeafCutter was performed using RNA-seq data from MB (13 SHHα-U1$^{g.3A>G}$ vs 39-SHHα U1^{wt} and 30-SHHδ U1$^{g.3A>G}$ vs 90 U1^{wt} from other SHH subtypes), CLL (11 U1$^{g.3A>C}$ vs 254 U1^{wt}) and HCC (20 U1$^{g.3A>C}$ vs 367 U1^{wt}). Thus, 3,193 and 533 differentially spliced introns were identified in 1,519 and 303 genes (LeafCutter $q < 0.1$ and absolute $\log_2$(effective size) > 1) in CLL and HCC, respectively. In consequence, we observed 2-3 times more novel splicing events in patients harboring the g.3A>G/C mutation than the wild-types, especially for splicing with the cryptic C/G6-5'SS in the three types of tumors. Furthermore, the direction of change for each intron was determined using the magnitude of the splicing change (ΔPSI). When comparing the base composition of the 5' splice site among introns with increased excision (ΔPSI > 0) and decreased excision (ΔPSI < 0) in samples with U1 mutation, we observed an enrichment of a dominant C or G base at the sixth position, as opposed to the T base observed in tumors with wild-type U1 snRNA (Figure 28).

To detect alternative splicing in individual genes, we focused on cryptic 5' splicing events with a C or G base at the sixth intronic position. For patients with SHHα and SHHδ subtypes of MB who harbor the U1$^{g.3A>G}$ mutation, cryptic and very specific splicing events with high effect sizes were detected in protein patched homolog 1 (*PTCH1*), zinc finger protein GLI2 (*GLI2*) and paired box 5 (*PAX5*). Splicing mediated by the U1$^{g.3A>G}$ mutant in *PTCH1* results in the inclusion of a cassette exon between exon 2 and exon 3, which causes a frameshift, and therefore predicted translation start from the ATG in exon 3 (Figure 29a). It has previously been reported that loss of expression of the 1,447-amino-acid isoform of PTCH1 promotes the derepression of Hedgehog signaling[157], an upregulated pathway in this MB subgroup. Similarly, the U1$^{g.3A>G}$ cassette exon in *GLI2* is spliced between exon 4 and exon 5, which results in a putative GLI2 protein that lacks the repressor domain (Figure 29b). Physiological GLI2 protein has a repressor domain at its amino terminus, and constructs that lack the amino terminus are much more potent at activating Hedgehog signaling than the full-length protein[158].

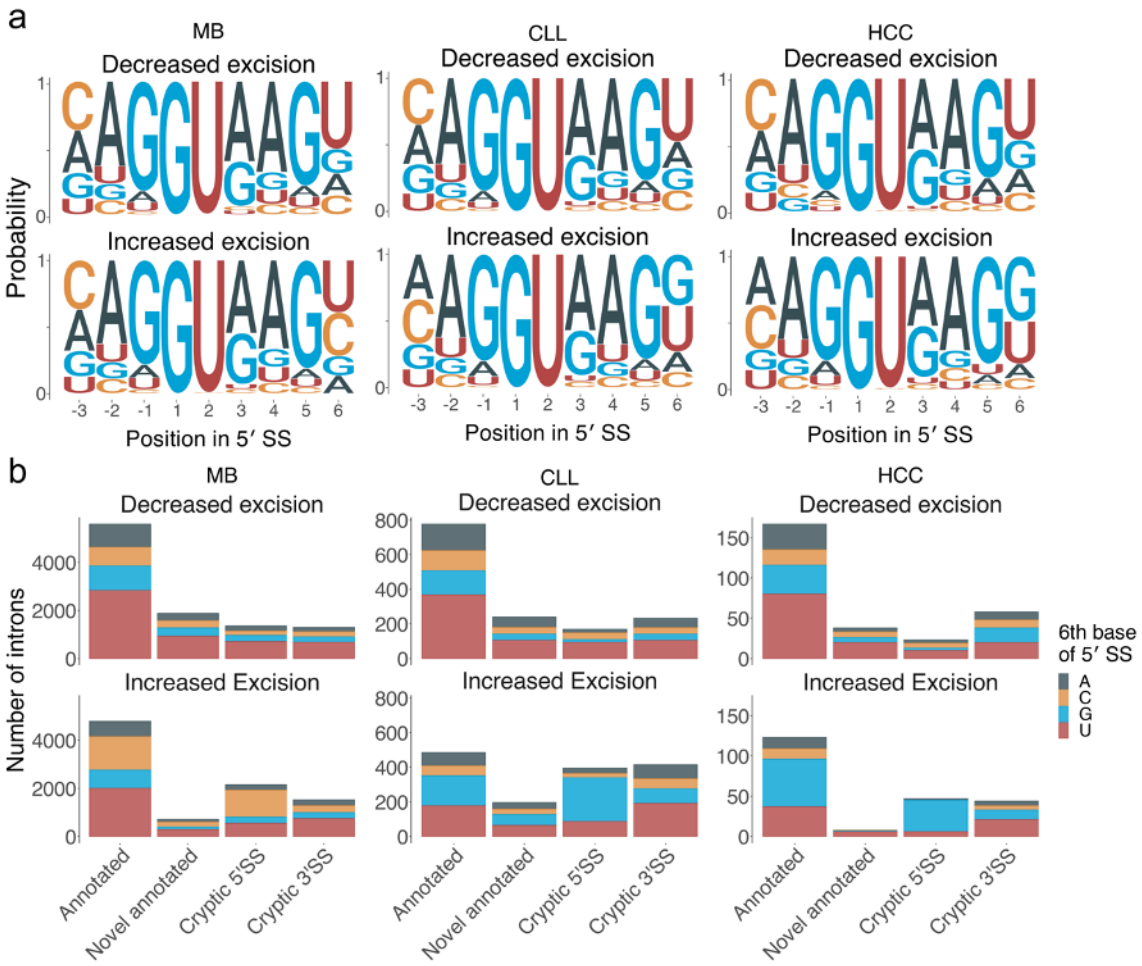

Figure 28. Global gene splicing changes associated with the g.3A>G and g.3A>C mutations in MB, and CLL and HCC, respectively. a) 5' splice site (5'SS) for introns with increased (n = 9,182 in MB; n = 1,657 in CLL; n = 239 in HCC) or decreased excision (n = 10,126 in MB; n = 1,536 in CLL; n = 294 in HCC) in cases with U1 mutation (n = 11 patients). Increased excision and decreased excision represent intron clusters that have significantly mis-spliced introns with ΔPSI > 0 and ΔPSI < 0, respectively. b) Category of mis-splicing events in MB, CLL and HCC, respectively. The number of introns is coloured by the sixth base of 5′ splice site. The 6th base of the sequence corresponds to the one interacting with base 3 of *U1*. The definitions of each category is: "annotated" if the junction matches any annotated introns; "novel annotated" if both splice sites are annotated but not paired; "cryptic 5′ splice site" if only the 3′ splice site is annotated; "cryptic 3′ splice site" if only the 5′ splice site is annotated.

In the case of the *PAX5* tumor suppressor gene, the isoform present in SHH MBs with U1^{wt} translates the complete DNA-binding domain. However, the cryptic exon in SHH medulloblastomas with mutant U1 snRNA causes a stop codon before the DNA-binding domain, resulting in loss of function (Figure 29c). Alternative splicing of the cell-cycle gene *CCND2*, a known downstream target of SHH signaling that is recurrently amplified in SHH MB, is detected in U1$^{g.3A>G}$ mutants of SHHδ but not in SHHα. The

CCND2 alternative isoform is prematurely terminated, which results in N-terminal sequences in which the PEST domain is predicted to be deleted (Figure 29d). Deletion of the PEST domain causes resistance to protein degradation and impaired export to the cytoplasm, which results in CCND2 accumulating in the nucleus to promote cell-cycle progression[159].

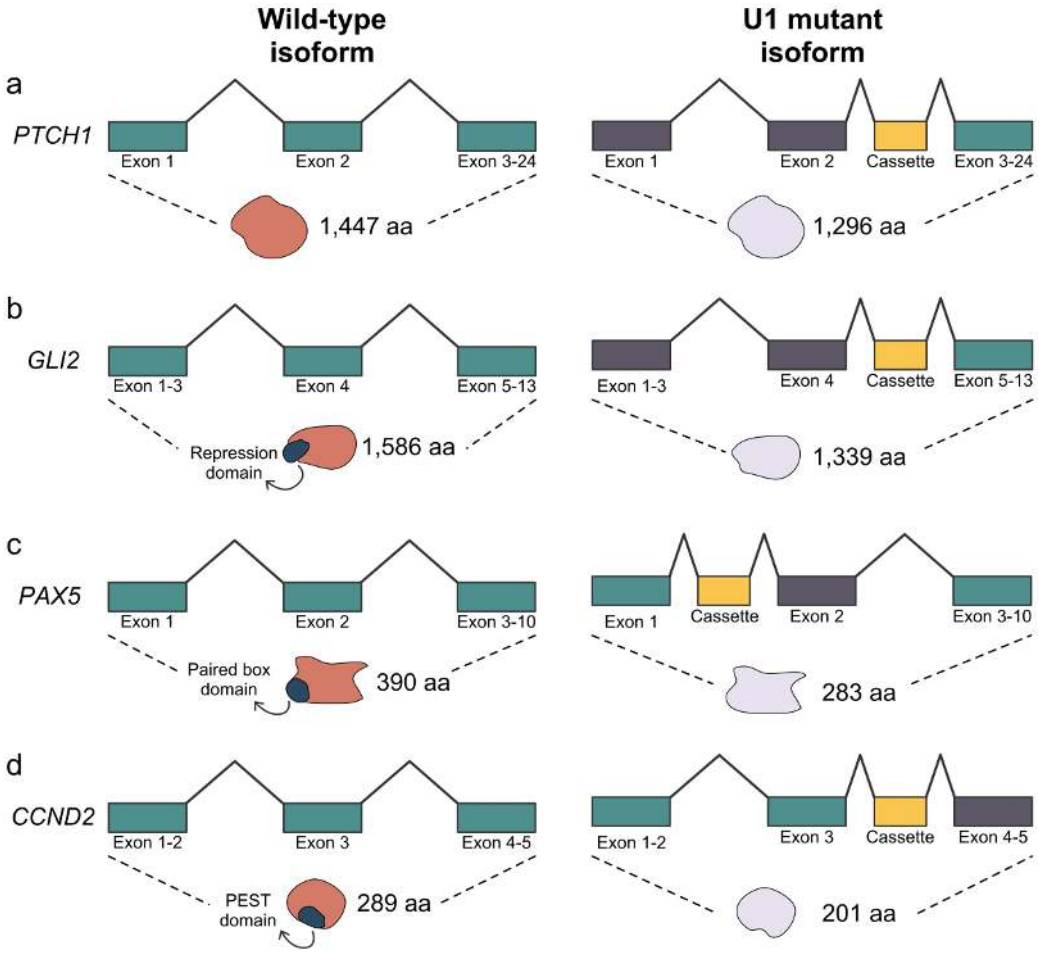

Figure 29. Illustration of canonical and cryptic alternative isoforms of genes affected by U1$^{g.3A>G}$ in MB. a) The new cassette in *PTCH1* causes transcription to begin in exon 3, losing exons 1 and 2. b) The new cassette between exons 4 and 5 of *GLI2* generates a frameshift that causes the protein to start transcribing in exon 5. This results in the loss of a repression domain in the mutant protein. c) The addition of a new exon between exons 1 and 2 of *PAX5* delays the initiation of transcription to exon 2, leading to the loss of a paired-box domain. d) In *CCND2* alternative splicing causes the loss of a PEST domain located in the C-terminal region of the protein. In green are represented the canonical exons, in yellow the cryptic exon generated by mutated U1 and in dark grey canonical exons affected by novel exons. Resulting proteins (and size) are displayed for each isoform. aa, amino acid.

For patients carrying the U1$^{g.3A>C}$ mutation in both CLL and HCC, 84 and 16 genes were mis-spliced according to Cancer Gene Consensus (CGC v.84), respectively. The most significant mis-spliced cancer gene associated with a G6 5′ splice site was musashi RNA binding protein 2 (*MSI2*) in CLL, where an isoform containing a cryptic exon that features a premature stop codon is exclusively expressed (Figure 30a). A similar pattern was observed for the gene DNA polymerase delta 1, catalytic subunit (*POLD1*). As the cryptic exon affected the polymerase, but not the exonuclease domain of POLD1 (Figure 30b), the g.3A>C mutation was not associated with a higher mutation burden. We also found mis-splicing in other genes not present in the CGC but related to CLL biology, such as the hyaluronic acid receptor gene CD44 molecule (*CD44*), which was the most significantly differentially spliced gene. Alternative splicing of *CD44* is tissue specific and has previously been associated with processes such as lymphocyte homing and tumorigenesis; the gene is also thought to regulate anti-apoptosis signaling in CLL[160,161]. Patients with wild-type CLL expressed predominantly the standard isoform (CD44s, which does not contain exon v2–v10), whereas cases of CLL with U1 mutations overexpressed multiple variant isoforms (CD44v)—presumably because the presence of several G6 5′ splice sites increased the excision rate of introns associated with variant exons (Figure 30c). Another similar example is ATP-binding cassette sub-family D member 3 (*ABCD3*), a fatty acid transporter for peroxisomes. Two cryptic exons were expressed exclusively in cases of CLL with U1 mutations (Figure 30d).Together, the alternative splicing results observed in these experiments support a model in which cryptic alternative splicing mediated by U1$^{g.3A>G}$ and U1$^{g.3A>C}$ might function as a driver in subsets of SHH MB, CLL and HCC, respectively.

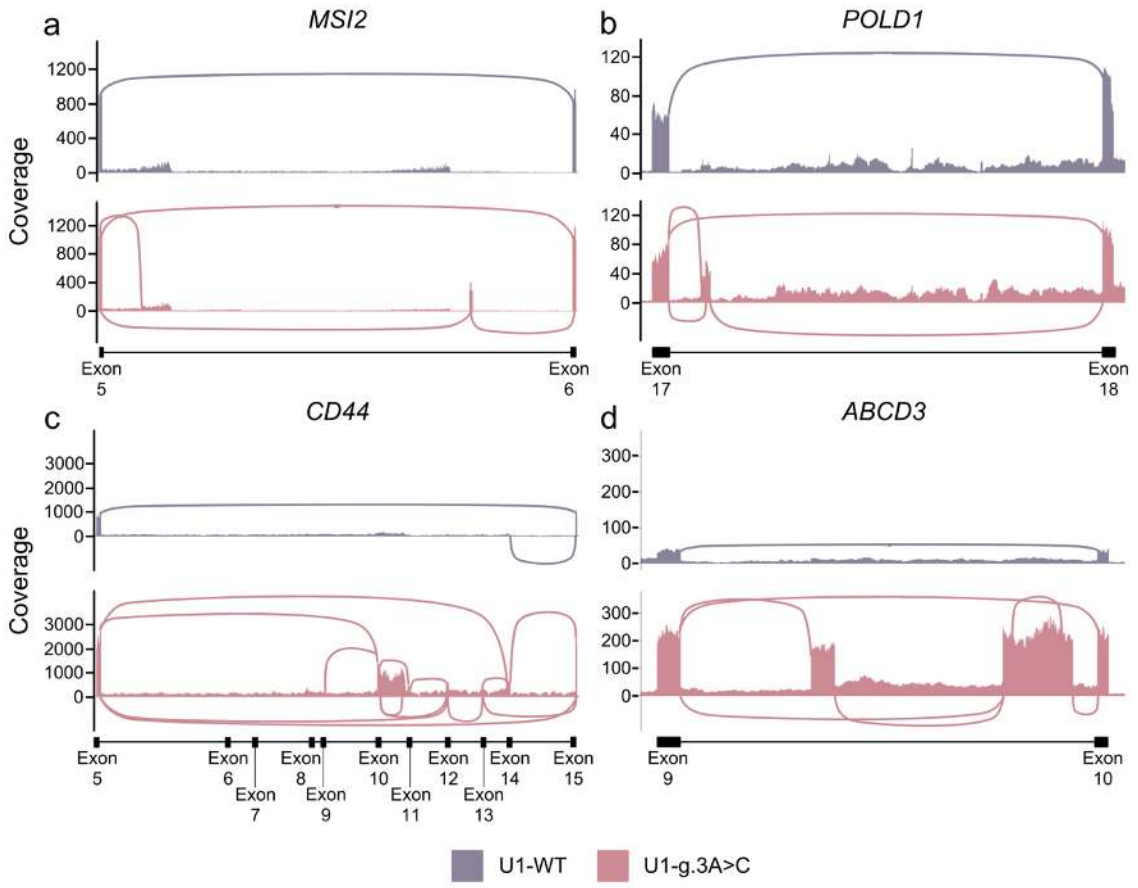

Figure 30. Sashimi plots of genes with alternative splicing due to U1$^{g.3A>C}$. a-b) The g.3A>C mutation in *U1* allows the detection of a cryptic exon in *MSI2* and *POLD1* genes. c) In the case of *CD44*, the mutation does not generate new exons but causes the expression of different isoforms, previously described, but not expressed in B cells. d) In contrast, in *ABCD3*, *U1$^{g.3A>C}$* results in the detection of two cryptic exons in intron 9 of the gene. The lines joining the exons represent the junctions that give rise to the different isoforms detected for each gene. Plots have been generated from RNA-seq data of patients with or whithout mutations in *U1* with the program ggsashimi.

Because splicing and expression frequently correlate[156,162], we also conducted differential expression analysis. In this regard, we found a wide range of deregulated genes in which we performed pathway analysis. An increase in non-sense mediated decay, which is consistent with the destruction of aberrantly spliced transcripts, was detected for SHH MB primary tumors (Figure 31a). However, in CLL this pathway was downregulated together with apoptosis, B cell receptor signaling and cytoplasmic ribosomes (Figure 31b). The downregulation of ribosomal genes may explain the reduced rates of nonsense-mediated decay observed in CLL. On the other hand, genes related to mRNA transcription, RNA splicing, protein ubiquitination and telomere maintenance were upregulated in CLL.

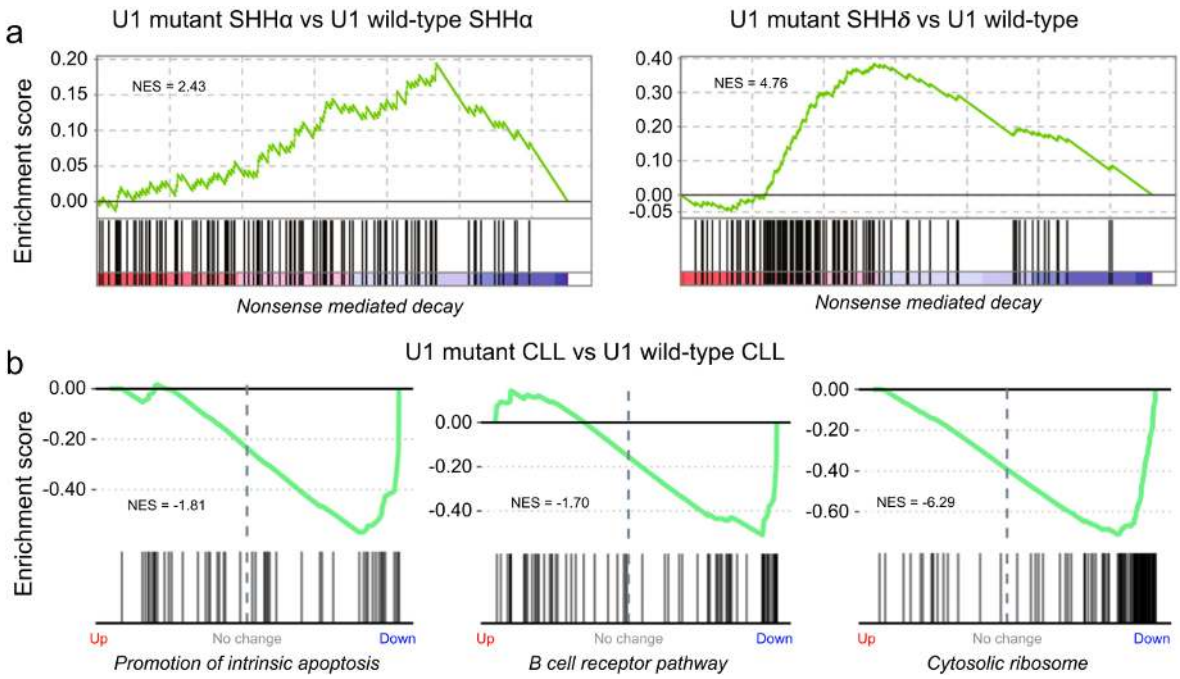

Figure 31. Pathways altered in MB and CLL with *U1* mutations. Enrichment plots by gene set enrichment analysis (GSEA) of a) "nonsense mediated decay" process between SHHα with mutant U1 (n = 13) and SHHα with wild-type U1 (n = 39), and SHHδ with mutant U1 (n = 30) and other subtypes of SHH medulloblastoma with wild-type U1 (n = 90). b) "promotion of intrinsic apoptosis", "B cell receptor" and "cytosolic ribosome" pathways between CLL with mutant (n = 11) and wild-type (n = 254) U1. Genes in the x axis are sorted from the most significantly upregulated gene to the most significantly downregulated gene. NES, normalized enrichment score.

Validation of $U1^{g.3A>G}$ and $U1^{g.3A>C}$ mutations

To determine whether there was a causal relationship between the U1 mutations and the splicing alterations observed in primary tumor samples, we performed an *in vitro* validation study. Thus, we constructed vectors carrying the U1 wild-type sequence as well as the two mutations previously described for position 3. To be as closely as possible physiological conditions, and since the *U1* promoter region is well characterized[163,164], we decided to clone not only the U1 gene but also its flanking regions. In this way we extracted 400 bp upstream of the start of the gene containing the promoter with the two required regions for U1 expression, the distal sequence enhancer and the proximal sequence enhancer. In addition, we also cloned 40 bp downstream of the gene, as this region is important for its post-transcriptional

processing and transport between the nucleus and the cytoplasm (Figure 32). In order to express the g.3A>C mutation in CLL cell lines (JVM3, HG3 and MEC1), and due to their low transfection efficiency, the plasmid used for these constructions was pLKO.1-puro, a lentiviral vector previously used in our laboratory. This plasmid is designed to insert the gene of interest under the U6 promoter. However, we replaced most of the U6 promoter as we used the U1 endogenous promoter (Figure 32).

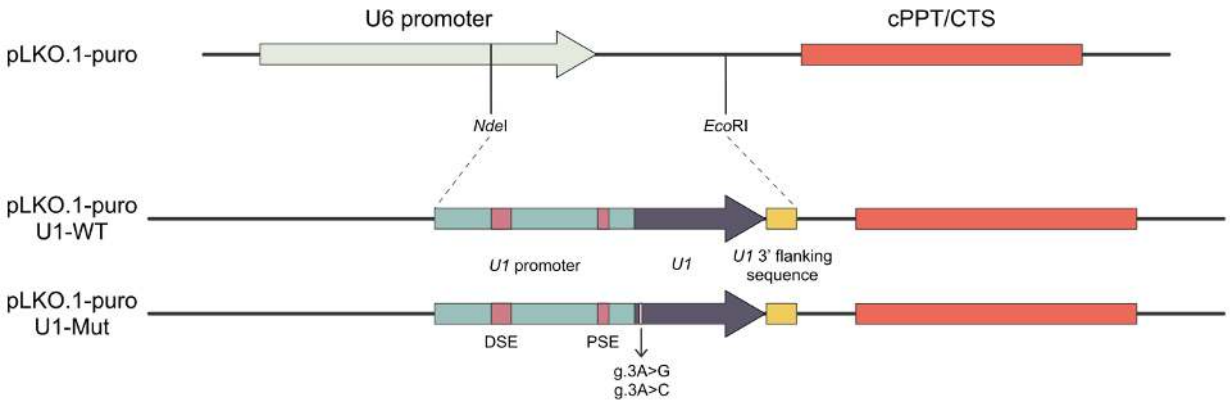

Figure 32. Representation of how the pLKO.1-U1^{wt} and pLKO.1-U1$^{g.3A>G/C}$ plasmids, used during cell lines experiments, were generated. The 3' region of the *U6* promoter was removed by inserting the promoter, gene and 3' flanking sequence of *U1*. The proximal (PSE) and distal (DSE) sequence enhancers necessary for *U1* expression are represented within the *U1* promoter.

To verify that the plasmid expressed the U1 mutations, pLKO.1-U1^{wt} and pLKO.1-U1$^{g.3A>G}$ vectors were transfected into human embryonic kidney 293T and pLKO.1-U1^{wt} and pLKO.1-U1$^{g.3A>C}$ were used to infect the CLL cell lines JVM3, HG3 and MEC1. Then, RNA was extracted to perform 5'RACE followed by Sanger sequencing, confirming that mutant U1 was expressed in the selected cell lines (Figure 33).

Previous analysis of RNA-seq data from MB, CLL and HCC patients resulted in the identification of numerous genes affected by mutations in *U1*. This allowed us to select some of them to validate the changes by RT-PCR. In this way, specific oligonucleotides were designed to amplify the affected region in the wild-type and mutated isoforms of the following genes: *PTCH1* and *MSI2* for *U1*$^{g.3A>G}$ and *ABCD3*, *MSI2* and *POLD1* for *U1*$^{g.3A>C}$. Thus, it was verified that cells expressing mutant *U1* also had altered splicing (Figure 34).

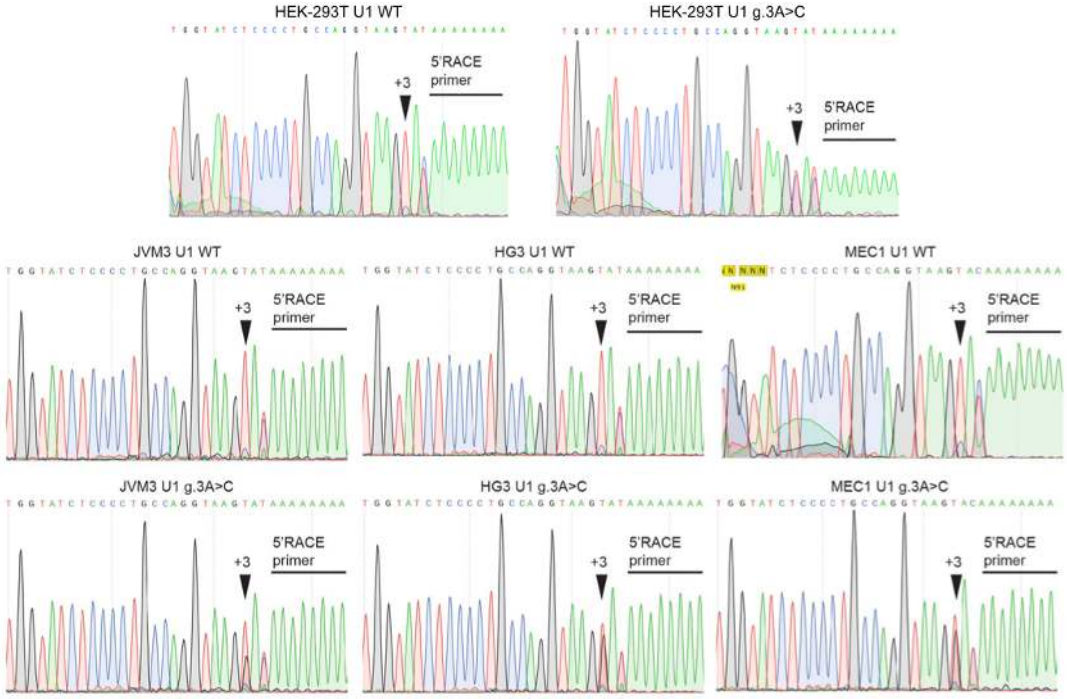

Figure 33. 5' RACE confirming the expression of the U1$^{g.3A>G}$ in HEK-293T and U1$^{g.3A>C}$ in three CLL cell lines. HEK-293T were transfected whereas CLL cell lines (JVM3, HG3 and MEC1) were infected with lentiviral particles that contain *U1* locus with or without the g.3A>G or g.3A>C mutations respectively. The electropherograms correspond to the sequence of the PCR product (reverse strand). The arrowheads indicate the location of the 3rd base of *U1*. The location of the 5' RACE primer is also indicated. The experiment was conducted once.

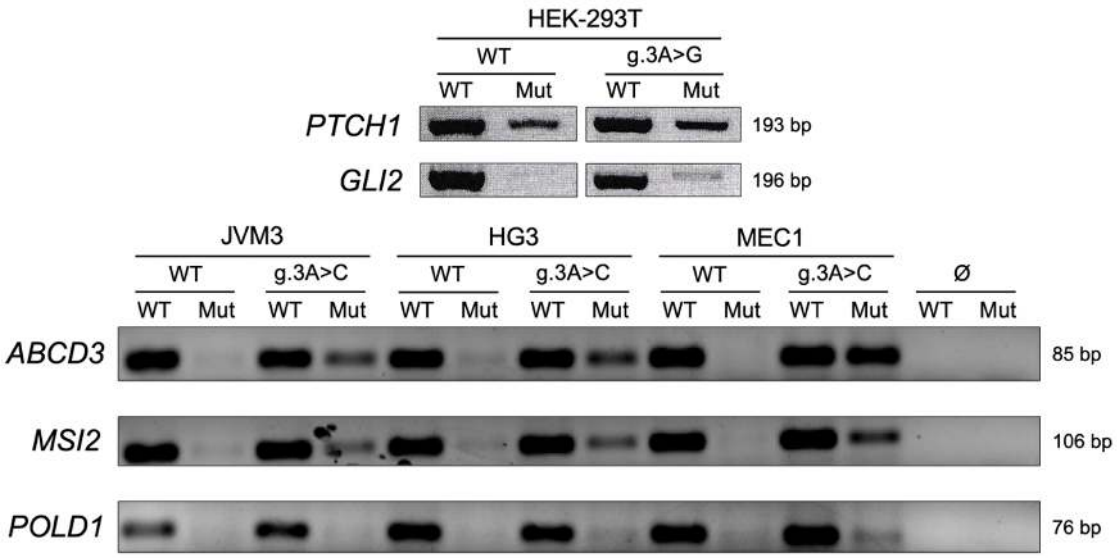

Figure 34. RT-PCR results for experimental validation of aberrant splicing after expression of mutant *U1*. To validate the g.3A>G mutation, *PTCH1* and *GLI2* genes were used, while for g.3A>C were *ABCD3*, *MSI2* and *POLD1*. The RNA used for this experiment was extracted from HEK-293T transfected with pLKO.1-U1$^{g.3A>G}$ and from CLL cell lines (JVM3, HG3 and MEC1) infected with pLKO.1-U1$^{g.3A>C}$. In the case of *PTCH1* and *POLD1* the mutated isoform appears in the wild-type cells at low levels and an increase in the intensity of the band is observed.

After confirming that cells transfected or infected with the pLKO.1-U1$^{g.3A>C}$ and pLKO.1-U1$^{g.3A>C}$ constructions, respectively, expressed the mutation in *U1* and that this was leading to changes in splicing, we decided to perform an RNA-seq experiment. Thus, the same transcriptome analysis for RNA-seq data was performed for these cell lines as was done for patients. In total, 826 introns in 318 genes in HEK-293T and 7,238 introns in 2,365 genes in CLL cell lines were differentially spliced when comparing U1^{wt} vs U1$^{g.3A>G}$ or U1$^{g.3A>C}$. Moreover, intron-centric analysis clearly demonstrates an enrichment of a G at the sixth intronic position when using the pLKO.1-U1$^{g.3A>C}$ vector (Figure 35). In the same way, a considerable increase in the incidence of cryptic 5′ splicing events was detected, with the finding of more introns with increased than decreased excision (Figure 35). In addition, CLL cell lines with *U1* mutations shared 39.1% of the G6 5′ splice site introns with increased excision and many differentially expressed genes of those detected in primary tumors. These results reproduce what was observed in primary tumors and validate a causal link between the g.3A>G and g.3A>C mutations and global splicing changes.

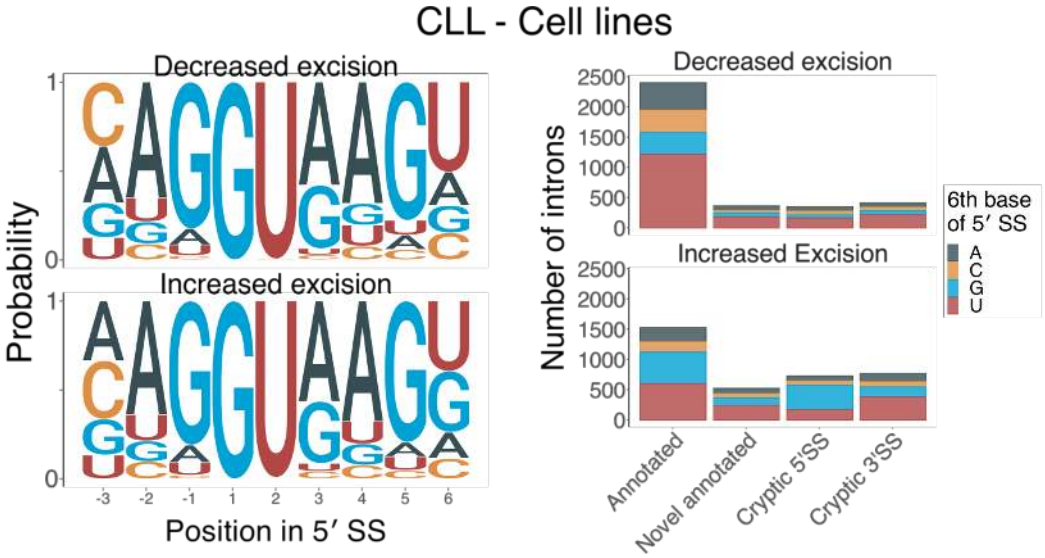

Figure 35. Exogenous expression of the g.3A>C mutation in cell lines induced global splicing changes. On the left, 5' splice site (5'SS) sequence usage for introns with increased (n = 3,645) or decreased excision (n = 3,593) in CLL cell lines infected with U1$^{g.3A>C}$. Increased excision and decreased excision represent intron clusters that have significantly mis-spliced introns with ΔPSI > 0 and ΔPSI < 0, respectively. On the right, category of mis-splicing events in CLL cell lines infected with U1$^{g.3A>C}$. The number of introns is coloured by the sixth base of 5′SS. The 6th base of the sequence corresponds to the one interacting with base 3 of *U1*. The definitions of each category is: "annotated" if the junction matches any annotated introns; "novel annotated" if both splice sites are annotated but not paired; "cryptic 5′ splice site" if only the 3′ splice site is annotated; "cryptic 3′ splice site" if only the 5′ splice site is annotated.

Characterization of U1$^{g.3A>C}$ CLL cell lines

The confirmation that CLL lines carrying the g.3A>C mutation in *U1* snRNA showed the same changes in splicing as patients with this same mutation, allowed us to use this cell model to explore the potential differences in cells carrying these mutations versus those expressing the U1^{wt}.

Proliferation and apoptosis in U1 g.3A>C cells

To test whether the expression of the U1$^{g.3A>C}$ mutation had an impact on proliferation when compared to control cells (those infected with lentiviral vector carrying the U1^{wt} allele), we performed a proliferation assay using MTT. We did not observe differences in cell proliferation between U1$^{g.3A>C}$ expressing-cells and U1^{wt} control cells in any of the three CLL cell lines analyzed (JVM3, HG3 or MEC1) (Figure 36a). These results indicate that the transcriptomic changes caused by the g.3A>C mutation in *U1* snRNA do not appear to have an impact on proliferation.

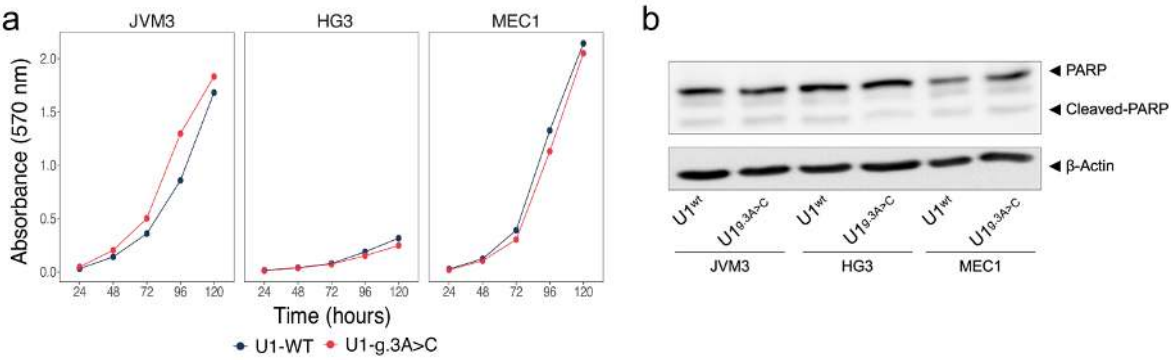

Figure 36. Characterization of CLL cell lines harboring the g.3A>C mutation in *U1*. A) Proliferation assay (MTT) carried out during five days with the three cell lines (JVM3, HG3 and MEC1) infected with pLKO.1-U1^{wt} (blue) or pLKO.1-U1$^{g.3A>C}$ (red). B) Western blot analysis of PARP to see differences in apoptosis induction (PARP cleavage) between cells expressing U1^{wt} and U1$^{g.3A>C}$ for the three cell lines (JVM3, HG3 and MEC1).

Complementary studies revealed that no differences in apoptosis could be detected between U1$^{g.3A>C}$ mutated cells and control ones. Thus, there was no change in the ratio between PARP and cleaved-PARP in any of the three studied cell lines when comparing *U1$^{g.3A>C}$* vs *U1^{wt}* (n = 2) (Figure 36b). This is in concordance with what was observed in the RNA-seq analysis, since although down-regulation of genes involved in apoptosis was detected in patients, this was not observed when cell lines were analyzed.

Sensitivity to Bortezomib and Pladienolide B in U1$^{g.3A>C}$ cells

The results obtained above show that the g.3A>C mutation in *U1* snRNA causes a major dysregulation of splicing. This process will presumably result in the generation of many aberrant proteins, many of them might not fold properly and are likely to be degraded. Due to the relevance of the proteasome in protein degradation and recycling, we hypothesized that the g.3A>C mutation in U1 might result in a higher dependency on the activity of the proteasome, making cells expressing U1$^{g.3A>C}$ more sensitive to its inhibition. To test this hypothesis, cells were incubated in the presence of the proteasome inhibitor Bortezomib (JVM3 and HG3 at 0 nM, 6 nM and 9 nM and MEC1 at 0 nM, 3 nM and 6 nM) and cell viability was measured at 24 h and 48 h (n = 2). We observed that JVM3-U1$^{g.3A>C}$ and HG3-U1$^{g.3A>C}$ cells had higher mortality than their respective controls at a concentration of 6 nM of Bortezomib although this result was not statistically significant (Figure 37a). When the experiment was repeated with an increased number of replicates (n = 6) JVM3-U1$^{g.3A>C}$ seemed to be more sensitive to 6 nM of Bortezomib than JVM3-U1^{wt} (Figure 37b). However, these results were still not robust, probably due to the high variability. This trend could not be validated either by proliferation assays or by looking at apoptosis through PARP activation (Figure 37c,d). On the other hand, for HG3 there was no difference in cell viability at 6 nM nor 9 nM of Bortezomib between those expressing U1$^{g.3A>C}$ or U1^{wt} (Figure 37b). Together, these results suggest that the U1$^{g.3A>C}$ mutation does not increase the sensitivity of cells to proteasome inhibition.

Furthermore, we tested the sensitivity of these cells to Pladienolide B, a spliceosome inhibitor. Specifically, this drug inhibits SF3B1, which is responsible for recognizing the 3'SS. Because it appears that mutations in *U1* and *SF3B1* are mutually exclusive in CLL, we reasoned that cells carrying the g.3A>C mutation in *U1* could be more sensitive to this drug. However, neither cell viability assays carried out in JVM3 and HG3 nor proliferation assays (n = 2) in JVM3, with PladB concentrations of 0 nM, 10 nM, 25 nM and 50 nM, at 24 h and 48 h showed differences between those expressing *U1^{wt}* and *U1$^{g.3A>C}$* (Figure 38a,b). The same results were obtained when the cell viability assays were repeated with lower concentrations of PladB (0 nM, 1.25 nM, 2.5 nM, 10 nM and 20 nM) (Figure 38c). There were also no differences in apoptosis (n = 2) after 2 h and 4 h of incubation with 50 nM PladB (Figure 38d).

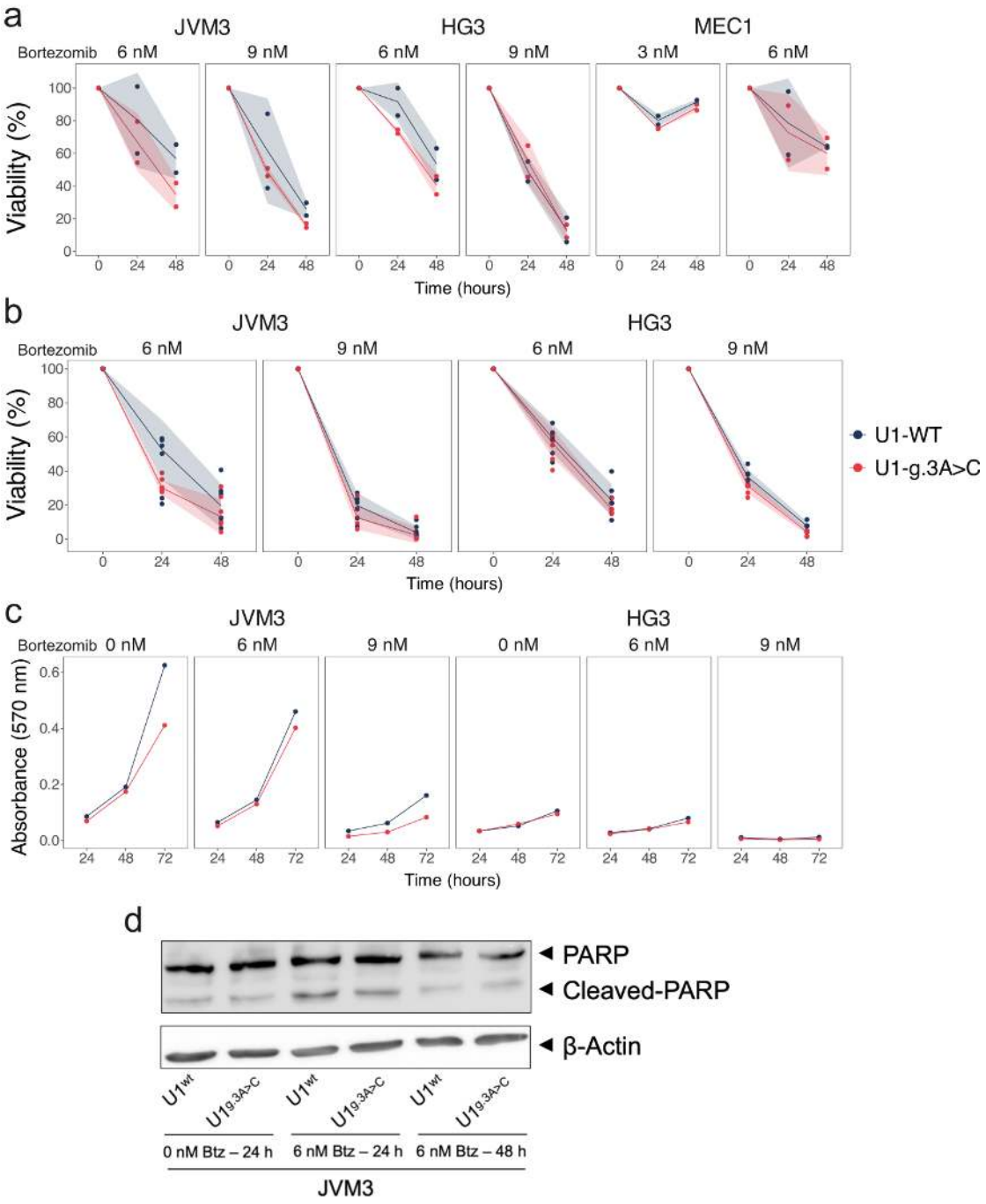

Figure 37. Cell viability, proliferation and apoptosis assays performed on JVM3, HG3 and MEC1, $U1^{wt}$ (blue) and $U1^{g.3A>C}$ (red), in the presence of different concentrations of Bortezomib (Btz). a) Cell viability assay (n = 2) on JVM3, HG3 and MEC1 cell lines harboring $U1^{wt}$ or $U1^{g.3A>C}$ at 24 h and 48 h, and with two diferent concentrations of Btz, 6 nM and 9 nM for JVM3 and HG3, and 3 nM and 6 nM for MEC1. b) Repetition of the cell viability assay for JVM3 and HG3 with the same conditions, but increasing the number of replicates (n = 6). c) MTT proliferation assay for JVM3 and HG3 ($U1^{wt}$ and $U1^{g.3A>C}$) cells without Btz (control condition) vs 6 nM and 9 nM of Btz. d) Western blot analysis of PARP cleavage as an indicator of apoptosis in control (0 nM) and at two different times (24h and 48h) after incubation with 6 nM of Btz in JVM3-$U1^{wt}$ and JVM3- $U1^{g.3A>C}$ cell lines.

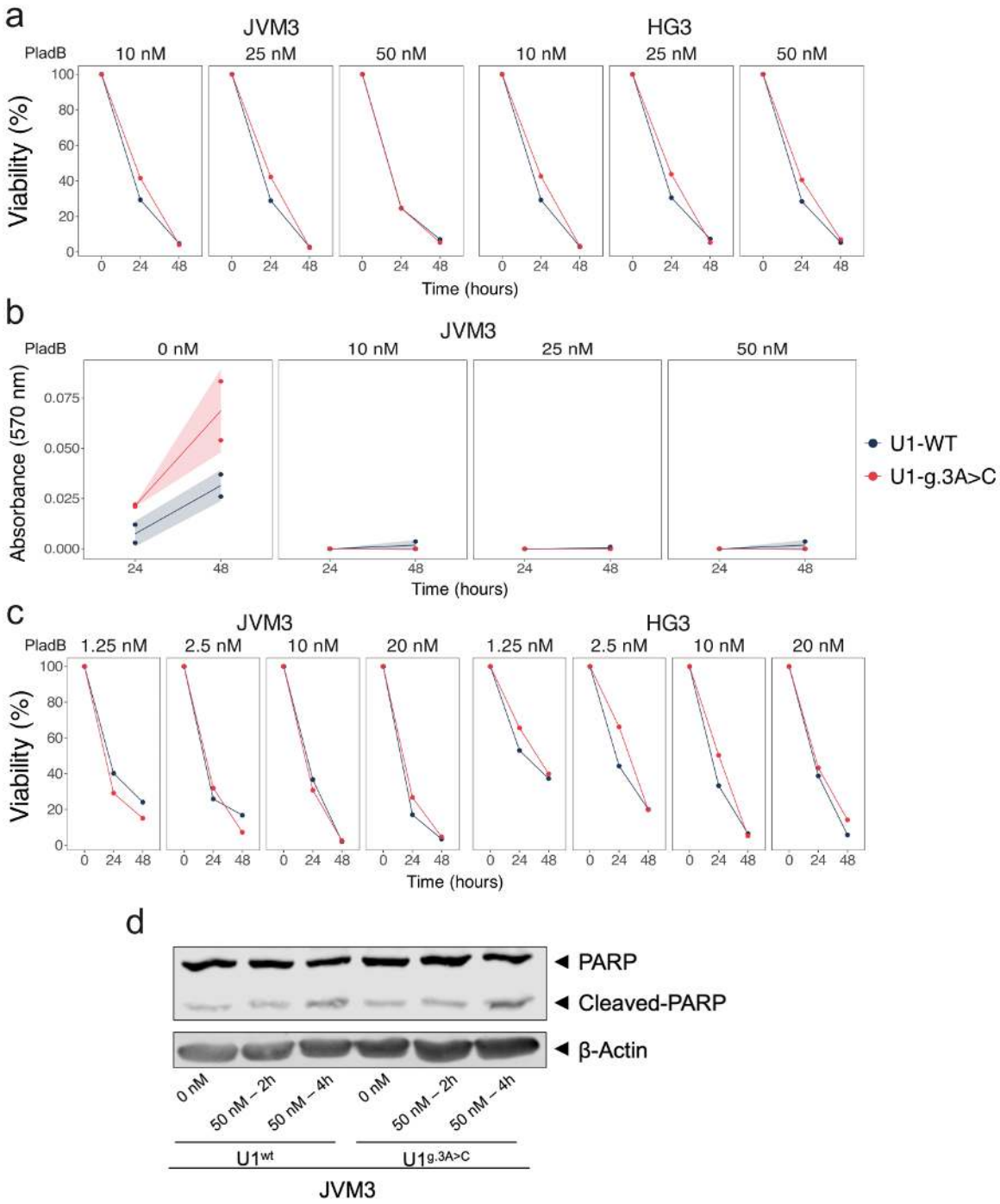

Figure 38. Cell viability, proliferation and apoptosis assays performed on JVM3 and HG3, U1^{wt} (blue) and U1$^{g.3A>C}$ (red), in the presence of different concentrations of PladienolideB (PladB). a,c) Cell viability assay on JVM3 and HG3 cell lines harboring *U1^{wt}* or *U1$^{g.3A>C}$* at 24 h and 48 h, and with three diferent concentrations of PladB, a) 10 nM, 25 nM and 50 nM or c) 1.25 nM, 2.5 nM, 10 nM and 20 nM. b) MTT proliferation assay (n = 2) for JVM3-*U1^{wt}* and JVM3-*U1$^{g.3A>C}$* cells without PladB (control condition) vs 10 nM, 25 nM and 50 nM of PladB. d) Western blot analysis of PARP cleavage as an indicator of apoptosis in control (0 nM) and at two different times (2h and 4h) after incubation with 50 nM of PladB in JVM3-*U1^{wt}* and JVM3- *U1$^{g.3A>C}$* cell lines.

Effect of U1$^{g.3A>C}$ mutation in lncRNA localization

Although the main function of U1 snRNA is the recognition of the 5'SS, it has recently been discovered that U1 also promotes the chromatin recruitment of long non-coding RNAs[165]. In particular, this has been seen in those lncRNAs with a higher density of predicted U1 recognition sites, as these are identified by the same 8 bases (from base 3 to 10 of the U1 snRNA) that recognize the 5'SS. Thus, similar to the effect of U1$^{g.3A>C}$ mutation altering the splicing of multiple genes, the presence of a mutation in a critical region involved in lncRNA localization might affect its localization in the chromatin. In order to test whether lncRNAs enriched with the new U1$^{g.3A>C}$ motif might be more associated to the chromatin than in normal conditions, we performed subcellular fractionation of chromatin, nucleoplasm and cytoplasm of JVM3 and HG3 cells infected with either pLKO.1-U1^{wt} or pLKO.1-U1$^{g.3A>C}$, followed by RNA sequencing.

RNA-seq analysis revealed that the chromatin-derived RNA was highly enriched in introns (52.3% vs. 19.4%) and intergenic regions (8.5% vs. 3.4%) when compared to total RNA. These results were expected, since RNA synthesis occurs in the chromatin, and therefore unprocessed RNA should be enriched in this fraction. We used ERCC RNA spike-in to normalize the data. After normalization and removing the cell line batch effect, the number of transcripts per million was calculated, filtering out genes with an expression of less than 3 TPMs. Only 11/259 lncRNAs had a chromatin retention ratio greater than 1.5 (Table 12). One of them, the Small Nucleolar RNA Host Gene 1 (*SNHG1*), was the only transcript in which a new strong donor site was recognized by the 5'SS sequence of U1$^{g.3A>C}$, while 5/11 transcripts had new medium donor sites. However, many other genes with several predicted new strong and medium donor sites did not appear associated to the chromatin, suggesting that the mutation might not have a major impact on this activity of U1.

Table 12. List of lncRNAs that appear associated to the chromatin in JVM3 and HG3-U1$^{g.3A>C}$ cell lines

Identifier	Symbol	Chromatin retention ratio	New strong site	New medium site	Density of strong sequences*	
					U1^{wt}	U1$^{g.3A>C}$
ENSG00000246582	LOC389641	1.92	-	-	-	-
ENSG00000205559	CHKB-DT	1.77	0	2	0.40	0.63
ENSG00000261520	DLGAP1-AS5	1.76	0	1	0.39	0.52
ENSG00000261879	LOC100130950	1.70	-	-	-	-
ENSG00000236618	PITPNA-AS1	1.65	0	1	0	0
ENSG00000233429	HOTAIRM1	1.63	0	1	0.22	0.43
ENSG00000234327	LOC101928000	1.63	-	-	-	-
ENSG00000254319	LOC101927815	1.62	-	-	-	-
ENSG00000249846	LINC02021	1.62	0	8	0.40	0.56
ENSG00000253633	LOC107986898	1.59	-	-	-	-
ENSG00000255717	SNHG1	1.59	1	0	1.51	1.76

**Prediction of the number of strong recognition sequences for U1 (wild-type or g.3A>C) per kilobase of gene.*

Analysis of U1$^{g.3A>C}$ -neoepitope-induced immune response

Due to the observed effect of the U1$^{g.3A>C}$ mutation on splicing affecting many proteins, we hypothesized that some of the aberrant proteins that are produced might contain neoepitopes that could be recognized by the immune system. To test our hypothesis, protein extracts from the three cell lines (JVM3, HG3 and MEC1), infected with either pLKO.1-U1^{wt} or pLKO.1-U1$^{g.3A>C}$, were separated in an SDS-PAGE gel and transfer to a PVDF membrane. Membranes were then incubated with 100 µL of serum containing the antibodies from 6 CLL patients with the *U1$^{g.3A>C}$* mutation and 5 patients without mutated *U1*. Due to the overlap between proteins with altered splicing between cell lines and patients, if antibodies against any of those neoantigens were present in serum from these patients, they might react against altered proteins produced by the cell lines expressing mutated U1. However, although serum from these patients recognized different proteins in the protein extracts derived from the cell lines, we could not detect any differences between lanes loaded with proteins derived from U1^{wt} or U1$^{g.3A>C}$ cell lines for any of the patients (Figure 39).

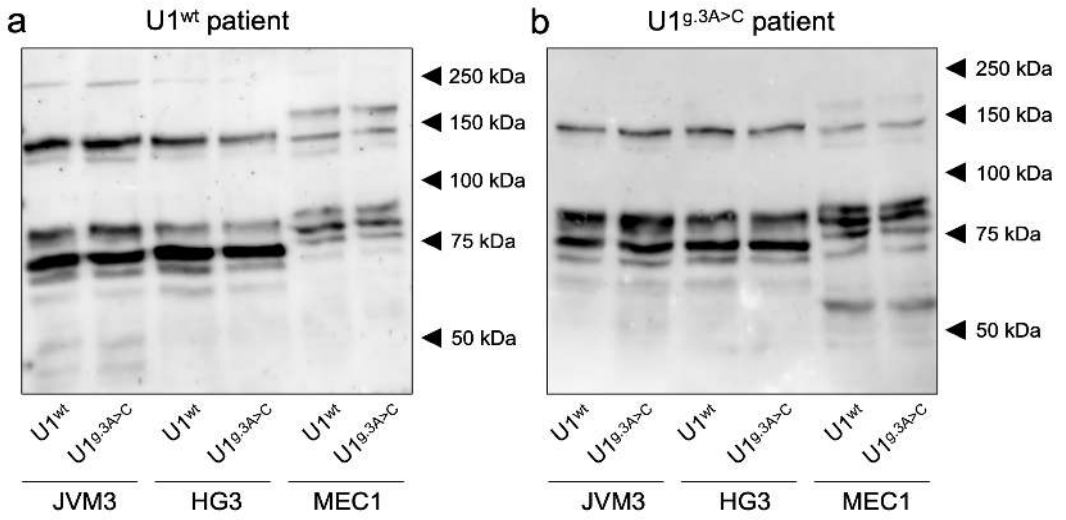

Figure 39 Western blot for detecting the presence of neoepitopes. Protein extracts from JVM3, HG3 and MEC1 expressing $U1^{wt}$ or $U1^{g.3A>C}$ were incubated with a) serum from patients with *U1*^{wt} (control) or b) patients with *U1*$^{g.3A>C}$. If there is a protein with neoepitopes that could be detected, it should appear as a specific band in the extracts of the mutated cell lines when the membrane is incubated with serum from patients who also have the mutation.

Elucidating the mechanism of action of mutations located in the 3'UTR of *NFKBIZ*

During the course of this thesis we participated in the integrative analysis of more than 1,100 CLL patients in collaboration with the group of Drs. Cathy Wu and Gaddy Getz. In a previous work by Dr. Gutiérrez-Abril focused on the analysis of mutations in non-coding regions, we identified five patients with mutations in the same highly conserved 3'UTR region of *NFKBIZ* (Figure 40), a gene that codifies for the NF-kappa-B inhibitor zeta (IκBζ), suggesting that it might constitute a putative driver gene in CLL. To further investigate the potential relevance of this non-coding region in cancer, we extended the analysis of somatic mutations in the 3'UTR of *NFKBIZ* to all available ICGC and TCGA cases. We discovered mutations affecting this region in eleven other cases of malignant lymphoma, four cases of diffuse large B-cell lymphoma (DLBCL) and at a lower frequency in other tumors. Previous studies by Gutiérrez-Abril with one of these mutations (3:g.101578254CAGTT>C) revealed that this mutation did not have a major impact on mRNA stability, but it suggested that its effect was related to post-transcriptional regulation. However, further experiments were needed to know what effects these mutations have on the cell and their mechanism of action[77].

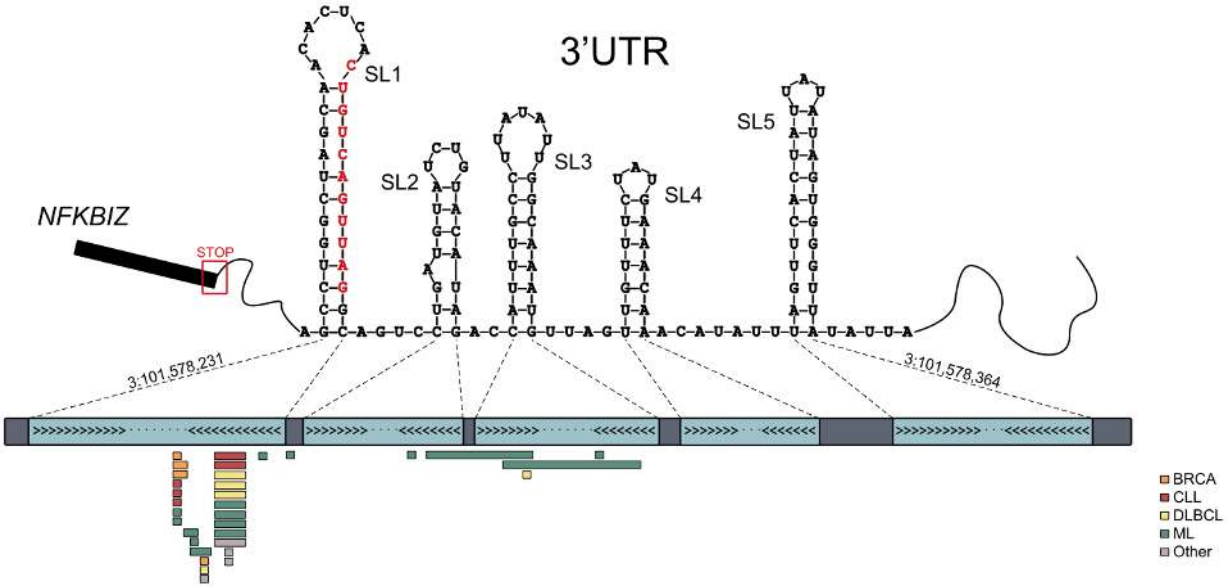

Figure 40. Secondary structures in the 3'UTR of *NFKBIZ*. On top are the five loops (SL1-5) that are formed at the beginning of the 3'UTR of *NFKBIZ*. The nucleotides in red are those belonging to the mutational hotspot. The distribution of mutations detected in different tumor types along this structure can be seen below. BRCA, breast cancer; CLL, chronic lymphocytic leukemia; DLBCL, Diffuse large B-cell lymphoma; ML, malignant lymphoma; Other, head and neck squamous cell carcinoma, colon cancer, biliary tract cancer and endometrial cancer.

Extending the analysis of gene expression in mutant 3'UTR constructions

In order to explore the mechanism by which these mutations in the 3'UTR of *NFKBIZ* might contribute to tumor development, six new mutations were cloned into the psiCHECK2 plasmid to test how they affected mRNA stability and post-transcriptional regulation (Table 13).

Table 13. List of NFKBIZ mutations analyzed by luciferase assays and qPCR

Chr	Position*	Ref	Alt	ID
3	101578250	C	G	250_C>G
3	101578250	C	A	250_C>A
3	101578250	C	CT	250_insT
3	101578254	CAGTT	C	254_delAGTT
3	101578285	GACCATTTGCCTT	G	285_del12
3	101578292	TGCCTTATATTGGCAAA	T	292_del16
3	101578304	G	A	304_G>A

**The mutation previously validated is underline.*

To determine the effect of these mutations on mRNA stability and protein production, we transfected those vectors in HeLa cells (n = 3), and luciferase mRNA expression was analyzed by RT-qPCR, while production of luciferase was assayed using the luciferase assay. This analysis revealed that all mutations caused a 5 to 20-fold increase in protein expression (Figure 41a), with the exception of g.101578304G>A, which was the only one located outside the loops. However, these mutations barely affected the stability of Renilla luciferase mRNA (n = 3) (Figure 41b).

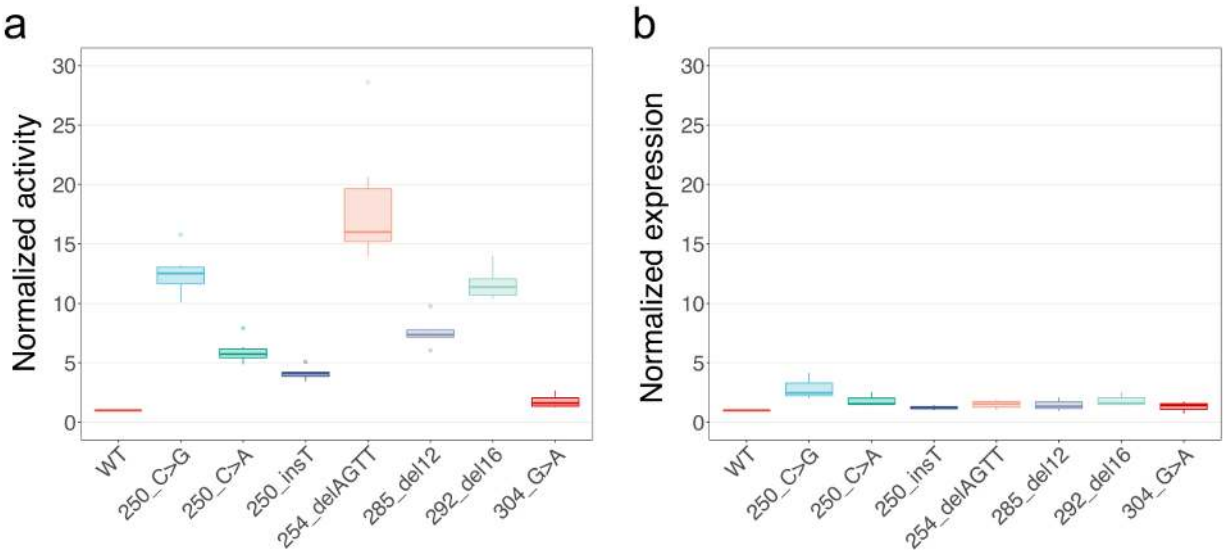

Figure 41. Effect of mutations in the 3'UTR of *NFKBIZ* on luciferase and mRNA expression. a) Normalized luciferase activity of Renilla protein in luciferase assays. b) Normalized expression of Renilla mRNA in luciferase assays. Data were normalized to NFKBIZ-3'UTR^{wt} values.

The expression of IκBζ is stimulated by interleukin 1 alpha through the MYD88 pathway, resulting in a rapid translation of previously synthesized mRNAs. Due to the importance of MYD88 mutations in CLL pathogenesis, we decided to test whether addition of IL1 could have an effect on the activity of the mutations observed in the 3'UTR of *NFKBIZ*. To test it, HeLa cells were transfected with different vectors containing the luciferase cDNA with the 3'UTR of *NFKBIZ* and the different mutations, and 4 h before the luciferase assay or the RNA extraction they were stimulated with IL1 (n = 3).

To perform the analysis, we compared how much Renilla luciferase activity, or its mRNA expression, was increased between cells stimulated with IL1 or unstimulated. In this regard, the conditions that showed the greatest increase in luciferase protein expression after IL1 stimulation were cells transfected with the wild-type or the g.10157830304G>A constructions. In fact, it was observed that those mutations that caused a greater increase in luciferase expression under normal conditions were those that achieved a smaller increase after stimulation with IL1 (Figure 42a). This might indicate that mutations in the 3'UTR of *NFKBIZ* have the same target with which IL1 modulates its translational levels. As under normal conditions, no change in mRNA stability was observed after IL1 stimulation (n = 2) (Figure 42b).

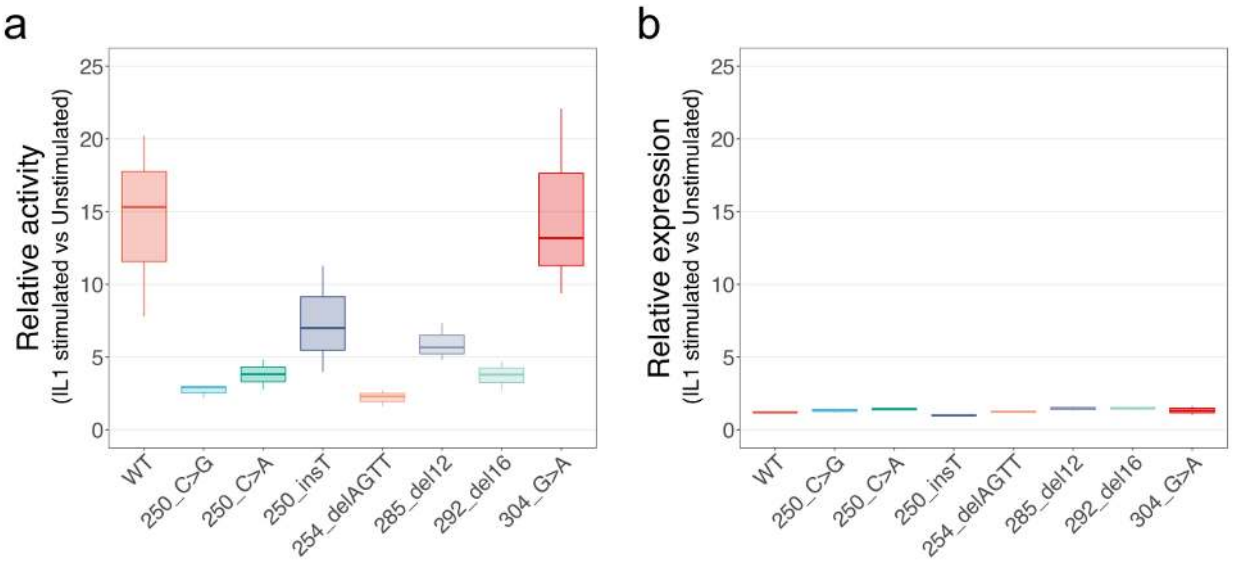

Figure 42. Effect of mutations in the 3'UTR of *NFKBIZ* on luciferase and mRNA expression under IL1 stimulation. a) Relative luciferase activity of Renilla protein in luciferase assays. b) Relative expression of Renilla mRNA in luciferase assays. For each mutation, the results correspond to the increase in activity or expression shown when comparing the values obtained under IL1 stimulation with respect to the unstimulated control.

Generation of clones with NFKBIZ-3'UTR mutations

To further study the downstream effects these mutations could have on cells, CRISPR knock-ins (KIs) were generated in HeLa and HEK-293T cell lines to produce the g.101578254CAGTT>C and g.101578250C>G mutations in the *NFKBIZ* gene, respectively (Figure 43a), as these mutations showed the greatest increase in protein expression. Although no clones with the desired mutations were obtained in HeLa cells, we did obtain a homozygous clone for the *NFKBIZ*-3'UTR$^{g.101578250C>G}$ mutation (Figure 43b) and another homozygous clone with *NFKBIZ*-3'UTR$^{g.101578255AGTTAC>A}$ (255_delGTTAC) (Figure 43) in HEK-293T. The latter, although it does not delete exactly the same four nucleotides as the g.101578254CAGTT>C mutation present in the patients, is located in the same region, so we decided to include it in the following studies. In addition, two other clones were used as negative controls: a clone with *NFKBIZ*-3'UTR^{wt} (no mutation was introduced after transfection with CRISPR-Cas9 and ssODN) and another homozygous clone that recurrently appeared with the *NFKBIZ*-3'UTR$^{g.101578254C>CA}$ mutation (254_insA) (Figure 43d).

Next, to verify that the clones had increased expression of IκBζ, both under normal conditions and after stimulation with IL1, as seen in luciferase assays, a western-blot assay was performed (n = 3). Thus, we could see how under normal conditions, IκBζ was not detectable in wild-type cells or in cells with the control g.101578254C>CA mutation, but it was highly produced in clones with the pathogenic variants (*NFKBIZ*-3'UTR$^{g.101578250C>G}$ and *NFKBIZ*-3'UTR$^{g.101578255AGTTAC>A}$) (Figure 44a). In contrast, stimulation with IL1 induced the expression of IκBζ in all analyzed clones (Figure 44b). Furthermore, due to the activity of IκBζ in the Nuclear Factor kappa B (NF-κB) pathway, we investigated whether increasing IκBζ, under normal conditions, could affect NF-κB expression, since IκBζ is upstream in the signaling pathway. In this regard, the overexpression of IκBζ caused by mutations in its 3'UTR (g.101578250C>G and g.101578255AGTTAC>A) results in a slight increase of p52, the active subunit of NF-κB, when compared to its expression in controls (Figure 44c).

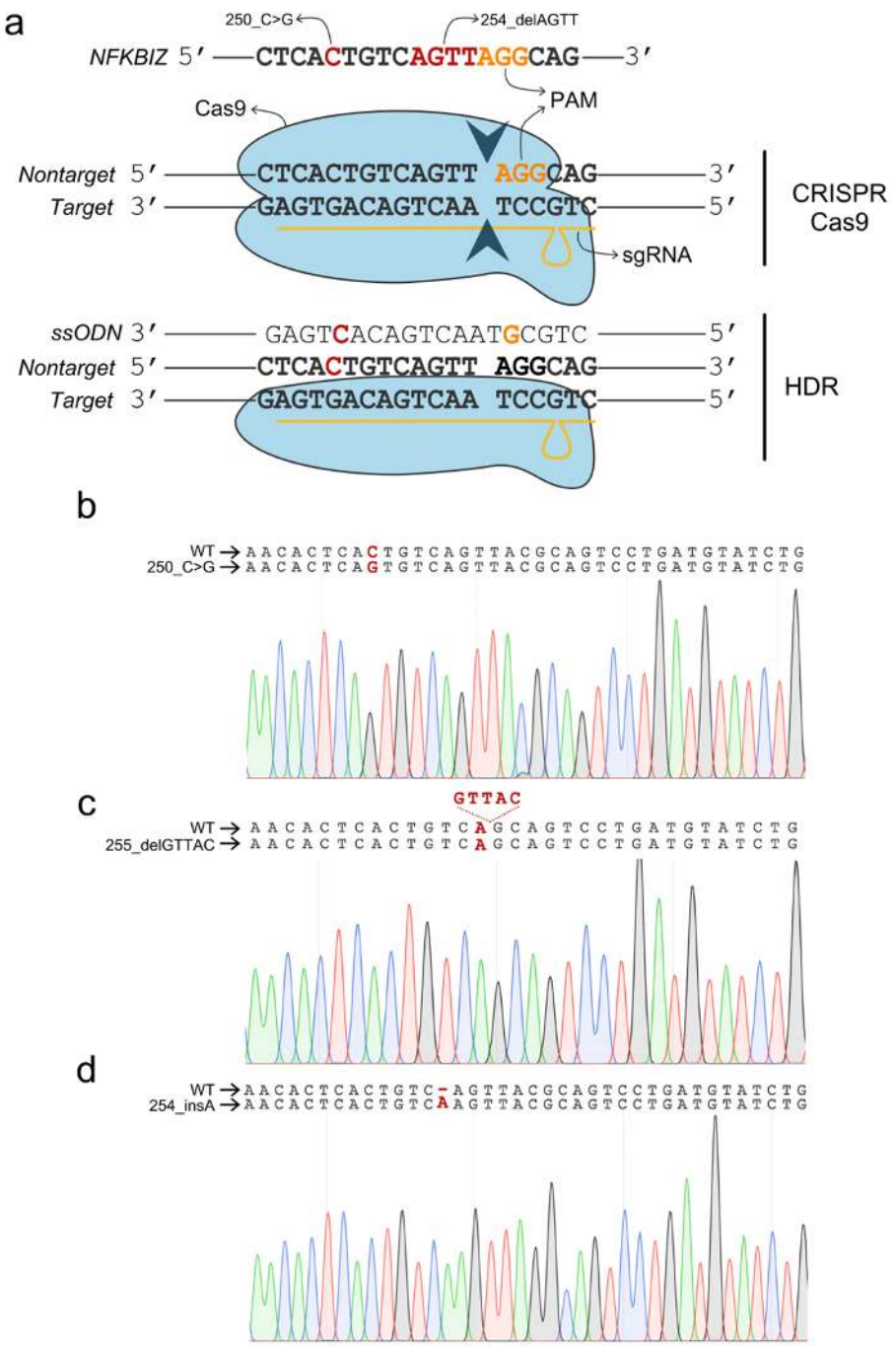

Figure 43. Generation of HEK-293T clones with mutations in the 3'UTR of *NFKBIZ* using the CRISPR-Cas9 system. a) On the top we find the sequence of the 3'UTR where the mutational hotspot is located. Mutations 250_C>G, 254_delAGTT and the PAM sequence that is recognized by the Cas9 protein have been indicated. Specifically, the sgRNA guides Cas9 to the complementary strand where the PAM sequence is located, also known as the target strand. The nuclease domains of Cas9 then cause a cut in the double strand just upstream of the PAM sequence. The ssODN binds to the nontarget strand by complementarity and the mutations are inserted into the genome by homology-directed repair (HDR) mechanism. In order to prevent Cas9 from repeatedly cleaving the target region, a change in the PAM sequence is inserted via ssODN. b-d) Electropherograms to genotype the changes produced by CRISPR-Cas9 for the clones with 250_C>G, 255_delGTTAC and 254_insA mutations, respectively.

We also tested whether the presence of mutations in the endogenous 3'UTR of *NFKBIZ* affected the stability of its mRNA (n = 2), but as we had observed in HeLa transfected with the 3'UTR fused to the Renilla luciferase CDS, no major differences were observed to explain the significant increase in the expression of IκBζ (Figure 44d).

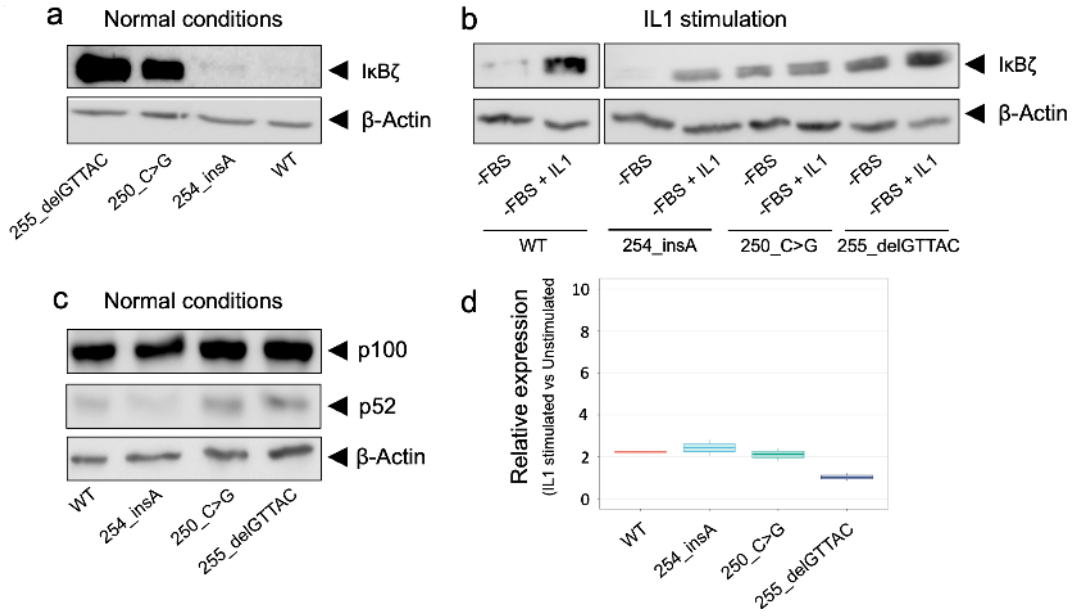

Figure 44. Analysis of HEK-293T knock-ins generated by CRISPR-Cas9. The same clones harboring the 255_delGTTAC, 250_C>G and 254_insA mutations or the wild-type isoform of *NFKBIZ* were used in all the experiments. a) Western blot analysis of IκBζ expression in normal conditions or b) under IL1 stimulation (cells were cultured without FBS and then incubated with IL1 for 4 h). c) Study to see if IκBζ overexpression alters NF-κB expression levels. P100 is the inactive protein, whereas p52 is the active subunit. d) Relative expression of *NFKBIZ* mRNA. For each mutation, the results correspond to the increase in expression shown when comparing the values obtained under IL1 stimulation with respect to the unstimulated control.

Next, we explored whether the presence of a mutation in the 3'UTR of *NFKBIZ* could lead to differences in proliferation. Thus, we compared clones with *NFKBIZ*-3'UTR$^{g.101578250C>G}$ and *NFKBIZ*-3'UTR$^{g.101578255AGTTAC>A}$ versus clones with *NFKBIZ*-3'UTR^{wt} or *NFKBIZ*-3'UTR$^{g.101578254C>CA}$ using a proliferation assay (n = 2). In addition to the clones used in the previous experiments, other independent clones for *NFKBIZ*-3'UTR^{wt} and *NFKBIZ*-3'UTR$^{g.101578254C>CA}$ were added for this assay to reduce the possibility of bias due to clone selection. No extra clones were added for the other mutations as they could not be obtained. Contrary to what it could be expected, clones with a higher expression of IκBζ were the ones with slower growth (Figure 45). Moreover, these are the only ones together with a wild-type clone (WT2) that do not show different proliferation rate

under normal conditions or with IL1 stimulation. On the other hand, the fact that the remaining clones show slower growth under IL1 stimulation is likely due to the fact that they are maintained without FBS, and it is possible that HEK-293T cells are not as dependent on IL1 for proliferation as immune cells.

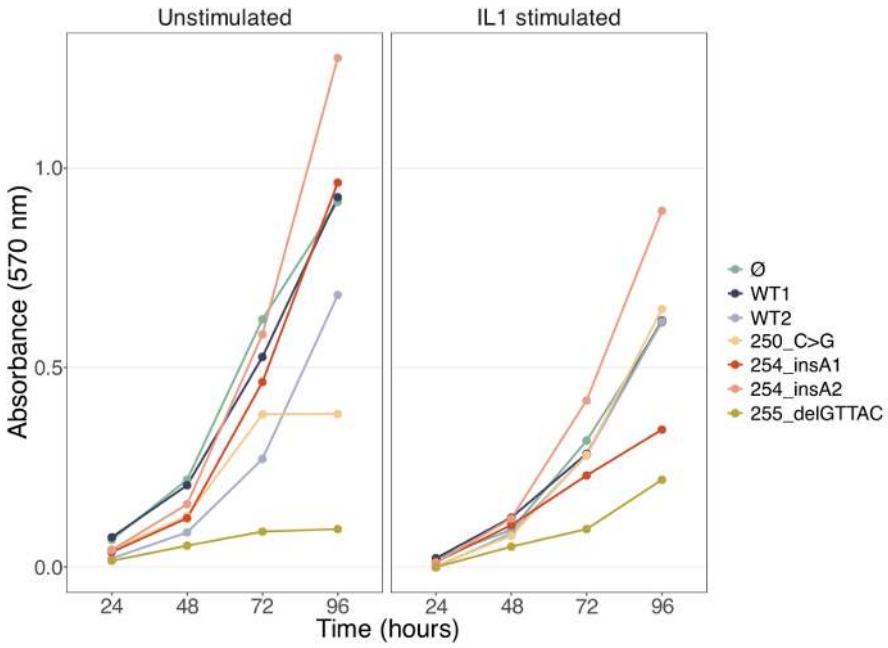

Figure 45. Proliferation assays performed on HEK-293T CRISPR clones with mutations in the 3'UTR of *NFKBIZ*. Unstimulated cells were maintained without FBS, while the stimulated cells were maintained without FBS but stimulated with IL1 (refreshed daily). For wild-type and 254_insA, two different clones were used. Ø represent cells that did not undergo clone selection.

Finally, RNA-seq (2 *NFKBIZ*-3'UTR^{wt} vs 2 *NFKBIZ*-3'UTR$^{g.101578250C>G}$) was performed to determine which genes and pathways might be altered due to overexpression of IκBζ. In total, we detected 797 genes that were significantly deregulated (539 down and 258 up). After pathway enrichment analysis we found that there was an over-representation of altered genes that participate in the "positive regulation of NF-κB transcription activity" gene ontology (C5 gene set from MSigDB). This would make sense since *NFKBIZ* is a regulator of this pathway. However, this result was no longer significant once we corrected for FDR. Although there were other altered pathways and categories, none of them were related to CLL. In this regard, it must be taken into account that the analysis was performed with RNA from HEK-293T cells, derived from embryonic kidney, rather than B cells. Further analyses in CLL-derived cells could help determine the putative alteration of the NF-κB pathway by mutations in the 3'UTR of NFKBIZ.

Elucidating the mechanism of action

The 3'UTR of *NKFBIZ* is highly conserved near the mutation hotspot, and it has been described that at least two endonucleases, MCPIP1 (also known as Regnase-1) and Roquin1/2 (coded by RC3H1/2 genes) bind to this region[166–168]. However, our studies do not support an effect of these mutations in mRNA stability. Furthermore, these proteins bind to loops 4 and 5, whereas the mutations detected in patients are mainly found in loop 1. We hypothesized that there is at least a third protein that binds to this first loop inhibiting mRNA translation. The presence of these mutations or signaling by IL1 might affect the binding of this protein to the 3'UTR of *NFKBIZ*, preventing its repressor activity (Figure 46).

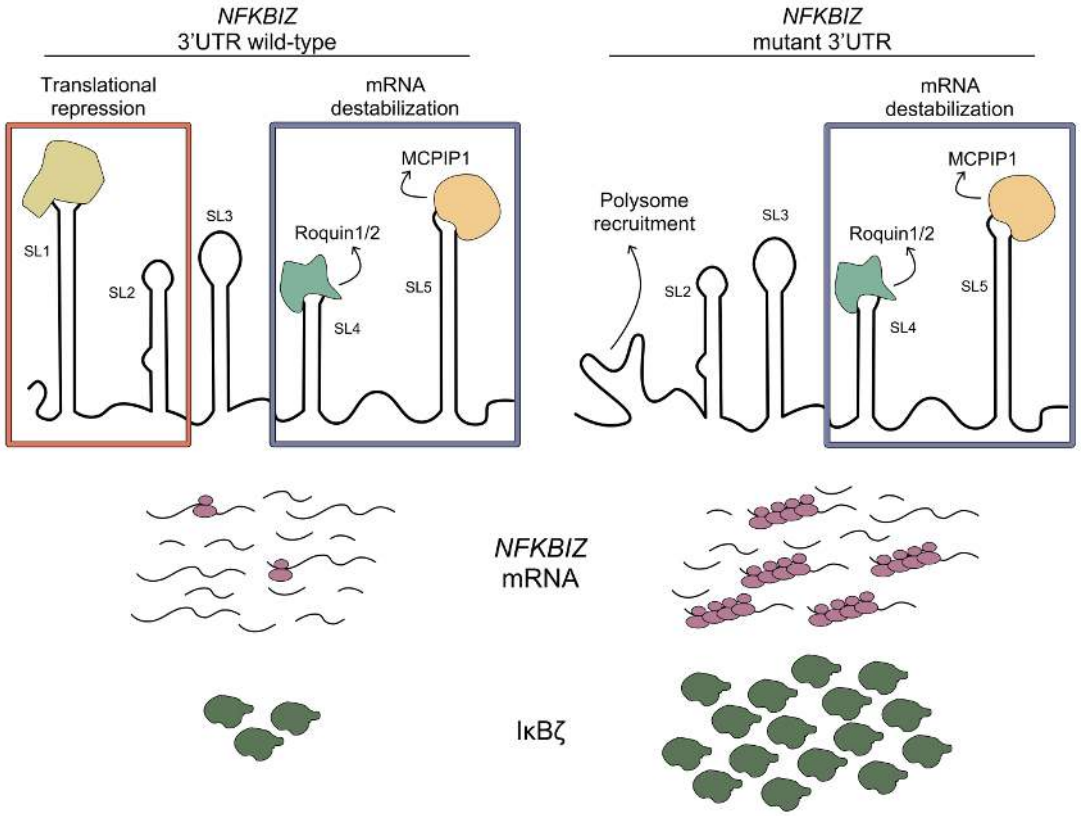

Figure 46. Model of the effect of the mutations in the 3'UTR of *NFKBIZ*. Under normal conditions MCPIP1 and Roquin1/2 bind to SL4 and SL5 promoting mRNA degradation. According to our hypothesis, there would also be a third protein recognizing SL1 and inhibiting protein translation. Mutations in SL1 destroy the loop structure preventing this putative protein from binding. This allows mRNA to be recruited to polysomes, causing overexpression of IκBζ.

To try to determine which proteins might be bound to this region of the 3'UTR of *NFKBIZ*, we decided to perform an *in silico* prediction using only the sequence corresponding to loops 1 and 2. For this purpose, we used two online tools: RBPmap and Attract. RBPmap predicted 41 regions (p-value < 0.05) in which 17 RNA binding proteins (RBPs) could bind, whereas Attract identified 60 regions for 31 RBPs. Among them, only 5 RBPs had been predicted by both programs (Table 14). However, most of them (4/5) are located in the nucleus and are related to splicing regulation. Only PUM2 exerts its function in the cytoplasm, the compartment in which the *NFKBIZ* mRNA is found. Moreover, this protein is involved in post-transcriptional repression[169], which is exactly the mechanism we wanted to investigate. In addition to PUM2, another protein of the same family and with the same function (PUM1) was also selected to perform a targeted screening to test whether these proteins are involved in the post-transcriptional regulation of *NFKBIZ*. We also chose IGF2BP2 and IGF2BP3 (predicted by Attract), two RBPs with translation repressor activity that are mainly located in the cytoplasm[170] and RPSAP52, a lncRNA that facilitates binding between IGF2BP2/3 and their target mRNAs[171]. Finally, MCPIP1 were added to the study to see how increased mRNA stability affected protein expression.

Table 14. RBPs that are predicted to bind to the loops 1 and 2 in the 3'UTR of NFKBIZ

RBP	Localization	Function
HNRNPA1	Nucleus	Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection
PUM2	Cytoplasm	Mediates post-transcriptional repression of transcripts via different mechanisms: acts via direct recruitment of the CCR4-POP2-NOT deadenylase leading to translational inhibition and mRNA degradation[169]
MBNL1	Nucleus	Mediates pre-mRNA alternative splicing regulation. Acts either as activator or repressor of splicing on specific pre-mRNA targets
PTBP1	Nucleus	Plays a role in pre-mRNA splicing and in the regulation of alternative splicing events
SRSF5	Nucleus	Plays a role in constitutive splicing and can modulate the selection of alternative splice sites

In order to knock-out these proteins, one shRNA plasmid from the MISSION RNAi library were selected per target protein. Then, they were co-transfected into HeLa cells together with the psiCHECK2-3'UTR^{wt} plasmids to perform luciferase assays (n = 4). Thus, we wanted to see the effect that these proteins had on luciferase expression and whether their inhibition led to increase luciferase activity. Two extra conditions were also transfected, psiCHECK2-3'UTR$^{g.101578254CAGTT>C}$ and psiCHECK2-3'UTR^{wt}, as positive and negative controls respectively (Figure 47). However, no differences in Renilla luciferase activity were observed between knock-down conditions.

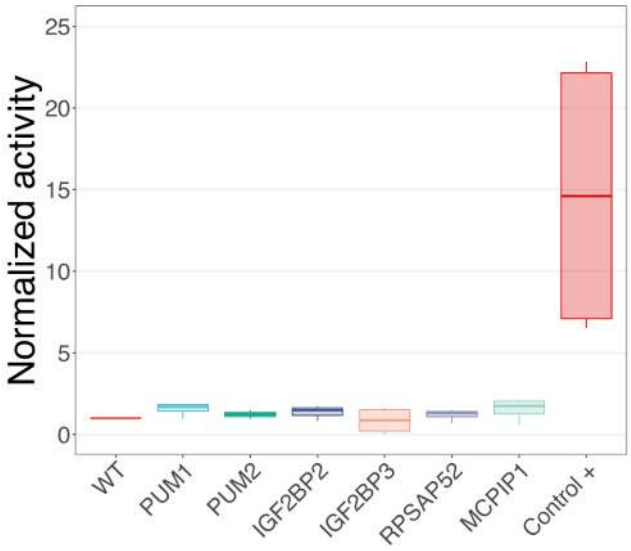

Figure 47. Effect of inhibition of proteins that might be interacting with the 3'UTR of *NFKBIZ* on Renilla luciferase expression. Assays were carried out in HeLa by co-transfecting shRNAs designed to knock out selected proteins together with plasmid psiCHECK2-3'UTR^{wt} (n = 4). "WT" and "Control +" conditions were transfected only with psiCHECK2-3'UTR^{wt} and 3'UTR$^{g.101578254CAGTT>C}$ vectors, respectively. Data were normalized to "WT" values.

DISCUSSION

New driver mutations and structural variants in MCL

The analysis of sequencing data, driven by the collaboration between the different groups that form the major international cancer genome sequencing consortia, has given us a very complete picture of the main molecular alterations present in the main types of cancer[8,21,36,56,70,78]. The integration of all this data has not been an easy task and has required more than five years and 10 million hours of CPU time[36]. However, there are still pending tasks, such as the analysis of other types of tumors not included within the ICGC or TCGA projects, and ensuring that all this knowledge reaches patients in the form of more precise diagnosis and new and better treatments. During this thesis, I have contributed to the mutational analysis of more than 60 patients with mantle cell lymphoma, a tumor type which had not been studied in depth to date[121,141,142]. The ability to analyze more than 60 tumors by WGS has allowed the characterization of this tumor type at an unprecedented resolution, allowing the detecting of new driver genes, and identifying a large number of structural alterations.

In this regard, we were able to observe that the t(11;14), an indispensable feature for the classification of both MCL subtypes, occurs by the same mechanism in cMCL and nnMCL. In fact, most breakpoints on chromosome 11 are located within the same 100 bp region (MTC) associated with the recognition motifs of the AID, an enzyme involved in the processes of somatic hypermutation and class switch recombination during the maturation of B lymphocytes. This, together with the fact that the MTC corresponds to an open chromatin region, allows the rearrangement of the immunoglobulin promoter (chr14) upstream of *CCND1* (chr11) during B cell development, resulting in overexpression of this cyclin and tumor transformation by deregulating the cell cycle[135]. However, it has been shown that *CCND1* overexpression is not sufficient to induce B-cell transformation, requiring the accumulation of additional oncogenic events[172]. This might explain the big difference between the two subtypes when tumor transformation is initiated by the same initial event.

To understand which other driver events might contribute to MCL development, we performed mutation calling of somatic mutations in these cases, followed by statistical analysis to identify genes with a significant number of mutations. We were able to discover genes altered in important mechanisms such as cell cycle regulation (*CDKN1B*), DNA replication (*SAMHD1*), RNA processing (*HNRNPH1*) or chromatin modification (*SMARCB1*), among others. In total, this resulted in the identification of 30 driver genes, 5 more than previously described[121]. An interesting fact is that, despite the increase statistical power to identify novel driver genes due to the larger number of analyzed cases, the number of novel driver genes identified was somehow modest, in comparison to other tumor types such as CLL. However, the parallel analysis of structural alterations in those WGS cases revealed the presence of numerous recurrent alterations, including several targeting specific genes, such as deletion of 13q14 (*RB1*) and 9p21 (*CDKN2A*) or amplification of 8q24 (*MYC*). These data suggest that contrary to other B-cell neoplasias such as CLL, MCL transformation appears to be more dependent on the accumulation of structural variants affecting genes involved in cell cycle, proliferation, and DNA damage repair, and less on point mutations in specific genes. From a chromosomal perspective, CLL is a very quiet entity, with four mayor chromosomal alterations (trisomy of chromosome 12, deletion of 13q14 and 6q15-21, and deletion of *TP53* or *ATM* at chromosomes 17 or 11, respectively). And at a minor frequency, t(14;18), t(14;19) and t(2;14)[155]. On average, each CLL tumor contains 2 structural variants at diagnosis. In contrast, MCL tumors have a higher number of structural variants (average 13) and with increase complexity, including alterations such as chromothripsis, kataegis or chromoplexia. The biological reason for this difference is currently unknown. Future studies could focus on timing their appearance in order to better understand the effect they have on tumor transformation.

In conclusion, this analysis has contributed to reveal the complexity of MCL and to explain at the genomic level the differences seen in disease progression between the two subtypes. Moreover, the identification of new driver genes and the mechanisms by which they are altered, opens the door to develop novel strategies and/or drugs for their treatment.

Improving the detection of somatic mutations

The identification and further revision of mutations in MCL whole genomes allowed us to identify weaknesses and potential room for improvement of some of the bioinformatics programs used for somatic mutation calling. In order to avoid program-specific bias, international groups usually combine several tools and keep the common results from all of them. This approach, while providing more accurate datasets, represents a very costly procedure both in time and computing resources, resulting in a somehow redundancy that increases the economic cost of this fundamental process of mutation calling. Nevertheless, the application of NGS techniques for clinical diagnosis in tumor samples requires procedures that provide enough sensitivity and specificity, while at the same time do not require large computing resources to achieve the analysis in a reasonable amount of time and under a tight economic budget. To increase accuracy, a final step of manual review through visual inspection is usually carried out for mutations that might be clinically informative. This manual revision increases the specificity, but at the cost of a labor intensive process. Recent advances in machine learning approaches are suitable to incorporate features that experts consider when distinguishing between *bona fide* mutations and false positives. However, most available programs that use machine learning approaches for somatic mutation calling have been trained with high depth of coverage WES, using *in silico*[99,100] or orthogonal validated mutations[101] and cannot be used for whole-genome analysis.

In this work, we have taken advantage of a manually curated dataset of real mutations with features that an expert curator might consider when manually reviewing a mutation in a research or clinical context. Thus, we have achieved a high sensitivity to detect SSNVs and small indels, while at the same time maintaining a low footprint, with low CPU and RAM consumption, being able to analyze a whole genome in less than 5 hours. Even more, this time can be reduced up to 3 hours when analyzing more than one case at the same time, since RFcaller has been designed to process multiple samples in parallel instead of analyzing them independently. This feature has not been seen in other variant callers and is very useful when analyzing many cases simultaneously. Another characteristic of RFcaller is the use of basic filters such as the minimum number

of reads, both wild-type and mutated, that have to appear to consider a mutation. These filters have two objectives, to eliminate false positives in the first steps in order to accelerate the pipeline, and to remove mutations for which we would not be sure they are real, even after manual review. In this sense, there are programs that identify a mutation with total coverages of 4 or 5 reads or when there are only 1 or 2 mutated reads. Despite the potential advantage of detecting those mutations, the use of those calls in the clinical practice should be supported by stronger evidence, and if the evidence is too scarce for an expert to make a call, the inclusion in a final report is somehow debatable.

Moreover, although RFcaller has been trained with WGS data, it has shown a good performance in exome samples. In fact, after manually revision of some of the false negatives, we saw that the main problem originates from the first call made with bcftools and not from the RFcaller regression algorithm. Upon further exploration we detected that most errors were due to a very high coverage of the region, which made it impossible to identify the mutation. Although the number of false negatives detected by WES was not very high, being able to improve the detection of mutations in this type of regions would be a great advance, not only for WES analysis but also to be able to use RFcaller with other techniques such as target deep-sequencing.

On the other hand, even though our selected features are often used by similar programs, most of them process SSNVs and indels following the same rules, when clearly the two types of mutations have different characteristics. In this regard, we analyzed SSNVs and indels separately, which allowed us to detect indels with higher accuracy without affecting the ability to detect SSNVs. Indeed, we have shown that RFcaller performance is similar to that of a combination of complex pipelines used in the PCAWG project to detect clonal mutations, with the ability to detect new ones. In fact, some of them have been found in driver genes, what might contribute to improve the detection of actionable mutations or mutations with clinical prognosis significance[173]. Furthermore, we showed that RFcaller was able to detect mutations even in the presence of some tumor contamination in the normal sample, a common problem in some hematological tumors that usually leads to false negatives with other pipelines.

However, in solid tumors, where the purity of the sample was not so high, RFcaller lost some sensitivity. This was caused by the fact that, due to the presence of normal DNA in tumor sample, the VAF of the mutations was slightly lower than when tumor purity is 100%. Therefore, some mutations were not detected or were detected with a very low quality. When compared to other pipelines, we demonstrated that most RFcaller false negatives were subclonal mutations with very low VAF, whose analysis might require additional tools and might not be as critical for taking clinical decisions. The most likely explanation for this loss of sensitivity in subclonal mutations in solid tumors is probably the fact that the training was performed using MCL samples with close to 100% tumor purity. Therefore, in order to recover this type of mutations in future versions of the tool, it could be trained using this reviewed set of subclonal mutations, so it can work in parallel to that of SSNVs and indels.

Additionally, an important point that tends to be ignored when evaluating a program is the difficulty of installing it and using it. In this sense, many of the programs used during this thesis were tedious to install, had requirements that were incompatible with other programs already installed or with newer versions of the operative system, had hardly any documentation or even did not work correctly and modifications in the source code were needed, modifications that are not suitable to the average user. In the last decade, tools such as Conda, which allows the creation of private environments with the needed requirements for the program to work, have been emerging. Another alternative that has become very popular is the use of containers such as Docker. This is based on the idea of creating a container, which can be understood as an independent virtual machine with its own lightweight operating system, that comes with all the requirements pre-installed along with the program. So, just by downloading this image, the program is ready to use, independently of the operative system or program versions installed in the machine where it is going to be executed.

Furthermore, RFcaller has been uploaded to the GitHub public repository allowing its use by other researchers and facilitating its improvement and addition of novel capabilities by building them upon existing versions. In conclusion, we have developed a pipeline called RFcaller, that is provided under a Docker system, which

allows its easy and fast installation without version incompatibilities. This tool allows the identification of clonal mutations with the same efficiency as *state-of-the-art* pipelines, but with a smaller footprint in computing resources, which might facilitate its implementation for clinical usage.

Improving the monitoring of pediatric MB tumors

Although RFcaller was still under development when we carried out this project in collaboration with Drs. Laura Escudero and Joan Seoane, it represents a good example of why it is necessary to have sensitive and specific mutational callers. Thus, during this thesis we tried to improve the diagnosis and follow-up of patients with pediatric medulloblastoma by analyzing the circulating tumor DNA present in the cerebrospinal fluid, and its comparison to WES of the primary tumor and normal DNA. Nowadays, the molecular characterization of MB provides the most accurate classification of this disease. The ability to detect those alterations before surgery would constitute a step forward in the management of these patients.

The analysis of WES mutations obtained from CSF-derived ctDNA and its comparison to WES from primary MB tumors revealed that ctDNA alterations recapitulated the main genomic alterations detected in the primary tumor, such as mutations in *TP53* and *PTCH1*, *MYCN* and *GLI2* amplifications, *SUFU* deletion or loss of 17p. This ctDNA was sufficient to detect not only point mutations (SSNVs and indels), but also copy number alterations. Furthermore, we were able to observe the same subclonal genomic landscape, which would allow the characterization of the intratumour heterogeneity, and the detection of germline mutations in known cancer genes. The data derived from these analyses provided sufficient information for MB molecular diagnosis and risk stratification, obtaining the same results as after direct analysis of tumor samples. This procedure has the advantage of using a much less invasive method that can be performed before surgery, as most of those patients have hydrocephalus and routine CSF extraction is performed, allowing a rapid diagnosis that can help to plan surgery.

On the other hand, WES analysis of CSF ctDNA also allowed us to characterize two cases with relapses. In fact, in one of the patients, who suffered from Li-Fraumeni syndrome, we were able to identify a completely new primary tumor that did not share somatic mutations with the original tumor. This fact is very important, since common follow up procedures, based on more targeted approaches such as RT-qPCR or droplet-digital PCR, will follow a particular mutation or a small set of mutations detected in the primary tumor. The presence of a completely novel tumor will be missed by these molecular approaches, but can be readily identified by WES from CSF-ctDNA. Therefore, being able to observe tumor evolution would allow us to identify new therapeutic targets, which would greatly benefit patients by offering more personalized therapies with fewer side effects.

Being able to classify tumors by CSF ctDNA prior to surgery has its advantages. In this way those patients with a better prognosis could benefit from a more conservative tumor resection followed by chemo- and radiotherapy, while in those with a shorter progression-free survival it would be much more important to achieve a complete tumor resection. We have also demonstrated that the analysis of sequential CSF ctDNA samples after surgery is useful for disease monitoring, as this would allow us to follow tumor evolution and adapt the intensity and duration of treatment. In this way, we detected two cases with minimal residual disease, not seen by radiologic imaging, which would allow them to be monitored more closely.

The use of sequencing technologies also allowed us to detect individuals with germline mutations in cancer predisposition genes. Thus, we discovered 10 mutations in 5 genes (*TP53*, *PTCH1*, *BRCA1*, *BRCA2* and *BLM*) related to hereditary cancer syndromes. In this regard, it was already known that mutations in *TP53*, *PTCH1* and *BRCA2* were associated with an increased predisposition to MB[174]. Specifically, mutations in *TP53* give rise to Li-Fraumeni syndrome, a syndrome that increases the predisposition to suffer various types of cancer at a very young age. On the other hand, alterations in *BLM* give rise to Bloom syndrome, which causes various pathologies and increases the risk of developing different types of tumors[175]. Recently, two cases with mutations in this gene were detected in a study of more than 1,000 patients with MB.

Due to its low prevalence, this gene was not considered to be related to a greater predisposition to MB[176]. However, in our study three patients with mutated *BLM* were detected. Taking into account that our cohort corresponded entirely to pediatric patients, where the importance of genetic predisposition is supposed to be higher, and that germline mutations in this gene have also been detected in other cohorts, our results suggest that *BLM* should be included as a predisposition gene for MB. Finally, we found a case that presented a germline mutation in *BRCA1*, a gene that has not been found to be related to an increased predisposition to MB. Nevertheless, this same patient also had a germline mutation in *TP53*, making it difficult to link *BRCA1* with an increased predisposition to MB. Finally, 2 of the 3 patients presenting germline mutations in *BRCA2* also had mutated *BLM*. In this sense, although the number of cases is too small to be able to confirm it, it seems that there is a relationship between mutations in these two genes and the probability of developing MB.

In summary, this work demonstrated that CSF ctDNA analysis is a non-invasive and accurate strategy for the characterization, including subtyping and risk stratification, and monitoring of pediatric patients with MB. Therefore, its application to other neoplasias of the central nervous system could impact the diagnosis and management of these patients.

Characterization of non-coding mutations in CLL

The search for driver mutations and structural alterations constitutes the initial step in our understanding of cancer at the molecular level. However, as important as identifying driver mutations is understanding the mechanism of action and the effects they will have on the biology of the cell. The last decade has seen an exponential increase in the number of cancer driver genes identified thanks to the advances in sequencing. However, only a tiny fraction of those statistically significant genes has been studied experimentally. In addition, these regions tend to be studied more often, since mutations in the coding sequence usually suggest a straightforward explanation, while the mechanism of action of non-coding mutations is more complex.

During this thesis, we have identified a recurrent mutation in *U1*, a repetitive and non-protein coding gene that codes for a snRNA that is part of the spliceosome. Interestingly, a recent study by our group has identified mutations in another repetitive and non-protein coding gene (*U2*), that also encodes a snRNA that is part of the spliceosome[177]. This highlights the importance of analyzing these type of regions, which had been excluded until now due to the difficulty in analyzing them and the more protein-centric approach when trying to interpret the plethora of mutations present in tumor genomes.

The mutations detected in *U1* are located in the third base of the transcript, modifying the 5' splice site recognition sequence. As expected, these mutations generated a global change in the splicing pattern of the cells, both *in vivo* and *in vitro*. Although the interpretation of mutations located in non-coding regions is a non-trivial process, our first hypothesis was that the g.3A>C mutation in *U1* could represent a gain of function mutation. This was supported by the recurrent mutation of the same base in many independent tumors, as well as for the existence of a large number of wild-type transcripts that continued to be expressed. Thus, the mechanism of action appears to be abnormal splicing. Recognition of the 5'SS is given by complementarity between bases 3-10 of *U1* and the last two bases of the exon together with the first six bases of the intron[178] (Figure 27). As we have seen in the results of the RNA-seq analysis, these mutations modified the preference for base 6 of the intron (complementary to the third base of *U1*), favoring a C for the $U1^{g.3A>G}$ mutation detected in MB or a G in the case of $U1^{g.3A>C}$ in CLL and HCC. Furthermore, among the different categories in which alternative splicing can occur, the mutations only caused changes in the cryptic 5'SS with an increased excision. That is, mutations caused the recognition of novel 5'SS, without participating in the selection of the 3'SS or other previously annotated sites, nor fails to recognize the canonical 5'SS. This confirmed the gain-of-function hypothesis, as the changes observed at the global level were very specific.

As we have already mentioned, both mutations cause a major change in the splicing pattern. However, if random dysregulation of multiple proteins through splicing were responsible for tumor transformation, we would not have seen that specificity of *U1*$^{g.3A>G}$ in MB or *U1*$^{g3.A>C}$ in CLL and HCC, but would expect to find both mutations in all three types of tumors. This suggests that what is important is not the overall deregulation of splicing, but the alteration of specific genes in each of the diseases. In this regard, just as in a tumor where driver and passenger mutations exist, mutations in *U1* will also generate driver and passenger aberrant splicing events. For example, in MB with the *U1*$^{g.3A>G}$ mutation, several driver genes were found deregulated for this tumor type (*PTCH1*, *GLI2*, *PAX5*, *CCND2*), whose alterations were not detected when we looked at RNA-seq data with the *U1*$^{g.3A>C}$ mutation. In addition to changes in splicing, the expression of multiple genes was also altered. Nevertheless, this is more an indirect effect of mutations in *U1*, since it is logical that dysregulation at the splicing level of so many proteins also translates into changes in the expression of many other genes. It would be necessary to carry out a more exhaustive analysis of the data to verify whether, additionally to any driver gene with altered splicing, there are others whose activation mechanism is due to direct changes in their expression.

To demonstrate that all these changes observed in patients were due to *U1* mutations, different cell lines were transfected or infected with *U1*$^{g.3A>G}$ or *U1*$^{g.3A>C}$ mutations, respectively. We saw the same change in the splicing pattern, as well as many deregulated genes. However, the percentage of altered genes that overlapped in both splicing and differential expression analyses between cell lines and patients was not very high. These results could be explained by the fact that there were not many patients with RNA-seq data, and they were not homogeneous, since each of them presented different driver mutations in addition to those in *U1*. This means that deregulated genes may appear or disappear due to the mutational background already present in each patient and not solely due to the effects of *U1*. Although the cell lines also exhibit other alterations, we use the same cells as controls in the experiments, whereas in the case of patients the controls are other cases with a very different mutational background.

In order to determine if the *U1*$^{g.3A>C}$ mutation could confer an advantage to cell lines over those with *U1*^{wt}, proliferation and apoptosis assays were performed. However, and similar to other mutations in splicing genes[179,180], no significant differences were found. There are several reasons to explain these results. First, we were using previously stablished cell lines that contained and depended on other driver events for their growth and survival. Therefore, introduction of a novel mutation, such as the *U1*$^{g.3A>C}$, might not result in an effect as the cell lines were already dependent on other pathways[181]. Second, it is possible that tumor growth depends on different stimuli while growing *in vitro vs in vivo*, and the *U1* mutation might be required only for certain processes not recapitulated *in vitro*. We were surprised not to observe differences in cell death when cells were treated with proteasome inhibitors. Thus, due to the large number of aberrant splicing transcripts formed in *U1*$^{g.3A>C}$ cells, it is very likely that the amount of truncated proteins to be degraded by the proteasome would be higher in *U1*$^{g.3A>C}$ cells when compared to control cells. Therefore, proteasome inhibition should generate some toxicity in the cells. However, the absence of differences suggests that some compensatory mechanisms might be acting. In this regard, when we compared differentially expressed genes caused by *U1*$^{g.3A>C}$, we observed that pathways related to protein synthesis and apoptosis were downregulated in CLL patients, suggesting that this compensatory mechanism could attenuate the expected toxicity generated by the inhibition of the proteasome. However, these changes were not observed in cell lines, suggesting that other, as yet unknown mechanism, might be responsible for this unexpected tolerance to proteasome inhibition.

Another interesting observation is that mutations in *U1* and *SF3B1* were mutually exclusive. As both genes are involved in the splicing process, recognizing the 5'SS and the 3'SS respectively, we explored the effect of inhibition of SF3B1 on cell survival in the presence or absence of the *U1*$^{g.3A>C}$ mutation. However, we did not observe major differences between *U1*-mutated and –wild-type cells in response to PladB, suggesting that mutations in *U1* and *SF3B1* might be mutually exclusive because they impact a similar pathway, and not because of synthetic lethality.

The main role of *U1* is the identification of the 5'SS, so it is logical to think that mutations in this gene will mainly affect splicing. However, it is known that this snRNA also has other functions on RNA transcriptional regulation by suppressing premature cleavage and polyadenylation[182,183]. Thus, mutations in *U1* could also be altering the expression of certain genes through these mechanisms. For example, the selection of premature cleavage sites may result in the degradation of some transcripts through the nonsense-mediated decay pathway[184]. On the other hand, the selection of alternative polyadenylation sites by the mutated *U1* can give rise to 3'UTRs of different lengths. Thus, in case of generating longer 3'UTRs, transcripts could present new motifs recognized by RNA biding proteins or miRNAs that might inhibit their expression, or on the contrary, these sites could disappear, increasing the stability of the transcripts[185]. Although the study of these processes would have improved our understanding of the mechanism of action of these mutations, the available data were not suitable for this type of analysis. Future experiments with other sequencing strategies could help us with this purpose.

Moreover, it has recently been described that *U1* also causes the arrest of some lncRNAs in chromatin, a process dependent on residues 3-10 of *U1* and associated with the number of U1 recognition sequences present in each lncRNA[165]. Based on this study, the presence of the g.3A>C mutation in *U1* could result in the recognition of additional lncRNAs that might be retained in the chromatin. However, we were unable to identify differences in lncRNA chromatin localization induced by $U1^{g.3A>C}$. The fact that the presence of a limited number of new U1 recognition sites may not be sufficient to cause chromatin retention, could explain why we did not detect any new chromatin-associated lncRNAs in cells with $U1^{g.3A>C}$.

One of the most interesting consequences of the $U1^{g.3A>C}$ mutation is the potential generation of novel antigens derived from its major alteration of splicing in numerous genes. Due to the relevance of tumor neoantigens in immunotherapy[186,187], *U1* mutations could constitute a relevant marker for this novel therapeutic approach, as despite *U1* being non-coding, it has an impact on many different proteins, and they should be similar in different patients. During this thesis we were able to perform a

preliminary approach by using serum from patients with *U1*^{wt} or *U1*$^{g.3A>C}$ against protein extracts from three cell lines (JVM3, HG3 and MEC1) either expressing *U1*^{wt} or *U1*$^{g.3A>C}$. The absence of patient antibodies specifically recognizing proteins in these cell lines indicates that either the amount of proteins with neoepitopes produced by these cell lines is too low to be detected by this technique, or that there is not a B-cell response to the potential neoantigens generated by these tumors. Further studies aimed at characterizing the T-cell response in these patients might provide additional information regarding the immune response generated by the *U1*$^{g.3A>C}$ mutation.

In addition to the mutations detected in *U1*, during this thesis we have also validated a series of mutations located in a highly conserved region of the 3'UTR of *NFKBIZ* in CLL. This hotspot was first identified by Dr. Gutiérrez-Abril after analysis of the 3'UTR regions in two cohorts of more than 1,000 patients with CLL[77]. More recently, Arthur et al. also detected mutations in this same region in cases with diffuse large B-cell lymphoma[188]. Because these mutations did not generate changes in transcript levels or mRNA stability, a more direct assay was necessary to see that the effect of, at least, the most recurrent mutation was the de-repression of translation by ribosomes[77]. Extension of these assays to six other mutations detected in this hotspot confirmed the same results, except for the g.10157830304G>A mutation, which neither altered mRNA stability nor led to increased protein expression. This difference might be due to the fact that while all the other mutations destroyed the SL1, this one was located in SL3 and it was not predicted to affect its structure (Figure 40). Similar results were obtained by CRISPR-Cas9-generated mutations, suggesting that alteration of SL1 within the 3'UTR of *NFKBIZ* is the target of those recurrent mutations detected in hematological tumors.

The fact that upon stimulation with IL1 there was an overexpression of IκBζ, with the same effect as mutations in its 3'UTR, made us explore in more detail the implication of MYD88 signaling pathway, as this is one of the most important pathways activated by interleukins. Furthermore, *MYD88* is one of the most frequent driver genes in CLL. Thus, a previous work has demonstrated that IRAK1 and TRAF6, two downstream members of MYD88 signaling pathway, were essential for such activation[189]. However, *IRAK1* is a kinase and *TRAF6* is an E3 ubiquitin ligase without RNA-binding domains, so they cannot

be the proteins that are ultimately blocking IκBζ translation. Therefore, other strategies, such as non-hypothesis driven experiments, should be designed in order to clarify which proteins participate directly in IκBζ repression. Together, this suggests that both, mutations in *MYD88* and in the 3'UTR of *NFKBIZ*, are acting through the same pathway. This can also be supported by the fact that mutations in *MYD88* and *NFKBIZ* are mutually exclusive in DLBCL[188], where *MYD88* is also considered a driver gene.

On the other hand, mutations described in CLL for the 3'UTR of *NFKBIZ* were detected in only one of the two cohorts comprising the study. Specifically, these mutations appeared in a series associated with patients in progression and/or participating in a clinical trial, which might imply that they were associated with worse prognosis or with a more advanced stage of the disease. This would make sense if we consider that the mechanism by which mutations in *MYD88* and *NFKBIZ* act are related, as it has been described that mutant *MYD88* is associated with unfavorable prognosis[190–192], although other studies show that it is of favorable prognosis, suggesting that its role is still unclear. This, together with the fact that mutations in this same region have been detected in DLBCL, may suggest that *NFKBIZ* is a gene involved in disease progression.

In summary, during this doctoral thesis we have contributed not only to increase our knowledge about the molecular mechanisms involved in cancer development by defining driver genes in a malignant lymphoma, such as MCL, or developing novel tools for the identification of somatic mutations in tumor genomes, but also to improve the diagnosis and management of patients with MB thanks to the use of liquid biopsies. In this regard, the generation of affordable pipelines for the analysis of tumor genomes with enough sensitivity and specificity fulfills a recurrent problem, by most medium size laboratories and most clinical centers, when attempting to use NGS strategies for diagnosis. Finally, we have attempted to characterize the molecular mechanism by which mutations in non-protein coding regions might contribute to tumorigenesis. A precise understanding of these processes might lead to the development of novel therapeutic approaches, including those based on immunotherapy, that could help some of these patients with bad prognosis mutations.

CONCLUSIONS/ CONCLUSIONES

1. Analysis of mantle cell lymphoma genomes has provided insight into the complex genomic landscape of these tumors, confirming that the main driving events of this disease are structural alterations, rather than the accumulation of point mutations as is the case in other hematologic tumors.

2. RFcaller is a fast and low computational demanding tool that allows the detection of somatic mutations in whole genomes with the same precision as that obtained with a combination of several of the most commonly used callers.

3. The sensitivity of RFcaller increases with tumor purity, and most of its false negatives correspond to subclonal mutations.

4. The analysis of circulating tumor DNA derived from cerebrospinal fluid in pediatric medulloblastoma tumors is a minimally invasive technique that allows characterization, monitoring and detection of minimal residual disease in these patients, as well as intratumor heterogeneity.

5. *U1* snRNA has been identified as a novel cancer driver gene, it is recurrently mutated in multiple tumor types and results in transcriptome-wide splicing changes.

6. Mutations at position 3 of U1 recognize new splicing donor sites, allowing the formation of cryptic 5' sites in the introns of multiple genes.

7. Chronic lymphocytic leukemia cell lines expressing the U1 mutant do not exhibit changes in growth, apoptosis, or resistance to proteasome or splicing inhibitory drugs when compared to wild-type cells.

8. Mutations in the 3'UTR of *NFKBIZ* destroy a highly conserved structure involved in translational repression, resulting in overexpression of the protein, but does not affect the stability of its mRNA.

9. Overexpression of *NFKBIZ* alters the expression of multiple genes, activating the NF-κB signaling pathway in HEK-293T cells.

10. Mutations in the 3'UTR of *NFKBIZ* mimic the effect of IL1 stimulation regarding translational repression of IκBζ, suggesting that these mutations act through the MYD88/IRAK1 pathway.

1. El análisis de genomas de linfoma de células del manto ha permitido comprender el complejo panorama genómico de estos tumores, confirmando que el principal evento conductor de esta enfermedad son las alteraciones estructurales, y no la acumulación de mutaciones puntuales como ocurre en otros tumores hematológicos.

2. RFcaller es una herramienta rápida y con bajos requerimientos computacionales que permite la detección de mutaciones somáticas en genomas completos, con la misma precisión que la obtenida con la combinación de varias de las herramientas más utilizadas.

3. La sensibilidad del RFcaller aumenta con la pureza del tumor, y la mayoría de los falsos negativos se corresponden con mutaciones subclonales.

4. El análisis del ADN tumoral circulante derivado del líquido cefalorraquídeo en tumores pediátricos de meduloblastoma constituye una técnica poco invasiva y que permite la caracterización, monitorización y detección de enfermedad mínima residual en estos pacientes, así como la heterogeneidad intratumoral.

5. Se ha identificado *U1* como nuevo gen conductor del cáncer, se encuentra recurrentemente mutado en múltiples tipos tumorales y provoca cambios de splicing a nivel de todo el transcriptoma.

6. Las mutaciones en la posición 3 de U1 provocan el reconocimiento de nuevos sitios donadores de splicing, lo que permite la formación de sitios 5' crípticos en los intrones de múltiples genes.

7. Las líneas celulares de leucemia linfática crónica que expresan la mutación g.3A>C de U1 no muestran alteraciones en el crecimiento, apoptosis ni resistencia a fármacos inhibidores del proteasoma o del splicing cuando se comparan con células control.

8. Las mutaciones en el 3'UTR de *NFKBIZ* destruyen una estructura altamente conservada implicada en la represión de la traducción, lo que da lugar a la sobreproducción de la proteína sin afectar a la estabilidad de su mRNA.

9. La sobreexpresión de *NFKBIZ* altera la expresión de múltiples genes, activando la ruta de señalización de NF-κB en células HEK-293T.

10. Las mutaciones en el 3'UTR de *NFKBIZ* provocan un efecto similar al ejercido por la estimulación con IL1 en cuanto a la represión traduccional de IκBζ, sugiriendo que estas mutaciones actúan a través de la vía de MYD88/IRAK1.

BIBLIOGRAPHY

1. Der, R., Epstein, C. L. & Plotkin, J. B. Generalized population models and the nature of genetic drift. *Theor. Popul. Biol.* **80**, 80–99 (2011).
2. Tigano, A. & Friesen, V. L. Genomics of local adaptation with gene flow. *Mol. Ecol.* **25**, 2144–2164 (2016).
3. Darwin, C., 1809-1882. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. (London : John Murray, 1859, 1859).
4. Darwin, C. & Wallace, A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Zool. J. Linn. Soc.* **3**, 45–62 (1858).
5. Clancy, S. Genetic mutation. *Nat. Educ. Knowl.* **1**, 187 (2008).
6. Carlin, J. L. Mutations Are the Raw Materials of Evolution. *Nat. Educ. Knowl.* **3**, 10 (2011).
7. Meyerson, W., Leisman, J., Navarro, F. C. P. & Gerstein, M. Origins and characterization of variants shared between databases of somatic and germline human mutations. *BMC Bioinformatics* **21**, 227 (2020).
8. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
9. Arana, M. E. & Kunkel, T. A. Mutator phenotypes due to DNA replication infidelity. *Semin. Cancer Biol.* **20**, 304–311 (2010).
10. Bębenek, A. & Ziuzia-Graczyk, I. Fidelity of DNA replication-a matter of proofreading. *Curr. Genet.* **64**, 985–996 (2018).
11. Ohno, M. Spontaneous de novo germline mutations in humans and mice: rates, spectra, causes and consequences. *Genes Genet. Syst.* **94**, 13–22 (2019).
12. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
13. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).
14. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
15. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
16. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
17. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
18. Blagosklonny, M. V. Molecular theory of cancer. *Cancer Biol. Ther.* **4**, 621–627 (2005).
19. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
20. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
21. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
22. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
23. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
24. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
25. Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* **10**, 25–50 (2015).

26. Shen, L., Shi, Q. & Wang, W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis* **7**, 25 (2018).

27. Yang, L., Han, Y., Suarez Saiz, F., Saurez Saiz, F. & Minden, M. D. A tumor suppressor and oncogene: the WT1 story. *Leukemia* **21**, 868–876 (2007).

28. Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269–271 (2004).

29. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

30. Nicolas, M. *et al.* Notch1 functions as a tumor suppressor in mouse skin. *Nat. Genet.* **33**, 416–421 (2003).

31. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154–1157 (2011).

32. Seoane, J. & Gomis, R. R. TGF-β Family Signaling in Tumor Suppression and Cancer Progression. *Cold Spring Harb. Perspect. Biol.* **9**, a022277 (2017).

33. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

34. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

35. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

36. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

37. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

38. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

39. Chen, X. *et al.* Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **335**, 1235–1238 (2012).

40. Li, C. & Luscombe, N. M. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat. Commun.* **11**, 1363 (2020).

41. Li, F. *et al.* The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell* **153**, 590–600 (2013).

42. Monroe, J. G. *et al.* Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature* **602**, 101–105 (2022).

43. Futuyma, D. J. *Evolutionary biology*. (Sinauer Associates, 1986).

44. 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).

45. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

46. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinforma. Oxf. Engl.* **27**, 175–181 (2011).

47. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

48. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

49. Evans, P., Avey, S., Kong, Y. & Krauthammer, M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. *IEEE Trans. Nanobioscience* **12**, 150–157 (2013).

50. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5486-5495 (2015).

51. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).

52. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).

53. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

54. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu & Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341.e23 (2017).

55. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

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

57. Kim, S., Kim, K., Choe, J., Lee, I. & Kang, J. Improved survival analysis by learning shared genomic information from pan-cancer data. *Bioinforma. Oxf. Engl.* **36**, i389–i398 (2020).

58. Nagy, Á., Munkácsy, G. & Győrffy, B. Pancancer survival analysis of cancer hallmark genes. *Sci. Rep.* **11**, 6047 (2021).

59. The Cancer Target Discovery and Development Network. Transforming Big Data into Cancer-Relevant Insight: An Initial, Multi-Tier Approach to Assess Reproducibility and Relevance. *Mol. Cancer Res.* **14**, 675–682 (2016).

60. Hahn, W. C. *et al.* An expanded universe of cancer targets. *Cell* **184**, 1142–1155 (2021).

61. Braun, T. P., Eide, C. A. & Druker, B. J. Response and Resistance to BCR-ABL1-Targeted Therapies. *Cancer Cell* **37**, 530–542 (2020).

62. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288 (2019).

63. Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).

64. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

65. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

66. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).

67. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).

68. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

69. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).

70. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

71. Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget* **6**, 35922–35930 (2015).

72. Wang, L. *et al.* Silencing of diphthamide synthesis 3 (Dph3) reduces metastasis of murine melanoma. *PloS One* **7**, e49988 (2012).

73. Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).

74. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041 (2003).

75. Zhang, W. *et al.* A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).

76. Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).

77. Gutierrez-Abril, J. Identification of novel genes and molecular mechanisms involved in tumorigenesis through the use of next generation sequencing technologies. (Universidad de Oviedo, 2018).

78. PCAWG Transcriptome Core Group *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

79. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).

80. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15**, 689–701 (2014).

81. Pagani, F. New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum. Mol. Genet.* **12**, 1111–1120 (2003).

82. Aretz, S. *et al.* Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum. Mutat.* **24**, 370–380 (2004).

83. Moulson, C. L. *et al.* Increased progerin expression associated with unusual LMNA mutations causes severe progeroid syndromes. *Hum. Mutat.* **28**, 882–889 (2007).

84. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

85. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

86. Wai, H. A. *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 1005–1014 (2020).

87. Qian, X. *et al.* Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. *Front. Genet.* **12**, 647400 (2021).

88. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).

89. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinforma.* **56**, (2016).

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

91. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).

92. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).

93. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

94. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).

95. 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.* **6**, 36540 (2016).

96. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

97. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16**, 197 (2015).

98. Ainscough, B. J. *et al.* A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.* **50**, 1735–1743 (2018).

99. Anzar, I., Sverchkova, A., Stratford, R. & Clancy, T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics* **12**, 63 (2019).

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

101. Spinella, J.-F. *et al.* SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics* **17**, 912 (2016).

102. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

103. Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).

104. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).

105. Target Guide Sequence Cloning Protocol. https://media.addgene.org/data/plasmids/52/52961/52961-attachment_B3xTwla0bkYD.pdf.

106. Mayer, A. & Churchman, L. S. A Detailed Protocol for Subcellular RNA Sequencing (subRNA-seq). *Curr. Protoc. Mol. Biol.* **120**, 4.29.1-4.29.18 (2017).

107. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168–e168 (2014).

108. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/ARXIV.1303.3997.

109. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).

110. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinforma. Oxf. Engl.* **30**, 2503–2505 (2014).

111. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

112. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

113. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinforma. Oxf. Engl.* **28**, 2184–2185 (2012).

114. Danecek, P., Schiffels, S. & Durbin, R. Multiallelic calling model in bcftools (-m). (2016).

115. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

116. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

117. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinforma. Oxf. Engl.* **35**, 421–432 (2019).

118. Valdés-Mas, R., Bea, S., Puente, D. A., López-Otín, C. & Puente, X. S. Estimation of copy number alterations from exome sequencing data. *PloS One* **7**, e51422 (2012).

119. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

120. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).

121. Beà, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18250–18255 (2013).

122. Wang, S., Tao, Z., Wu, T. & Liu, X.-S. Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis. *Bioinformatics* **37**, 1590–1592 (2021).

123. Anders, S., Pyl, P. T. & Huber, W. *HTSeq - A Python framework to work with high-throughput sequencing data*. http://biorxiv.org/lookup/doi/10.1101/002824 (2014) doi:10.1101/002824.

124. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

125. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

126. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

127. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).

128. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).

129. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).

130. Le, T. T., Fu, W. & Moore, J. H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* **36**, 250–256 (2020).

131. López-Ratón, M., Rodríguez-Álvarez, M. X., Suárez, C. C. & Sampedro, F. G. OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *J. Stat. Softw.* **61**, (2014).

132. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinforma. Oxf. Engl.* **21**, 3940–3941 (2005).

133. Nadeu, F. *et al.* Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–2130 (2016).

134. Nadeu, F. *et al.* Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia* **32**, 645–653 (2018).

135. Campo, E. & Rule, S. Mantle cell lymphoma: evolving management strategies. *Blood* **125**, 48–55 (2015).

136. Weltgesundheitsorganisation. *WHO classification of tumours of haematopoietic and lymphoid tissues*. (International Agency for Research on Cancer, 2017).

137. Martin, P. *et al.* Outcome of deferred initial therapy in mantle-cell lymphoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27**, 1209–1213 (2009).

138. Fernàndez, V. *et al.* Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma. *Cancer Res.* **70**, 1408–1418 (2010).

139. Puente, X. S., Jares, P. & Campo, E. Chronic lymphocytic leukemia and mantle cell lymphoma: crossroads of genetic and microenvironment interactions. *Blood* **131**, 2283–2296 (2018).

140. Navarro, A., Beà, S., Jares, P. & Campo, E. Molecular Pathogenesis of Mantle Cell Lymphoma. *Hematol. Oncol. Clin. North Am.* **34**, 795–807 (2020).

141. Wu, C. *et al.* Genetic heterogeneity in primary and relapsed mantle cell lymphomas: Impact of recurrent CARD11 mutations. *Oncotarget* **7**, 38180–38190 (2016).

142. Zhang, J. *et al.* The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells. *Blood* **123**, 2988–2996 (2014).

143. Bohlander, S. K., Kakadiya, P. M. & Coysh, A. Chromosome Rearrangements and Translocations. in *Encyclopedia of Cancer (Third Edition)* (eds. Boffetta, P. & Hainaut, P.) 389–404 (Academic Press, 2019). doi:https://doi.org/10.1016/B978-0-12-801238-3.65105-X.

144. Nadeu, F. *et al.* Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. *Blood* **136**, 1419–1432 (2020).

145. Schwalbe, E. C. *et al.* Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. *Lancet Oncol.* **18**, 958–971 (2017).

146. Hovestadt, V. *et al.* Medulloblastomics revisited: biological and clinical insights from thousands of patients. *Nat. Rev. Cancer* **20**, 42–56 (2020).

147. Ramaswamy, V. *et al.* Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta Neuropathol. (Berl.)* **131**, 821–831 (2016).

148. De Mattos-Arruda, L. *et al.* Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat. Commun.* **6**, 8839 (2015).

149. Pentsova, E. I. *et al.* Evaluating Cancer of the Central Nervous System Through Next-Generation Sequencing of Cerebrospinal Fluid. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **34**, 2404–2415 (2016).

150. Wang, Y. *et al.* Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 9704–9709 (2015).

151. Seoane, J., De Mattos-Arruda, L., Le Rhun, E., Bardelli, A. & Weller, M. Cerebrospinal fluid cell-free tumour DNA as a liquid biopsy for primary brain tumours and central nervous system metastases. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **30**, 211–218 (2019).

152. Toguchida, J. *et al.* Prevalence and spectrum of germline mutations of the p53 gene among patients with sarcoma. *N. Engl. J. Med.* **326**, 1301–1308 (1992).

153. Denison, R. A., Van Arsdell, S. W., Bernstein, L. B. & Weiner, A. M. Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 810–814 (1981).

154. Manser, T. & Gesteland, R. F. Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. *Cell* **29**, 257–264 (1982).

155. Nadeu, F., Diaz-Navarro, A., Delgado, J., Puente, X. S. & Campo, E. Genomic and Epigenomic Alterations in Chronic Lymphocytic Leukemia. *Annu. Rev. Pathol.* **15**, 149–177 (2020).

156. Wang, L. *et al.* Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell* **30**, 750–763 (2016).

157. Kogerman, P. *et al.* Alternative first exons of PTCH1 are differentially regulated in vivo and may confer different functions to the PTCH1 protein. *Oncogene* **21**, 6007–6016 (2002).

158. Sasaki, H., Nishizaki, Y., Hui, C., Nakafuku, M. & Kondoh, H. Regulation of Gli2 and Gli3 activities by an amino-terminal repression domain: implication of Gli2 and Gli3 as primary mediators of Shh signaling. *Dev. Camb. Engl.* **126**, 3915–3924 (1999).

159. Mirzaa, G. *et al.* De novo CCND2 mutations leading to stabilization of cyclin D2 cause megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome. *Nat. Genet.* **46**, 510–515 (2014).

160. Fedorchenko, O. *et al.* CD44 regulates the apoptotic response and promotes disease development in chronic lymphocytic leukemia. *Blood* **121**, 4126–4136 (2013).

161. Herishanu, Y., Gibellini, F., Njuguna, N., Keyvanfar, K. & Wiestner, A. CD44 Signaling Via PI3K/AKT and MAPK/ERK Pathways Protects CLL Cells from Spontaneous and Drug Induced Apoptosis. *Blood* **112**, 541–541 (2008).

162. Seiler, M. *et al.* Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep.* **23**, 282-296.e4 (2018).

163. Murphy, J. T. *et al.* Functional elements of the human U1 RNA promoter. Identification of five separate regions required for efficient transcription and template competition. *J. Biol. Chem.* **262**, 1795–1803 (1987).

164. Gunderson, S. I., Knuth, M. W. & Burgess, R. R. The human U1 snRNA promoter correctly initiates transcription in vitro and is activated by PSE1. *Genes Dev.* **4**, 2048–2060 (1990).

165. Yin, Y. *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, 147–150 (2020).

166. Behrens, G. *et al.* A translational silencing function of MCPIP1/Regnase-1 specified by the target site context. *Nucleic Acids Res.* **46**, 4256–4270 (2018).

167. Essig, K. *et al.* Roquin targets mRNAs in a 3'-UTR-specific manner by different modes of regulation. *Nat. Commun.* **9**, 3810 (2018).

168. Mino, T. *et al.* Regnase-1 and Roquin Regulate a Common Element in Inflammatory mRNAs by Spatiotemporally Distinct Mechanisms. *Cell* **161**, 1058–1073 (2015).

169. Van Etten, J. *et al.* Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *J. Biol. Chem.* **287**, 36370–36383 (2012).

170. Cao, J., Mu, Q. & Huang, H. The Roles of Insulin-Like Growth Factor 2 mRNA-Binding Protein 2 in Cancer and Cancer Stem Cells. *Stem Cells Int.* **2018**, 4217259 (2018).

171. Oliveira-Mateos, C. *et al.* The transcribed pseudogene RPSAP52 enhances the oncofetal HMGA2-IGF2BP2-RAS axis through LIN28B-dependent and independent let-7 inhibition. *Nat. Commun.* **10**, 3979 (2019).

172. Gladden, A. B., Woolery, R., Aggarwal, P., Wasik, M. A. & Diehl, J. A. Expression of constitutively nuclear cyclin D1 in murine lymphocytes induces B-cell lymphoma. *Oncogene* **25**, 998–1007 (2006).

173. Garcia-Prieto, C. A., Martínez-Jiménez, F., Valencia, A. & Porta-Pardo, E. Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. *Bioinformatics* btac306 (2022) doi:10.1093/bioinformatics/btac306.

174. Bourdeaut, F. & Delattre, O. Genetic predisposition to medulloblastomas: just follow the tumour genome. *Lancet Oncol.* **19**, 722–723 (2018).

175. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol. Syndromol.* **8**, 4–23 (2017).

176. Waszak, S. M. *et al.* Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort. *Lancet Oncol.* **19**, 785–798 (2018).

177. Bousquets-Muñoz, P. *et al.* PanCancer analysis of somatic mutations in repetitive regions reveals recurrent mutations in snRNA U2. *Npj Genomic Med.* **7**, 19 (2022).

178. Kondo, Y., Oubridge, C., van Roon, A.-M. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *eLife* **4**, (2015).

179. Zhou, Q. *et al.* A chemical genetics approach for the functional assessment of novel cancer genes. *Cancer Res.* **75**, 1949–1958 (2015).

180. Fei, D. L. *et al.* Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival. *PLoS Genet.* **12**, e1006384 (2016).

181. Levine, A. J., Jenkins, N. A. & Copeland, N. G. The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. *Cancer Cell* **35**, 10–15 (2019).

182. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).

183. Oh, J.-M. *et al.* U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat. Struct. Mol. Biol.* **24**, 993–999 (2017).

184. Nickless, A., Bailis, J. M. & You, Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell Biosci.* **7**, 26 (2017).

185. So, B. R. *et al.* A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. *Mol. Cell* **76**, 590-599.e4 (2019).

186. Biernacki, M. A. & Bleakley, M. Neoantigens in Hematologic Malignancies. *Front. Immunol.* **11**, 121 (2020).

187. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).

188. Arthur, S. E. *et al. Non-coding* NFKBIZ *3' UTR mutations promote cell growth and resistance to targeted therapeutics in diffuse large B-cell lymphoma*. http://biorxiv.org/lookup/doi/10.1101/2021.05.22.445261 (2021) doi:10.1101/2021.05.22.445261.

189. Ohba, T. *et al.* Identification of interleukin-1 receptor-associated kinase 1 as a critical component that induces post-transcriptional activation of IκB-ζ: Post-transcriptional activation of IκB-ζ by IRAK1. *FEBS J.* **279**, 211–222 (2012).

190. Baliakas, P. *et al.* Prognostic relevance of MYD88 mutations in CLL: the jury is still out. *Blood* **126**, 1043–1044 (2015).

191. Martínez-Trillos, A. *et al.* Clinical impact of MYD88 mutations in chronic lymphocytic leukemia. *Blood* **127**, 1611–1613 (2016).

192. Qin, S.-C. *et al.* MYD88 mutations predict unfavorable prognosis in Chronic Lymphocytic Leukemia patients with mutated IGHV gene. *Blood Cancer J.* **7**, 651 (2017).

PUBLICATIONS

List of publications related to the Thesis and my contribution

- Shuai S, Suzuki H, **Diaz-Navarro A**, Nadeu F, Kumar SA, Gutierrez-Fernandez A, et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature. 2019 Oct;574(7780):712–6. My contribution was:
    - Generate lentiviral vectors and cell models expressing $U1^{wt}$ or $U1^{g.3A>C}$ in three CLL cell lines (JVM3, HG3 and MEC1).
    - Confirm the expression of mutant *U1* by 5'RACE and validate its effect on splicing through the amplification of three transcripts with aberrant splicing (*ABCD3*, *MSI2* and *PODL1*).
    - Perform RNA-seq to confirm causal effect of the $U1^{g.3A>C}$ mutation on splicing.
- Suzuki H, Kumar SA, Shuai S, **Diaz-Navarro A**, Gutierrez-Fernandez A, De Antonellis P, et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. Nature. 2019 Oct;574(7780):707–11. My contribution during this project was:
    - Carry out experimental validations of the $U1^{g.3A>G}$ mutation in HEK-293T by transient transfection.
    - Confirm the expression of mutant *U1* by 5'RACE and validate its effect on splicing through the amplification of two transcripts with aberrant splicing (*PTCH1* and *GLI2*).
    - Perform RNA-seq to confirm causal effect of the $U1^{g.3A>G}$ mutation on splicing.
- Escudero L, Llort A, Arias A, **Diaz-Navarro A**, Martínez-Ricarte F, Rubio-Perez C, et al. Circulating tumour DNA from the cerebrospinal fluid allows the characterisation and monitoring of medulloblastoma. Nat Commun. 2020 Oct 27;11(1):5376. In this work my contribution was:
    - Exome analysis, from tumor and cerebrospinal fluid, of 13 patients with medulloblastoma.
    - Search for both somatic and germline mutations and copy number alterations.
    - Assist with the molecular classification of these patients, using previously generated data.

- Nadeu F, Martin-Garcia D, Clot G, **Díaz-Navarro A**, Duran-Ferrer M, Navarro A, et al. Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. Blood. 2020 Sep 17;136(12):1419–32. In this project my task was:
    - The analysis of somatic mutations, structural variations and copy number alterations to identify novel genes or driver events of MCL, using WGS from 61 patients.
- **Díaz-Navarro A**, Bousquets-Muñoz P, Nadeu F, López-Tamargo S, Beà S, Campo E, et al. RFcaller: a machine learning approach combined with read-level features to detect somatic mutations. bioRxiv. 2022. The steps I took to develop RFcaller were:
    - Design and development of the model to detect SSNVs and indels.
    - Review the somatic mutation datasets to obtain a database with curated and reliable mutations.
    - Train the machine learning algorithms and set the cut-off points.
    - Perform the comparison with the results published by the PCAWG for mutations detected in CLL and BRCA.
    - Test the performance of the tool with whole exomes.

Other papers published during the Thesis

- Álvarez-Eguiluz Á, **Díaz-Navarro A**, Puente XS. Dissecting Degradomes: Analysis of Protease-Coding Genes. Methods Mol Biol. 2018;1731:1–13.
- Ruiz de Garibay G, Mateo F, Stradella A, Valdés-Mas R, Palomero L, Serra-Musach J, et al. Tumor xenograft modeling identifies an association between TCF4 loss and breast cancer chemoresistance. Dis Model Mech. 2018 May 18;11(5):dmm032292.
- Gómez-Miragaya J, **Díaz-Navarro A**, Tonda R, Beltran S, Palomero L, Palafox M, et al. Chromosome 12p Amplification in Triple-Negative/BRCA1-Mutated Breast Cancer Associates with Emergence of Docetaxel Resistance and Carboplatin Sensitivity. Cancer Res. 2019 Aug 15;79(16):4258–70.
- Nadeu F, **Diaz-Navarro A**, Delgado J, Puente XS, Campo E. Genomic and Epigenomic Alterations in Chronic Lymphocytic Leukemia. Annu Rev Pathol. 2020 Jan 24;15:149–77.
- Bousquets-Muñoz P, **Díaz-Navarro A**, Nadeu F, Sánchez-Pitiot A, López-Tamargo S, Shuai S, et al. PanCancer analysis of somatic mutations in repetitive regions reveals recurrent mutations in snRNA U2. npj Genom Med. 2022 Dec;7(1):19.
- Kontio J, Soñora VR, Pesola V, Lamba R, Dittmann A, **Navarro AD**, et al. Analysis of extracellular matrix network dynamics in cancer using the MatriNet database. Matrix Biology. 2022 Jun;110:141–50.