UNIVERSIDAD
DE GRANADA

Bioinformatic approaches for the discovery of non-coding alterations in cancer

PhD Thesis. Programa de Doctorado en Bioquímica y Biología Molecular (B16.56.1). Universidad de Granada.

PhD candidate:

Álvaro Andrades Delgado

Supervisors:

Prof. Pedro Pablo Medina Vico

Prof. Marta Eugenia Cuadros Celorrio

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alvaro Andrades Delgado
ISBN: 978-84-1117-888-4
URI: https://hdl.handle.net/10481/82466

Abstract

Introduction. Cancer is one of the main causes of premature death worldwide. Cancers arise when cell genomes accumulate driver mutations, which are mutations that improve cell fitness. Driver mutations are a minority among the thousands of mutations present in a typical cancer genome. Although major efforts have been made to identify driver mutations in various cancers, most of them have focused on the protein coding genome, which only represents ~1.1% of the human genome. Part of the ~98.9% of the human genome that does not code for protein contains functional elements, such as regulatory DNA elements, intronic splice regions, untranslated regions of protein coding genes, and non-coding RNA genes. Among non-coding RNAs, microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) may participate in the regulation of gene expression and their expression is often altered in cancer. However, efforts to identify non-coding driver mutations have been rare, and sample sizes in lung adenocarcinoma (LUAD) have been low. In addition, the largest mutational study in diffuse large B-cell lymphoma (DLBCL) to date omitted mutations in intronic splice regions.

Objectives. We aimed to computationally identify and characterize novel driver mutations in non-coding DNA in in-house and external LUAD cohorts, with special focus on miRNA genes, lncRNA genes, and intronic splice regions. In addition, we aimed to identify and characterize previously missed mutations in intronic splice sites in external DLBCL datasets.

Methods. We performed targeted sequencing of genomic DNA in an in-house cohort of 70 LUAD primary tumors, 27 matched normal samples, and 37 LUAD cell lines. Our design included all human miRNA genes (n = 1881), as well as exons of cancer-related lncRNA genes (n = 908) and protein coding genes (n = 1307). We developed computational pipelines to identify high-confidence somatic variants by combining multiple variant calling tools, and we also applied them to external whole-genome sequencing data of LUAD samples from The Cancer Genome Atlas (N = 59 tumor-normal pairs). In addition, we applied state-of-the-art driver discovery tools to find putative

drivers in coding sequences, lncRNAs, miRNAs, intronic splice regions, proximal promoters, and untranslated regions. We assessed the functional relevance of the identified candidate drivers using external genomic, gene expression, and clinical data as well as functional impact scores. Furthermore, we developed a novel pipeline to annotate variants in a miRNA-centric manner, identifying variants that affect seeds and those that disrupt or create sequence motifs that mediate the processing of miRNA primary transcripts. Finally, in DLBCL, we reanalyzed external datasets (combined N = 1711) to identify previously missed recurrent mutations at intronic splice sites, we analyzed the impact of the splice site mutations on RNA processing, and we functionally characterized the most recurrent splice site mutation, which affected *CD79B*.

Results. We successfully detected high-confidence somatic variants in all analyzed datasets. However, driver discovery tools did not perform adequately in our targeted sequencing cohorts of limited size, as one based on functional impact predictions lacked sensitivity in non-coding regions whereas one based on mutation clustering had a high false positive rate. Still, we identified three candidate driver lncRNAs that accumulated mutations so that at least one mutation had high predicted functional impact: *TUSC7*, *SOX2-OT*, and *ZEB2-AS1*. However, the affected lncRNAs had very low expression in external LUAD datasets. This, together with their mutational patterns and the genomic characteristics of their loci, argued against an RNA sequence-dependent effect of their mutations. In miRNAs, a mutation in the seed of miR-133b was predicted to have high functional impact and to prevent it from targeting the oncogene *EGFR*. In addition, we identified mutations that disrupted or created processing motifs in miRNA primary transcripts, such as mutations that disrupted mismatched GHG motifs in mir-7-1, mir-7-2, and mir-139. In intronic splice regions, we found mutations that altered RNA splicing in LUAD driver genes such as *MET* and *RBM10*. In promoters and untranslated regions, we detected no high-confidence drivers. In DLBCL, intronic splice site mutations recurrently affected cancer driver genes and caused major RNA aberrations in *cis*. The most recurrent RNA alteration was intron 4 retention in *CD79B*. The alteration was caused by recurrent mutations at the fourth splice donor site of *CD79B*, and it was associated with

an increase in the number of B cell receptors in the cell surface and a subsequent increase in oncogenic signaling.

Conclusions. Non-coding variants with high predicted functional impact were rare in our LUAD datasets. In addition, it was unclear whether the candidate driver non-coding RNAs in LUAD had RNA sequence-dependent functions. Experimental work will be necessary to confirm whether the candidate driver non-coding RNAs have biological activity in LUAD and whether their activity is altered by the observed mutations. In DLBCL, intronic splice site mutations are recurrent and they can cause major cancer-promoting aberrations in driver genes.

Resumen

Introducción. El cáncer es una de las principales causas de muerte prematura mundialmente. El cáncer se origina cuando los genomas celulares acumulan mutaciones conductoras, que son mutaciones que confieren ventaja selectiva a la célula. Las mutaciones conductoras son una minoría entre las miles de mutaciones que contiene un genoma tumoral promedio. Aunque se han llevado a cabo grandes esfuerzos para identificar mutaciones conductoras en una gran variedad de cánceres, la mayoría de los esfuerzos se han centrado en el genoma codificante de proteína, que tan solo supone ~1,1% del genoma humano. Parte del ~98,9% del genoma humano no codificante de proteína contiene elementos funcionales, tales como ADN regulador, regiones intrónicas de corte y empalme, regiones no traducidas de genes codificantes de proteína y genes de ARN no codificante. Entre los ARNs no codificantes, los microARNs (miARNs) y los ARNs largos no codificantes (ARNlncs) pueden participar en la regulación de la expresión génica y su expresión está a menudo alterada en cáncer. Sin embargo, los esfuerzos para identificar mutaciones conductoras en secuencias no codificantes han sido escasos, y los tamaños de muestra para adenocarcinoma de pulmón (ADC) han sido bajos. Además, el mayor estudio hasta la fecha en linfoma difuso de células B grandes (LDCBG) omitió las mutaciones en regiones intrónicas de corte y empalme.

Objetivos. Nuestro objetivo principal fue analizar y caracterizar computacionalmente nuevas mutaciones conductoras en secuencias no codificantes en cohortes propias y externas de ADC, con especial interés en miARNs, ARNlncs y regiones intrónicas de corte y empalme. Un objetivo adicional fue identificar y caracterizar mutaciones previamente no descritas en sitios intrónicos de corte y empalme en datos externos de LDCBG.

Métodos. Realizamos secuenciación de DNA genómico dirigida a todos los genes de miARNs humanos (n = 1881), así como a exones de genes de ARNlncs relacionados con cáncer (n = 908) y de genes codificantes de proteína relacionados con cáncer (n = 1307) en una cohorte propia de 70 tumores primarios de LUAD, 27 muestras normales pareadas y 37 líneas celulares de LUAD. Desarrollamos métodos computacionales para identificar variantes

somáticas con alta confianza mediante la combinación de múltiples herramientas. Además, aplicamos dichos métodos para analizar datos de secuenciación de genoma completo de muestras de ADC de The Cancer Genome Atlas (N = 59 parejas tumor-normal). Asimismo, aplicamos herramientas de descubrimiento de mutaciones conductoras en secuencias codificantes, ARNlncs, miARNs, regiones intrónicas de corte y empalme, promotores proximales y regiones no traducidas. Determinamos la relevancia funcional de los elementos candidatos a conductores utilizando datos externos genómicos, transcriptómicos y clínicos. Además, desarrollamos una nueva metodología para anotar variantes de una forma miARN-céntrica, pudiendo identificar variantes que afectan a secuencias semilla y aquellas que crean o destruyen motivos de secuencia que median el procesamiento de los transcritos primarios de miARNs. Finalmente, en LDCBG, reanalizamos conjuntos de datos externos (N combinada = 1711) para identificar mutaciones recurrentes en sitios intrónicos de corte y empalme no detectadas en estudios anteriores. Analizamos el impacto de las mutaciones halladas en el procesamiento del ARN afectado, y caracterizamos funcionalmente la mutación más recurrente, que afectaba a *CD79B*.

Resultados. Detectamos exitosamente variantes somáticas con alta confianza en todos los conjuntos de datos analizados. Sin embargo, las herramientas de descubrimiento de mutaciones conductoras no tuvieron un rendimiento adecuado en nuestras cohortes de secuenciación dirigida de tamaño limitado: una herramienta basada en predicciones de impacto funcional tuvo baja sensibilidad en regiones no codificantes, mientras que otra basada en el agrupamiento de mutaciones tuvo una tasa elevada de falsos positivos. No obstante, identificamos tres ARNlncs candidatos a conductores que acumulaban mutaciones tal que al menos una de ellas tenía un alto impacto funcional predicho: *TUSC7*, *SOX2-OT* y *ZEB2-AS1*. Sin embargo, los ARNlncs afectados tenían una expresión extremadamente baja en datos externos de ADC. Esto, unido a sus patrones mutacionales y a las características genómicas de sus *loci*, hizo improbable que el efecto de sus mutaciones fuese dependiente de la secuencia de ARN. En miARNs, una mutación en la semilla de miR-133b tenía un alto impacto funcional predicho, impidiendo la unión de miR-133b al oncogén *EGFR*. Asimismo, identificamos mutaciones que destruían o

creaban motivos de procesamiento en los transcritos primarios de miARNs, destacando las mutaciones que afectaban a motivos GHG desapareados en mir-7-1, mir-7-2 y mir-139. En regiones intrónicas de corte y empalme, hallamos mutaciones que alteraban el corte y empalme de genes conductores de ADC como *MET* y *RBM10*. En promotores y en regiones no traducidas, no encontramos ninguna mutación conductora con alto nivel de confianza. En LDCBG, las mutaciones en sitios intrónicos de corte y empalme afectaban recurrentemente a genes conductores de la enfermedad y causaban grandes aberraciones a nivel de ARN en *cis*. La aberración más recurrente a nivel de ARN fue la retención del intrón 4 de *CD79B*. La alteración estaba causada por mutaciones recurrentes en el cuarto sitio intrónico donador de corte y empalme de *CD79B*, y estaba asociada a un incremento en el número de receptores de células B en la superficie celular y un consiguiente aumento en la señalización oncogénica.

Conclusiones. Las variantes en ARNs no codificantes con impacto funcional predicho elevado fueron infrecuentes en nuestros conjuntos de datos. Asimismo, no se pudo determinar de manera concluyente que los ARNs no codificantes candidatos a conductores de ADC tuviesen funciones dependientes de la secuencia de ARN. Se requerirá trabajo experimental para confirmar si los ARNs no codificantes candidatos a conductores tienen actividad biológica en ADC y si las mutaciones detectadas en los mismos alteran dicha actividad. En LDCBG, las mutaciones en sitios intrónicos de corte y empalme son recurrentes y pueden causar grandes aberraciones en los principales genes conductores de la enfermedad.

Table of contents

Figure index

Table index

List of abbreviations

Abbreviation	Meaning
ABC	Activated B cell-like
AID	Activation induced-cytidine deaminase
BAM	Binary alignment map
BCR	B cell receptor
BMR	Background mutation rate
CADD	Combined Annotation–Dependent Depletion
CCLE	Cancer Cell Line Encyclopedia
CDS	Coding sequence
ceRNA	Competing endogenous RNA
CI	Confidence interval
CPM	Counts per million
dsRNA	Double-stranded RNA
DLBCL	Diffuse large B-cell lymphoma
FACS	Fluorescence-activated cell sorting
FATHMM-MKL	Functional Analysis Through Hidden Markov Models and Multiple Kernel Learning
FC	Fold change
FDR	False discovery rate
FPKM	Fragments per kilobase of exon per million mapped fragments
GCB	Germinal center B cell-like
GDC	Genomic Data Commons

Abbreviation	Meaning
GTEx	Genome Tissue Expression
ICGC	International Cancer Genome Consortium
Ig	Immunoglobulin
ITAM	Immunoreceptor tyrosine-based activation motif
lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
MAF	Mutation annotation file
miRNA	MicroRNA
MNV	Multi-nucleotide variant
ncRNA	Non-coding RNA
NSCLC	Non-small cell lung cancer
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCG	Protein-coding gene
PSI	Percentage spliced in
qPCR	Quantitative polymerase chain reaction
RISC	RNA-induced silencing complex
SNP	Single nucleotide polymorphism
ssRNA	Single-stranded RNA
TCGA	The Cancer Genome Atlas
TPM	Transcripts per million
UTR	Untranslated region
VAF	Variant allele frequency
VCF	Variant call format

Abbreviation	Meaning
WES	Whole exome sequencing
WGS	Whole genome sequencing

Comments on notation and terminology

Throughout this text, gene symbols, DNA elements, and RNAs are italicized (e.g., *KRAS* gene, *KRAS* promoter, *KRAS* mRNA). Protein names are not italicized (e.g., KRAS protein).

For microRNAs (miRNAs), miRBase nomenclature has been followed (https://www.mirbase.org/help/nomenclature.shtml). Precursor miRNAs are not capitalized (e.g., hsa-mir-133b), whereas mature miRNAs are written with a capitalized R (e.g., hsa-miR-133b). For convenience, because this text refers exclusively to humans, the "hsa-" prefix has been omitted in most cases.

To describe differences between a DNA sequence of interest and a reference sequence, the terms "variant" and "mutation" can be used almost interchangeably. However, due to its neutral connotation, "variant" is preferred when describing any type of sequence difference regardless of its functional significance, and "mutation" is reserved for variants detected in a tumor that may have functional impact. "Mutation" is also used instead of "variant" for well-established terms (e.g., "background mutation rate").

A "genomic feature" (or simply "feature") is any sequence of nucleotides that share some property of interest. Features can be whole genes, but they can also be specific regions within a gene (e.g., exons, introns, coding sequences, or untranslated regions), promoters, enhancers, etc.

Computer code is displayed in a fixed-width typography (e.g., `bash`).

Chapter 1. Introduction

In this work, we have searched for cancer-promoting (or "driver") mutations in multiple types of genomic elements that do not code for protein in various cancers. This ***Introduction*** begins by summarizing the worldwide incidence of cancer. Next, it introduces the concept of driver mutations in the context of cancer evolution. Then, it outlines the composition of the human genome, especially of the regions that do not code for protein, and it summarizes the conceptual framework for how to determine whether a non-coding DNA sequence has a biological function. Afterward, it outlines the statistical and computational methods that are used to identify driver mutations in the non-coding genome. Finally, it introduces each type of non-coding element that will be addressed in this work, highlighting their known alterations in cancer.

1.1. The global burden of cancer

Cancer is a group of diseases in which abnormal cells grow uncontrollably, often invading adjacent or distant parts of the body (Wild et al., 2020). Cancer is a broad term, and a wide variety of cancer types with vastly different molecular characteristics can originate in almost any organ or tissue of the human body. Cancer is the first or second cause of premature death in most high- and medium-income countries. In 2018, 9.6 million people died from cancer worldwide, accounting for one in six deaths (World Health Organization, 2022). In the same year, the most diagnosed cancers in both sexes were those from the breast, lung, colorectum, prostate, and stomach (Wild et al., 2020) (***Figure 1A***). In addition, the deadliest cancers were those from the lung, followed by colorectum, liver, stomach, and breast (***Figure 1B***).

The World Health Organization estimates that up to 50% of cancer deaths could be prevented by multiple courses of action (World Health Organization, 2022). First, new cases can be reduced by promoting healthy lifestyles and reducing exposure to carcinogens, such as tobacco smoke (Wild et al., 2020). In addition, survival rates can be increased by diagnosing new cases at early stages, developing better therapies, and improving clinical procedures for

classifying, treating, and following up cancer patients. In fact, over the last two decades, improvements in clinical approaches have reduced cancer mortality in most high- and medium-income countries (Wild et al., 2020). Still, more research is urgently required to reduce the death toll of cancer.

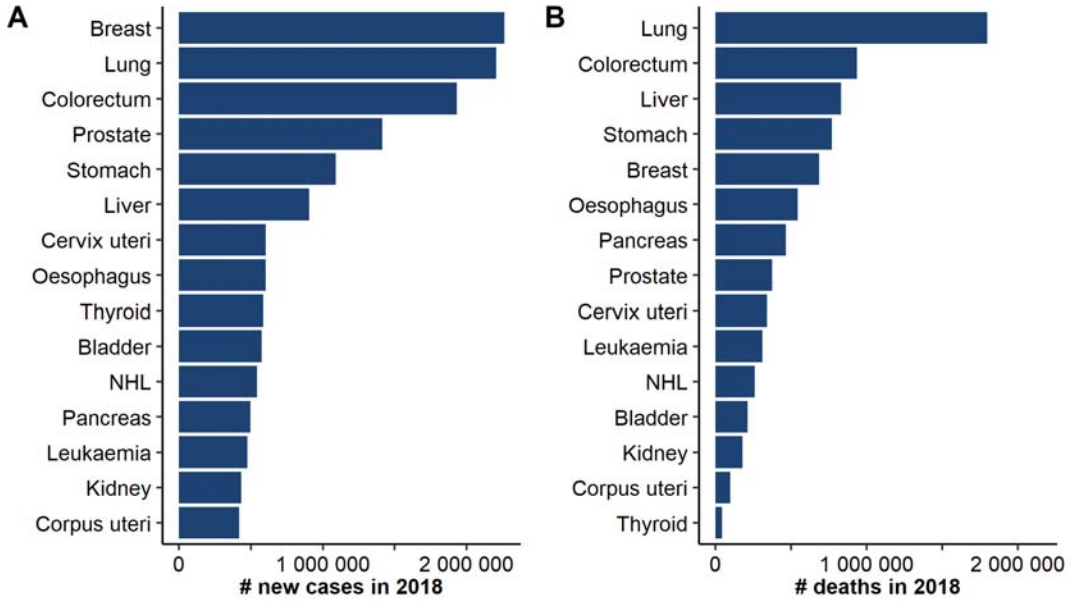


Figure 1. Worldwide cancer statistics by primary site. ***A.*** *Number of newly diagnosed cancers in 2018.* ***B.*** *Number of cancer deaths in 2018. NHL: Non-Hodgkin lymphoma. Data from the International Agency for Research on Cancer, https://gco.iarc.fr/today.*

1.2. Molecular drivers of cancer

To better diagnose, classify, treat, and follow up cancer patients, it is essential to decipher the molecular mechanisms by which cancer initiates and evolves. Such mechanisms may involve genetic, epigenetic, and environmental factors (Alizadeh et al., 2015). Any factor that promotes cancer must improve cell fitness, for example by stimulating cell growth, preventing cell death, or increasing resistance to treatment (Hanahan, 2022). In this text, we focus on genetic factors, which involve permanent changes (known as "variants" or "mutations") in the DNA sequence of cancer cell genomes. In particular, we focus on somatic variants, which are those specifically present in cancer cells but not in non-cancer cells from the same patient. However, germline variants

(i.e., those that are inheritable and are present in all cells of the body) also play a crucial role in some cancers (Campbell et al., 2020).

Cancer development can be understood as a Darwinian evolution process (Campbell et al., 2020) (***Figure 2***). Over time, cells in a tissue accumulate somatic variants in their DNA. Each variant has a minuscule probability of conferring a selective advantage to the cell. Those that do are known as driver mutations, and cells that harbor them are positively selected. However, most variants in the genome of a cancer cell are passenger mutations, which have a neutral or, occasionally, a detrimental effect on cell fitness. In particular, tumor genomes typically contain ~10^4-10^6 somatic variants, but only ~5 of them are estimated to be drivers (Rheinbay et al., 2020).

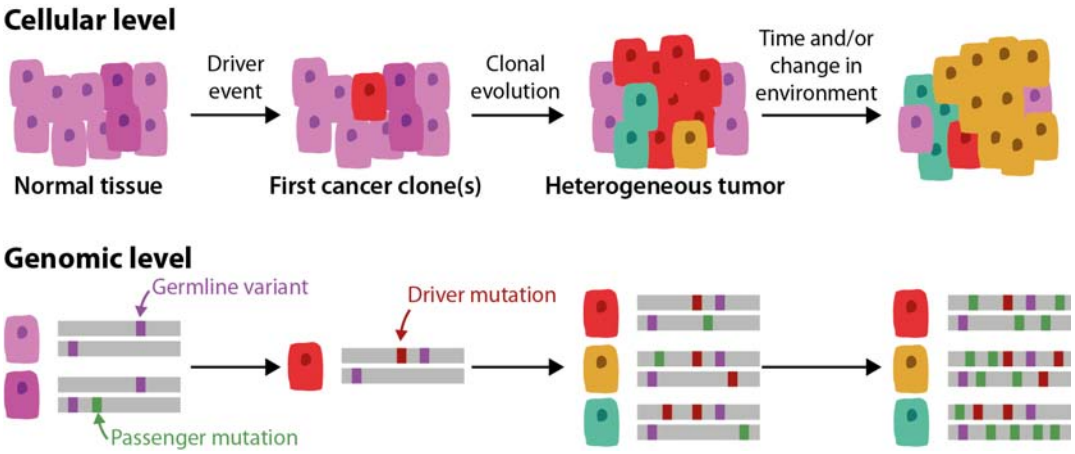


Figure 2. Clonal evolution of cancer. *Normal cells (purple) may acquire somatic variants throughout their lifespan. Most of these variants are expected to have a neutral effect on cell fitness (passenger mutations), but some may confer a competitive advantage (driver mutations), creating aberrant cells that outgrow their neighbors. As cancer cells proliferate, they may accumulate different sets of driver mutations, generating different clones (red, yellow, and blue). As time passes, or as the external conditions change, a new clone or an initially minor clone may become the dominant one, and previously dominant clones may become a minority or disappear.*

Although this simplified model of "driver" and "passenger" mutations explains the basic processes that govern cancer evolution, more layers of complexity can be added for better accuracy:

- **Intra-tumor heterogeneity and the dynamics of tumor evolution**. The distribution of mutations in a tumor changes over space and time (Gerstung et al., 2020). At a given time, a tumor is usually a mixture of different clones, which are groups of cells that have originated from a common ancestor and share the same set of driver mutations. As time passes, new driver mutations may appear and, with them, new clones may originate, expand, and outgrow the others. Changes in the external conditions, such as the patient starting a new treatment, can also change the equilibrium between clones. As a consequence, when a bulk tumor sample is sequenced, the detected variants represent a snapshot of a specific time point and, possibly, a mixture of different clones.
- **Drivers can be context-dependent**. For example, variants that confer resistance to a specific drug may be selectively advantageous as long as the patient is being treated with the drug, and otherwise they may be passengers (Kumar et al., 2020).
- **Some drivers may not be genetic**. Some researchers suggest that the definition of driver should incorporate epigenetic alterations and alterations that are not cell-autonomous (Alizadeh et al., 2015).
- **The definition of drivers and passengers may not be dichotomous**. Each variant in each cell is under a certain degree of selective pressure, which can be positive, negative, or neutral (Kumar et al., 2020):
    - **Positive selection** occurs when DNA sequence changes increase fitness, and therefore they are favored. Driver mutations are positively selected. Contrary to species evolution, cancer evolution is more driven by positive than by negative selection (Martincorena et al., 2017; Melton et al., 2015).
    - **Negative (or purifying) selection** occurs when DNA sequence changes decrease fitness, and therefore they are disfavored. Some passenger mutations are negatively selected. In cancer, negative selection has been rarely observed, except for hemizygous essential genes (Van den Eynden et al., 2016).

- **Neutral evolution** occurs when changes in the sequence are neither favored nor disfavored. Neutral processes largely shape the variability between species, populations, and individuals (Doolittle, 2013). In cancer, most passenger mutations are under neutral evolution. In some cases, the mutation truly has a neutral effect (neutral passenger). In other cases, however, it has a slightly positive or negative effect that is too small for selection to measurably act upon it (weak driver or weak deleterious mutations) (Kumar et al., 2020).

1.3. The non-coding genome

Human cells encode their genetic information in DNA, which contains the necessary instructions for their development, growth, differentiation, maintenance, and division. The collection of genetic information of a cell is known as its genome, and a human haploid genome has a size of ~$3.1 \cdot 10^9$ base pairs (bp) (Nurk et al., 2022). Part of the genome is transcribed into RNA, and part of the RNA is translated into protein (***Figure 3A***). In its original definition, the central dogma of molecular biology correctly posits that "the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible" (Crick, 1958).

In principle, cancer genomes could harbor variants in each of their thousands of millions of bases. However, analyzing every single nucleotide in the genome in search for driver mutations is a technically challenging and resource-demanding task that is only being made feasible in recent years thanks to whole-genome sequencing (WGS) technologies, which for now have only been adopted by some of the largest projects to date (Elliott and Larsson, 2021). As a consequence, in most mutational studies, it has been necessary to focus mutational analyses on the parts of the genome that are the most likely to harbor easily identifiable driver mutations. Therefore, studies of driver mutations have been influenced by the debate over which parts of the genome are expected to have any biological function at all, as well as by technical and methodological limitations.

As the end products of part of the genome, proteins have long been appreciated as crucial effectors that can perform a wide variety of biological functions (Crick, 1958). Accordingly, the protein coding genome has been thoroughly searched for driver mutations, and the vast majority of currently known driver mutations affect PCGs (Bailey et al., 2018; Martínez-Jiménez et al., 2020). However, there are only ~20 000 PCGs in humans, and only ~1.1% of the nucleotides of the human genome encode proteins (Frankish et al., 2019; Nurk et al., 2022) (***Figure 3B***). Therefore, several questions arise about the remaining ~98.9% of the human genome that does not code for protein: (i) what types of elements constitute it? (ii) Are these elements functional? And, (iii) are these elements implicated in cancer?

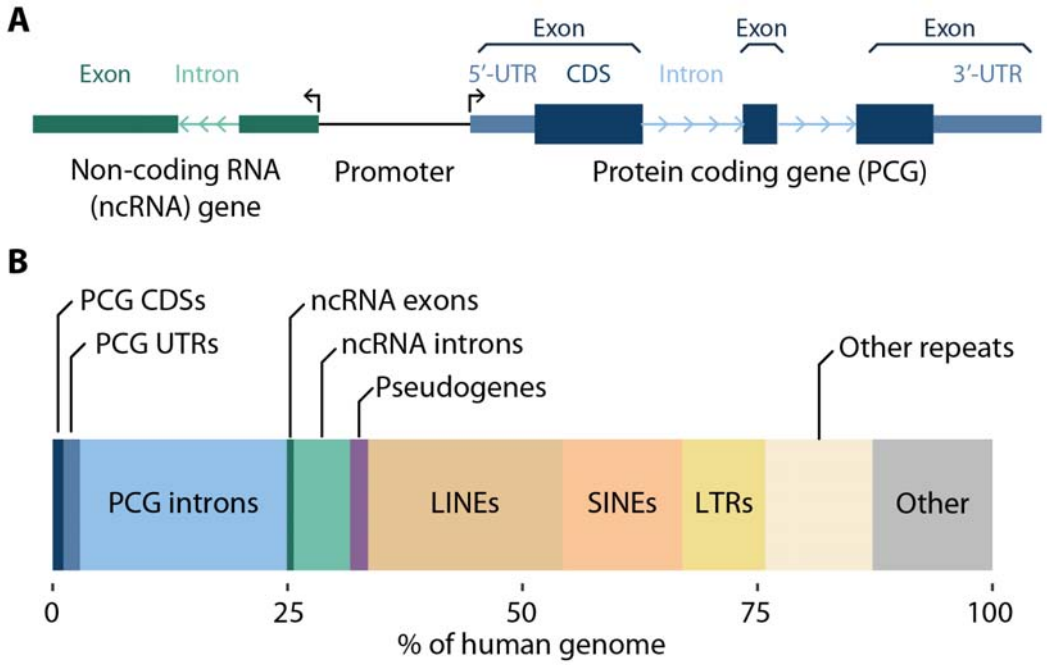


Figure 3. Components of the human genome. A*. Schematic representation of a hypothetical bidirectional promoter that modulates the expression of a non-coding RNA (ncRNA) gene in the 3'-5' direction and a protein coding gene (PCG) in the 5'-3' direction. Exons, introns, untranslated regions (UTRs), and coding sequences (CDSs) are labeled.* ***B****. Percentage of the human genome that is constituted by each type of genomic element. LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), and LTRs (long terminal repeats) are all repetitive DNA. Data sources: (Gregory, 2005; Lander et al., 2001; Nurk et al., 2022). Sizes of components of coding and non-coding genes were estimated from GENCODE v29.*

As an answer to the first question, there are many types of non-coding sequences in humans (Frankish et al., 2019; Nurk et al., 2022) (***Figure 3B***):

- **As part of PCGs**. Not all the DNA within PCGs encodes the protein sequence. When a PCG is transcribed, it generates a primary transcript that is then processed to generate a mature messenger RNA (mRNA). During this processing, some internal sequences, known as introns, are removed and the remaining sequences, known as exons, are joined together in a process known as splicing (***Figure 3A*** and ***Section 1.8***). Introns of PCGs are much longer than exons, and they constitute ~22% of the human genome (Frankish et al., 2019; Lander et al., 2001). Mature mRNAs also contain sequences that are not translated into proteins at their 5' and 3' ends. These sequences are known as untranslated regions (UTRs), and they are ~1.8% of the human genome (Frankish et al., 2019) (***Section 1.10***).
- **Non-coding RNA genes**. Non-coding RNA genes are genes that are transcribed into RNA but not translated into protein. Although the number of non-coding RNA genes in humans is still a controversial matter, current estimates roughly agree that there are a few tens of thousands, and their exons constitute ~0.7% of the human genome (Frankish et al., 2019; Nurk et al., 2022). Non-coding RNAs are a highly heterogeneous group. For example, they include transfer RNAs and ribosomal RNAs, which play well-known roles in protein translation (Frankish et al., 2019; Nurk et al., 2022). However, the remaining classes of non-coding RNAs have been historically less studied, and they are coarsely and arbitrarily classified based on the length of their RNA products. Thus, long non-coding RNAs (lncRNAs) are longer than 200 nt, whereas small non-coding RNAs are shorter than 200 nt (Slack and Chinnaiyan, 2019). Expectedly, the two groups are highly heterogeneous. On the one hand, short non-coding RNAs include microRNAs (miRNAs), PIWI-interacting RNAs, and small nuclear RNAs, among others (Slack and Chinnaiyan, 2019). On the other hand, lncRNAs are even more heterogeneous than their small counterparts and, currently, they cannot be classified into well-defined functional subclasses. Importantly, as will be detailed in later sections, miRNAs

and lncRNAs can regulate gene expression and some are consistently altered in cancer (***Sections 1.6-1.7***). Many non-coding RNA genes, especially lncRNA genes, contain introns, which are usually much longer than exons and may constitute ~6% of the human genome.

- **Pseudogenes**, which are inactive copies of PCGs, usually generated by gene duplication. Pseudogenes may or may not be transcribed, and some of them are even translated, but their products are believed to be non-functional, possibly with very few exceptions (Frankish et al., 2019; Slack and Chinnaiyan, 2019). Pseudogenes constitute ~1.9% of the human genome (Frankish et al., 2019).
- **DNA that plays a structural role**, such as telomeres and centromeres.
- **Regulatory DNA**, such as promoters and enhancers, which modulate gene expression.
- **Other intergenic non-coding DNA**, such as inactive transposons and viral sequences, which are classes of repetitive DNA.

About 54% of the human genome is made of repetitive sequences (Nurk et al., 2022) (***Figure 3B***). The majority of repetitive sequences in the human genome consist of inactive transposons and defective viral sequences, which are unlikely to have a biological function. On the other hand, some repetitive sequences, such as ribosomal RNA genes, have well-established functions.

Regarding the questions of whether non-coding DNA is functional and whether it is implicated in diseases such as cancer, current evidence suggests that the answer may be yes for a small part of it. Remarkably, less than 10% of the disease-associated polymorphisms identified by genome-wide association studies are located in coding sequences (Maurano et al., 2012). As a result, recent studies on cancer drivers have become increasingly interested in the non-coding genome (Elliott and Larsson, 2021). However, before exploring cancer driver elements in the non-coding genome, it is essential to robustly define what makes a DNA sequence functional, and to use that definition to predict which non-coding DNA elements are likely to have a function. Furthermore, this conceptual framework will be essential to distinguish between driver ("functional") and passenger ("non-functional") mutations.

1.4. Selective pressure, biological function, and non-coding DNA

Strictly, a DNA sequence is functional if and only if it is under selection (Doolittle, 2013). In other words, if a sequence evolves neutrally, it is not functional. As discussed in ***Section 1.2***, selection can be positive or negative. Whereas positive selection is more important than negative selection in cancer evolution, the reverse is generally true in the evolution of species and populations. Therefore, negative selection is a useful indicator of whether a DNA sequence is functional in humans. Negative selection can be quantified based on evolutionary conservation or on the constraint of the sequence within populations (Karczewski et al., 2020; Pollard et al., 2010). In contrast to evolutionary methods, population-based methods can detect functions that have been acquired or lost recently in evolution. However, they require very large cohorts and, for now, analyses in the largest cohorts are mostly restricted to coding regions (Karczewski et al., 2020).

Decades of research on evolutionary biology and human population genetics have led to the estimate that approximately 10%, and no more than ~15%, of the human genome has a sequence-dependent biological function (Meader et al., 2010; Ponting and Hardison, 2011; Ward and Kellis, 2012). Because only ~1.1% of the human genome codes for proteins, even if assuming that all protein-coding DNA is functional, there would be at least ~8 times more functional non-coding DNA than functional coding DNA, confirming the non-coding genome as a vast resource of functional DNA that might be implicated in disease. On the other hand, these estimates also mean that ~90% of the human genome is not functional. This agrees with the fact that more than half of the human genome is made of inactive transposons, inactive viral sequences, and introns (Nurk et al., 2022) (***Figure 3***).

The estimates discussed above were challenged by large-scale next-generation sequencing-based studies. Most importantly, the Encyclopedia of DNA Elements (ENCODE) Project reported that ~80% of the human genome has a biochemical "function" (Dunham et al., 2012). ENCODE defined "function" as any type of biochemical activity, such as transcription, reproducible protein

binding, or histone marks. ENCODE's claims have been heavily criticized because they misuse the concept of biological function (Doolittle, 2013; Eddy, 2013; Graur et al., 2013). Most crucially, non-specific or non-functional interactions between biomolecules occur extensively *in vivo*, causing spurious transcription and protein-nucleic acid binding (Struhl, 2007; Willingham and Gingeras, 2006). ENCODE considered these events as functional, thus inflating their estimates (Doolittle, 2013; Eddy, 2013; Graur et al., 2013).

In conclusion, the non-coding genome might be a vast resource of functional and disease-related DNA sequences, and this is fully compatible with the notion that the majority of the human genome is functionless. In this text, we will use the term "function" in its strict definition (i.e., under selection), and otherwise we will use the term "activity" instead.

1.5. Identifying cancer drivers in non-coding sequences: a methodological overview

1.5.1. General rationale of driver discovery

Whereas negative selection is useful for identifying functional DNA sequences in humans, positive selection is the most important indicator of whether a variant promotes cancer. Indeed, driver mutations are defined as mutations that are positively selected in cancer. When mutations are positively selected in a cohort, they can be observed at higher frequencies than expected under a background model of neutral evolution. There are multiple methods for detecting positive selection, and they differ in how they measure the accumulation of variants and, most importantly, in how they model their background (Rheinbay et al., 2020).

Regarding how the accumulation of variants is measured, all methods consider recurrence directly or indirectly (Rheinbay et al., 2020). Some methods leverage additional information, such as the predicted functional impact of each variant (Martincorena et al., 2017; Mularoni et al., 2016). On the other hand, other methods rely on the distribution of variants alone, for example by searching for clusters of nearby variants (Arnedo-Pac et al., 2019).

However, even some of the latter require a definition of "silent" variants to estimate their background, and therefore they require defining which variants are expected to have no functional impact (Tamborero et al., 2013).

Constructing a correct background is the main challenge of driver discovery. The background mutation rate (BMR) is not constant across the genome because it is influenced by multiple factors at various levels (Gonzalez-Perez et al., 2019; Supek and Lehner, 2019). At a genome-wide level, the BMR is mainly influenced by replication timing. At a gene level, the BMR is affected by chromatin state and by transcription levels. For example, highly transcribed tissue-specific genes have high indel rates (Imielinski et al., 2017). At a single nucleotide level, the BMR depends on the trinucleotide context (i.e., the mutated nucleotide and its two immediately adjacent nucleotides). Moreover, in some tumor types, DNA or RNA editing enzymes such as the activation induced-cytidine deaminase (AID) and the apolipoprotein B mRNA-editing enzyme catalytic subunit (APOBEC) mutate certain sequence motifs or DNA hairpins, respectively (Gonzalez-Perez et al., 2019; Supek and Lehner, 2019). Furthermore, transcription factor binding sites in highly active promoters have an increased mutation rate, in part because bound transcription factors physically hinder DNA repair (Perera et al., 2016). This is also true for other genomic positions that are regularly bound to proteins, such as CTCF binding sites (Supek and Lehner, 2019). Overall, knowledge on local determinants of the BMR is recent and, most likely, incomplete (Elliott and Larsson, 2021).

Although recent driver discovery methods account for global determinants of the BMR, their ability to account for local determinants is limited (Rheinbay et al., 2020). As a result, the "hits" reported by driver discovery tools may contain false positives, and they must be carefully evaluated in search for artifacts and signs of local mutational processes. Further complicating the issue, not all hits originating from background mutational processes are necessarily false positives. For example, we previously reported that recurrent somatic variants affecting a splice donor site of *BCL7A* in lymphoma were caused by AID activity, but they were also under positive selection (Baliñas-Gavira et al., 2020). Moreover, we showed that the variants improved cell fitness *in vivo* and *in vitro* compared to wild type *BCL7A*.

Most driver detection methods aim to identify driver features, defined as genomic features whose patterns of somatic variants suggest positive selection. However, not all variants that affect a driver feature are driver mutations (i.e., not all of them are positively selected and contribute to the cancer phenotype) (Martincorena et al., 2017). To our knowledge, few methods have been developed to systematically distinguish driver from passenger mutations within a driver feature, and only in PCGs (Muiños et al., 2021). Otherwise, pipelines to distinguish driver from passenger mutations rely on external information, such as functional impact scores, conservation, prediction of transcription factor binding sites, or the effect of the variant on gene expression, which must be evaluated on a case-by-case basis (Fu et al., 2014; Rheinbay et al., 2020).

1.5.2. Challenges of searching for non-coding drivers

Some driver discovery methods rely on predictions of the functional impact of variants. In coding sequences (CDSs) of PCGs, such predictions are straightforward thanks to the genetic code. For example, the ratio between non-synonymous and synonymous variants (dN/dS) quantifies selection in PCGs (Martincorena et al., 2017). As a result, PCG-centric driver discovery tools are numerous and well-tested. Remarkably, multiple state-of-the-art driver discovery tools have been integrated to obtain comprehensive "consensus" collections of pan-cancer and cancer-specific driver PCGs (Bailey et al., 2018; Martínez-Jiménez et al., 2020). However, even consensus-based approaches require manual curation to remove false positives and rescue false negatives (Bailey et al., 2018).

In contrast, methods to predict non-coding drivers have been less developed, and comprehensive reports of non-coding drivers are rare (Fredriksson et al., 2014; Fujimoto et al., 2016; Nik-Zainal et al., 2016; Puente et al., 2015; Rheinbay et al., 2020; Weinhold et al., 2014). Overall, there are multiple reasons why research on non-coding drivers has been slow:

- **Historical lack of cancer sequencing data in non-coding regions**. Large-scale WGS projects have only been feasible in recent years. Before the era of WGS, sequencing was mostly limited to the exome, or to parts of it. Although targeted sequencing of non-coding DNA was, in theory, possible, sequencing studies still neglected non-coding regions. It may be argued that non-coding regions were ignored, in part, because proteins have been historically considered as the main effectors of eukaryotic phenotypes. However, this explanation is not complete because many non-coding sequences, such as promoters, enhancers, ribosomal RNAs and transfer RNAs, have long been known to be functional. Instead, we propose the remaining reasons below.
- **Annotation of functional non-coding elements is incomplete and inaccurate**. For example, annotation of non-coding RNA genes has greatly evolved over the past decade thanks to technological and methodological improvements (Cao et al., 2018; Frankish et al., 2019; Lorenzi et al., 2021). However, it is currently unclear which annotated lncRNAs are functional and which ones are "transcriptional noise" (i.e., non-functional RNA generated by spurious transcription) (Cao et al., 2018).
- **Predicting the effect of non-coding variants is challenging** because the relationship between sequence and function of most non-coding regions is not fully understood. In addition, most non-coding sequences may tolerate variants better than coding sequences. In contrast, the functional impact of variants in coding sequences can be predicted using the genetic code, and this is exploited by PCG-centric driver discovery methods.
- **Low mappability**. Many non-coding sequences have multiple copies along the genome, which can cause mapping artifacts and impair analyses in these regions (Shuai et al., 2019; Suzuki et al., 2019).
- **Modeling the BMR in non-coding regions is challenging** because localized mutational processes in non-coding regions are poorly understood (Imielinski et al., 2017; Nik-Zainal et al., 2016; Rheinbay et al., 2020). Without a proper background model, signals of positive selection cannot be accurately distinguished from noise.

1.5.3. Challenges of predicting the functional impact of non-coding variants

The function of non-coding sequences, if any, resides in the DNA itself or in non-coding RNA products, and therefore the variety of criteria that can be used to predict the functional impact of non-coding variants is limited. One of the most important criteria is evolutionary conservation (Pollard et al., 2010). If the rate of substitutions in a nucleotide is slower than expected under neutral evolution, the nucleotide is likely to be functional, and variants in the nucleotide are likely to alter its function. However, a conserved nucleotide may only be essential in specific cell types, contexts, or developmental stages, and some sequences may have gained or lost their function recently in evolution. Therefore, a somatic variant affecting a highly conserved nucleotide in a specific tumor need not have any functional impact. In addition, conservation metrics are not allele-specific.

Besides conservation, other proposed criteria for quantifying the functional impact of non-coding variants are DNA accessibility, disruption or creation of transcription factor binding sites, histone modifications, expression, and distance to exon-intron boundaries (Kircher et al., 2014; Shihab et al., 2015). None of these metrics alone can accurately predict the functional impact of non-coding variants, and they must be used in conjunction with conservation metrics, as they may help circumvent the limitations of using conservation metrics alone. Thus, recent methods such as Combined Annotation–Dependent Depletion (CADD) and Functional Analysis Through Hidden Markov Models and Multiple Kernel Learning (FATHMM-MKL) integrate multiple metrics into a single allele-specific score using machine learning (Kircher et al., 2014; Shihab et al., 2015). In this way, both CADD and FATHMM-MKL can predict the functional impact of every single possible variant in the human genome.

1.5.4. Methods for non-coding driver discovery

As will be detailed in later sections, in our work we performed targeted sequencing of miRNAs, lncRNAs, and PCGs in search for novel non-coding cancer drivers. Therefore, we were especially interested in driver discovery tools that are valid, at least in theory, for targeted sequencing of non-coding regions. Here, we describe the two tools that we selected for our purposes.

Among the methods that incorporate the functional impact of variants, OncoDriveFML is of particular interest for our purposes because, in contrast to most other tools, it is applicable to targeted sequencing of non-coding DNA (Mularoni et al., 2016). OncoDriveFML computes the mean functional impact score (from any external source, such as CADD or FATHMM-MKL) of the observed variants in each feature of interest (***Figure 4A***). Then, to construct the background, it randomly simulates variants along each feature of interest and computes their mean functional impact score. It performs the simulations a large number of times (by default, 10 000) accounting for differences in the mutation rate in different trinucleotide contexts, estimated empirically using all variants in the cohort. Finally, for each feature, it compares the observed mean functional impact score with that of its background, obtaining a p value that is then corrected to control the false discovery rate. Because OncoDriveFML constructs the background locally, it is insensitive to factors that affect the BMR at a genomic scale. In addition, whereas other driver discovery methods require variants from intergenic regions or other "silent" variants to construct their background, OncoDriveFML does not, which makes it applicable to targeted sequencing (Rheinbay et al., 2020).

Among the methods that are independent of functional impact annotations, we selected OncoDriveCLUSTL (Arnedo-Pac et al., 2019). The tool identifies positive selection in any feature of interest by searching for clusters of variants (***Figure 4B***). Variant clusters are scored based on the number of variants that they contain and the shape of the smoothed distribution of variant density. OncoDriveCLUSTL constructs its background by simulating random mutations within the feature in a similar manner to OncoDriveFML. Therefore, OncoDriveCLUSTL is also applicable, in theory, to targeted sequencing of non-coding DNA (Arnedo-Pac et al., 2019).

Using both OncoDriveFML and OncoDriveCLUSTL, driver mutations can be searched for in targeted sequencing of various types of non-coding features. In the remaining sections of this ***Introduction***, we introduce the three main non-coding features analyzed in our work: lncRNAs, miRNAs, and intronic splice regions. Finally, we discuss two other types of non-coding elements that we analyzed in lower detail: promoters and UTRs.

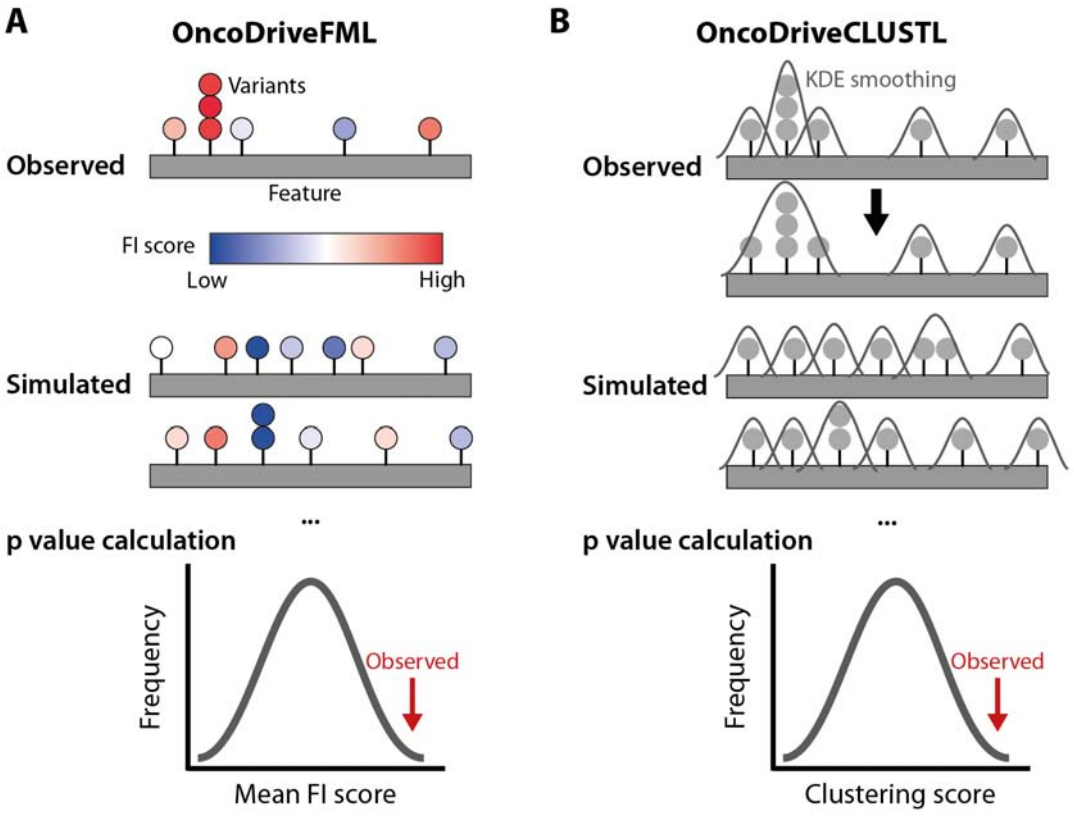


Figure 4. Methods for driver discovery used in this work. A. OncoDriveFML*. Variants are scored based on their predicted functional impact (FI), which is calculated by external tools. Then, simulations are performed in which variants are randomly generated and their FI score is estimated. Next, the mean FI score of the observed variants is compared to the background distribution obtained from the simulations, yielding a p value for each genomic feature of interest.* ***B. OncoDriveCLUSTL****. Kernel density estimator (KDE) smoothing is performed on the variants, and clusters of variants are searched for. A "clustering score" is calculated for each genomic feature of interest. Simulations are performed to estimate the background distribution of clustering scores. Finally, the observed clustering score is compared to the background distribution to obtain a p value. Figures adapted from (Arnedo-Pac et al., 2019; Mularoni et al., 2016).*

1.6. Long non-coding RNAs in cancer

1.6.1. Long non-coding RNAs are a heterogeneous group of biomolecules

Long non-coding RNAs (lncRNAs) are defined as non-coding RNAs that are longer than 200 nucleotides (Statello et al., 2021). The definition of lncRNAs is extremely broad, and as a consequence lncRNAs can have vastly different properties and functions. Most, but not all, lncRNAs are transcribed by RNA polymerase II, m^7G-capped at their 5' end, and polyadenylated at their 3' end, in a similar manner to mRNAs (Statello et al., 2021). However, several features distinguish lncRNAs from mRNAs. The most important one is that, by definition, lncRNAs are not translated into functional proteins. In addition, lncRNAs have lower evolutionary conservation, number of exons, expression levels, and splicing efficiency than mRNAs. Moreover, lncRNAs tend to be more localized in the nucleus than mRNAs.

Because lncRNAs are such a heterogeneous group, they can be classified in multiple ways (Kopp and Mendell, 2018) (***Figure 5***):

- **Based on their genomic location relative to their closest gene** (***Figure 5A***), lncRNA loci can be intergenic, antisense (if they overlap or they are close to another gene in the opposite strand), or sense (if they overlap another gene in the same strand).
- **Based on the locality of their action** (***Figure 5B***), lncRNAs can act in *cis* or in *trans*. On the one hand, some lncRNAs never abandon the chromatin fraction, but they modulate nearby genes in *cis* (Kopp and Mendell, 2018). These lncRNAs are strictly nuclear, and their steady-state expression levels tend to be low because they are degraded quickly. On the other hand, some lncRNAs leave the chromatin fraction and act in *trans*, modulating distant genes or participating in cellular processes outside of the chromatin fraction. These lncRNAs can be nuclear or cytosolic, and their steady-state expression levels tend to be higher than those of *cis*-acting lncRNAs because they must be stable enough to reach their destination.

- **Based on where their functionality resides** (***Figure 5C***). The function of some lncRNA loci is dependent on the RNA sequence itself. However, this is not always the case. Because regulatory DNA elements such as promoters and enhancers are transcribed bidirectionally in humans, a lncRNA locus may be functional only because it contains or it is downstream of a regulatory DNA element, and not because of the RNA product (Ibrahim et al., 2018). In other cases, the sequence of the lncRNA is irrelevant for its function, but its act of transcription or splicing modulates nearby genes in *cis* (Kopp and Mendell, 2018). For example, when a lncRNA locus overlaps another gene, transcription of the lncRNA may physically interfere with transcription of the overlapping gene. Moreover, the function of some lncRNAs may be dependent on their secondary structure, and therefore their RNA sequence may tolerate variants as long as they do not disrupt the structure. Moreover, all these mechanisms are not mutually exclusive (Marchese et al., 2017). Therefore, when characterizing a novel lncRNA, all possibilities must be considered, and it should not be assumed by default that their function is dependent on the RNA sequence.
- **Based on their mechanism of action** (***Figure 5D***). LncRNAs can have extremely diverse mechanisms of action by binding to chromatin, to other RNAs, or to proteins. LncRNAs often modulate gene expression, either via chromatin remodeling, transcriptional regulation in *cis* or in *trans*, or by acting as scaffolds for proteins (reviewed by (Statello et al., 2021)). At a post-transcriptional level, lncRNAs may mediate the formation of RNA-protein complexes, organize nuclear architecture, or sequester proteins or miRNAs to impair their function (***Section 1.7***). The lncRNAs that bind to miRNAs, preventing them from binding to their other targets, are named competing endogenous RNAs (ceRNAs).

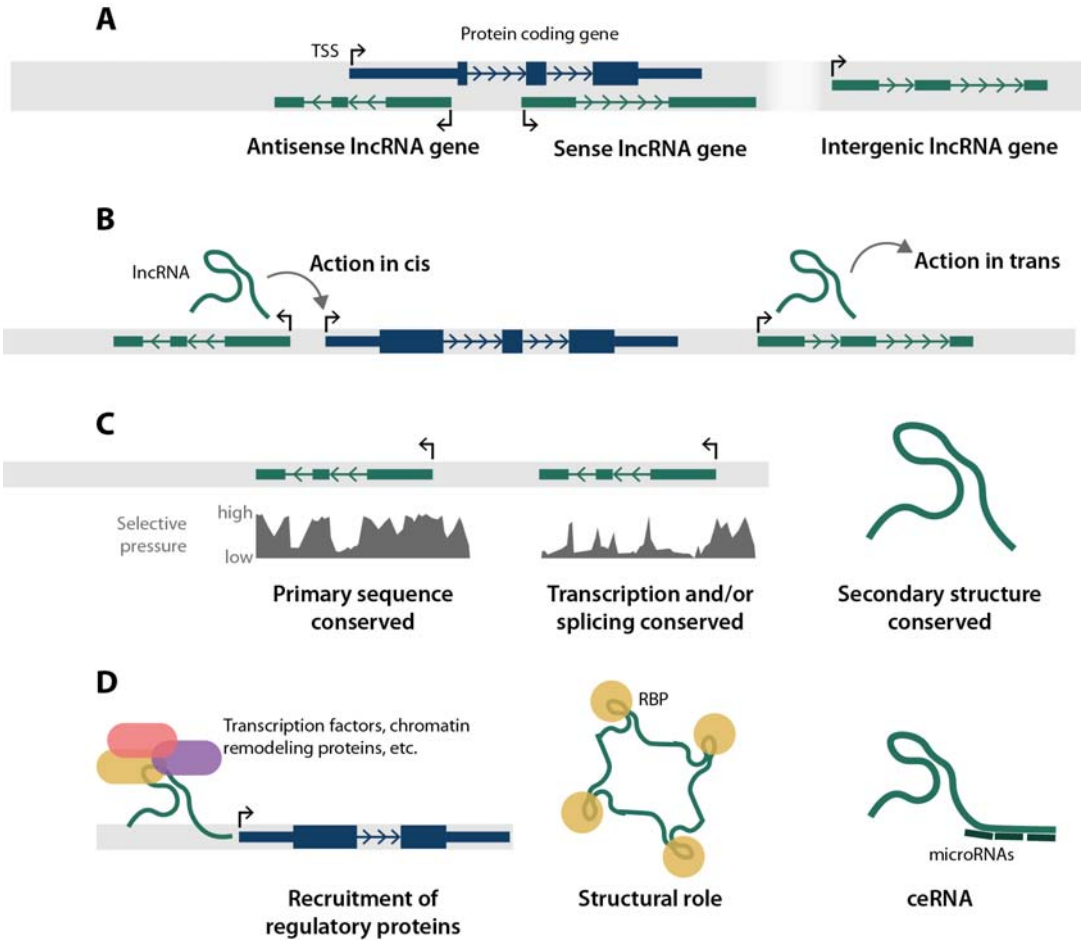


Figure 5. The diversity of long non-coding RNAs (lncRNAs). ***A****. Classification of lncRNAs based on their genomic location relative to protein coding genes. Transcription start sites (TSSs) are marked with black arrows.* ***B****. lncRNAs may act in cis, i.e., directly on nearby genes, or in trans, i.e., on distant loci or subcellular compartments.* ***C****. The function of lncRNAs may be dependent on their primary sequence (left), on their act of transcription and/or splicing (middle), or on their secondary structure (right).* ***D****. Some proposed molecular mechanisms for lncRNA function include: recruitment of regulatory proteins, such as transcription factors and chromatin remodeling proteins, to specific loci (left); structural roles, e.g., by forming subcellular structures, such as speckles and paraspeckles, or by mediating contacts between chromosomes (middle); and sponging of microRNAs, preventing their function (right). RBP: RNA binding protein; ceRNA: competing endogenous RNA.*

1.6.2. LncRNAs are altered in cancer

LncRNAs have diverse biological functions that are usually related with the regulation of gene expression. As a consequence, lncRNAs are often dysregulated in cancer, directly participating in all hallmarks of cancer (Goodall and Wickramasinghe, 2021; Huarte, 2015). The Cancer LncRNA Census (CLC), a curated compendium of cancer-related lncRNAs, currently contains 492 entries (Vancura et al., 2021). Some representative lncRNAs that have been linked to each cancer hallmark are summarized in ***Table 1***.

Besides their direct role in tumorigenesis, lncRNAs may also be useful as cancer biomarkers. For example, the lncRNA *PCA3* is expressed specifically in prostate cancer (Hessels et al., 2003), and a diagnostic test that detects *PCA3* expression has been approved by the U.S. Food and Drug Administration (FDA) to help determine whether it is necessary to repeat a biopsy in patients who have had a previous negative biopsy (FDA approval number: P100033). Although, to our knowledge, *PCA3* is the only approved lncRNA biomarker to date, expression of other lncRNAs may have clinical value in the future. For example, expression of *MALAT1* is associated with patient survival and tumor metastasis in non-small cell lung cancer (Ji et al., 2003). In addition, expression of *TCL6* is strongly associated with specific subtypes of childhood leukemia (Cuadros et al., 2019). Overall, reports on lncRNAs whose expression is associated with specific cancer subtypes or with patient prognosis have been growing exponentially over the last decades, but they have very rarely reached clinical application.

Although there are myriads of reports of lncRNAs whose expression is altered in cancer, few studies have explored other types of alterations, such as those at the genomic level. However, genomic alterations that involve lncRNAs may help uncover novel roles of lncRNAs. For example, by searching for genomic regions that are recurrently amplified or deleted in cancer, novel oncogenic or tumor suppressor lncRNAs can be discovered, respectively (Athie et al., 2020). In addition, in theory, point mutations could also affect lncRNA function in cancer, for example by altering their expression or by affecting their binding affinity to other biomolecules. However, driver point mutations in lncRNAs may be uncommon (Elliott and Larsson, 2021; Rheinbay et al., 2020).

Based on their accumulation of point mutations and small indels, *NEAT1* and *MALAT1* used to be strong candidate driver lncRNAs (Fujimoto et al., 2016; Nik-Zainal et al., 2016). However, their accumulation of variants may not be due to positive selection, but due to background processes (Rheinbay et al., 2020). Other lncRNAs, such as *G029190* and *CTD-2105E13.15*, were recently reported as possible pan-cancer drivers, but their signal was weak and their functional significance was unclear (Rheinbay et al., 2020).

Table 1. Representative examples of lncRNAs that have been reported to promote (+) or suppress (-) cancer hallmarks.

Hallmark	lncRNA	Effect	Cancer	Reference
Sustained proliferation	*MEG3*	-	Various	(Zhou et al., 2007)
Evasion of growth suppressors	*CDKN2B-AS1*	+	Various	(Kotake et al., 2011)
Immune evasion	*ALAL-1*	+	Lung	(Athie et al., 2020)
Replicative immortality	*TERC*	+	Cervix	(Andersson et al., 2006)
Inflammation	*PACERR*	+	Pancreas	(Liu et al., 2022)
Invasion and metastasis	*HOTAIR*	+	Breast	(Gupta et al., 2010)
Angiogenesis	*TUG1*	+	Glioblastoma	(Cai et al., 2017)
Genome instability and mutagenesis	*PCAT1*	+	Prostate	(Prensner et al., 2014)
Avoidance of cell death	*PANDAR*	+	Various	(Hung et al., 2011)
Metabolism deregulation	*TP53COR1*	+	Various	(Yang et al., 2014)

Hallmarks of cancer are described in (Hanahan, 2022). Newly emerging hallmarks have not been compiled here.

1.6.3. The controversy over lncRNA function

1.6.3.1. Perspectives from evolutionary biology and population genetics

Over the last decades, lncRNAs have consolidated as a distinct but diverse group of biological molecules. However, a lingering question is how many of the tens of thousands of human lncRNAs have an actual biological function. This question is especially challenging to answer because annotations of lncRNAs are unstable and because their functions are sometimes sequence-independent, relying instead on their transcription, splicing, or secondary structure (Kopp and Mendell, 2018). Nevertheless, several remarkable approaches have been made to tackle the question of the functionality of lncRNAs (reviewed by (Ulitsky, 2016)).

From an evolutionary point of view, lncRNA exons and promoters are more conserved than introns and intergenic regions, but less conserved than CDSs (Guttman et al., 2009; Haerty and Ponting, 2013). In particular, when accounting for conservation of sequence, transcription, or splicing patterns, ~20-30% of high-confidence lncRNAs are conserved between humans and primates (Ulitsky, 2016). However, evolutionary approaches have encountered major limitations. First, conservation of lncRNA exons is only significant when analyzing curated sets of high-confidence lncRNAs, suggesting that general annotations may be plagued with non-functional entries (Haerty and Ponting, 2013). Second, evolutionary methods have mostly evaluated full sequences of lncRNAs, and therefore they are unable to detect small and highly conserved sequences within mostly non-conserved lncRNAs, an occurrence that has been described for some well-characterized lncRNAs (Ulitsky, 2016). Third, rapid evolution does not mean lack of function, as some lncRNAs may have acquired biological functions recently during evolution. For example, whereas human *HOTAIR* may be involved in the regulation of the *HOXD* cluster, there is no evidence for such activity in mice (Schorderet and Duboule, 2011). The latter limitation can be overcome by analyzing purifying selection within populations of the same species.

From an intra-specific point of view, little evidence of purifying selection has been found in lncRNA exons in humans (Haerty and Ponting, 2013). However, it has been robustly observed in other species, such as fruitflies (Haerty and Ponting, 2013). Because the effective population size of humans is relatively small (~2-3 orders of magnitude smaller than that of fruitflies), purifying selection may not be able to act on weakly deleterious variants in lncRNA exons, and therefore lncRNA exons may evolve neutrally in humans (Haerty and Ponting, 2013).

1.6.3.2. Inconsistent annotation of lncRNAs hinders research on lncRNA function

Studies on lncRNA function have been seriously limited by the lack of a robust annotation of human lncRNAs. Indeed, defining what is and what is not a lncRNA gene is a controversial subject. RNA polymerases and transcription factors spuriously bind to accessible DNA sequences across the whole genome and generate non-functional RNAs, causing what is known as transcriptional noise (Struhl, 2007; Willingham and Gingeras, 2006). Transcriptional noise is widespread in the genome, but its RNA products usually have low concentrations and they are quickly degraded. However, some bona fide lncRNAs have very low steady-state concentrations, even below one copy per cell (Seiler et al., 2017). Furthermore, regulatory DNA elements such as promoters and enhancers are usually transcribed in both directions, generating non-coding RNAs whose functionality is unclear in most cases and whose concentration is correlated with the activity of the promoter or enhancer (Ibrahim et al., 2018). Therefore, it is extremely challenging to distinguish between functional lncRNAs and non-functional RNAs generated by non-specific processes. As a consequence, using different criteria and different data to annotate lncRNA loci can lead to vastly different results. Illustratively, current estimates of the number of human lncRNA genes range from ~18 000 to ~100 000 (Frankish et al., 2019; Zhao et al., 2021a). However, without a proper and robust annotation, broad questions about the functionality of lncRNAs as a set cannot be answered accurately.

1.6.3.3. Transcription does not mean function

The debate on how many lncRNAs are functional was strongly reignited by the ENCODE Project Consortium, who claimed that ~80% of the human genome is functional and who classified all transcribed DNA sequences as functional (Dunham et al., 2012). ENCODE found that up to ~75% of the human genome was transcribed in at least one of their analyzed cell types, and ~39% was transcribed in a typical cell line (Djebali et al., 2012). Most of these transcripts were non-coding and not conserved, and ~80% of the non-coding transcripts were expressed at 1 or fewer copies per cell.

Then, how many of the non-coding transcripts identified by ENCODE were functional? Members of ENCODE argued that most of them were likely to be functional because they had tissue-specific expression, and that they may have sequence-independent functions that could explain their lack of evolutionary conservation (Mattick and Dinger, 2013). As a counterargument, transcriptional noise is also tissue-specific because each cell type in each context expresses a specific set of transcription factors and has specific patterns of chromatin accessibility (Eddy, 2013). Furthermore, although some lncRNAs have sequence-independent functions, they are believed to be rare (Ponting and Hardison, 2011), and they are still constrained at some other level (e.g., secondary structure, transcription, or splicing) (Kopp and Mendell, 2018; Ulitsky, 2016). Finally, although some functions in human genomes may have been acquired recently in evolutionary time, they are likely to represent only a small fraction of the human genome according to estimates from ENCODE researchers (Ward and Kellis, 2012).

In conclusion, for any uncharacterized non-conserved non-coding transcript, the simplest explanation is that it originates from transcriptional noise, even if its transcription is tissue-specific. The null hypothesis should always be that of no function, and any claim on functionality should be robustly proven.

1.7. MicroRNAs in cancer

1.7.1. MicroRNA biogenesis and function

MicroRNAs (miRNAs) are short (~22 nt) non-coding RNAs that modulate gene expression (Bartel, 2004). In the nucleus, the primary transcripts of miRNA genes (pri-miRNAs) form hairpin structures that are recognized by the Microprocessor complex, which is constituted by one copy of the RNase III endonuclease DROSHA and two copies of the RNA-binding protein DGCR8 (***Figure 6***) (Partin et al., 2020). Auxiliary proteins, such as SRSF3, may help Microprocessor bind to pri-miRNAs (Kim et al., 2021). Microprocessor cleaves pri-miRNAs at specific sites generating a stem-loop structure known as precursor miRNA (pre-miRNA). Pre-miRNAs are then exported to the cytosol, where another RNase III endonuclease, DICER, cleaves off the loop end of the pre-miRNA, generating a double-stranded RNA (dsRNA) duplex constituted by the mature miRNA and a quasi-complementary fragment of similar size. Finally, the mature miRNA, in its single-stranded RNA (ssRNA) form, binds to a multiprotein complex constituted by Argonaute and accessory proteins and forms the RNA-induced silencing complex (RISC). The mature miRNA guides RISC towards target mRNAs, promoting their cleavage or translational repression. Mature miRNA names end in either "-5p" or "-3p" based on whether they originated from the stem-loop arm closest to the 5' end or to the 3' end of the sequence, respectively.

The target specificity of miRNAs is mostly determined by their "seed" sequence, which is defined as nucleotides 2-7 in the mature miRNA (Agarwal et al., 2015). The model of canonical target recognition by miRNAs establishes that the seed sequence binds to the 3'-UTRs of target mRNAs by perfect base pairing, which can be strengthened by a match at position 8 and/or by an adenine opposite position 1. Models based on such simple assumptions can explain most experimentally validated miRNA-target interactions, and functional "non-canonical" binding is rare (Agarwal et al., 2015). Indeed, these models are the basis of one of the most popular miRNA target prediction tools, TargetScan (https://www.targetscan.org).

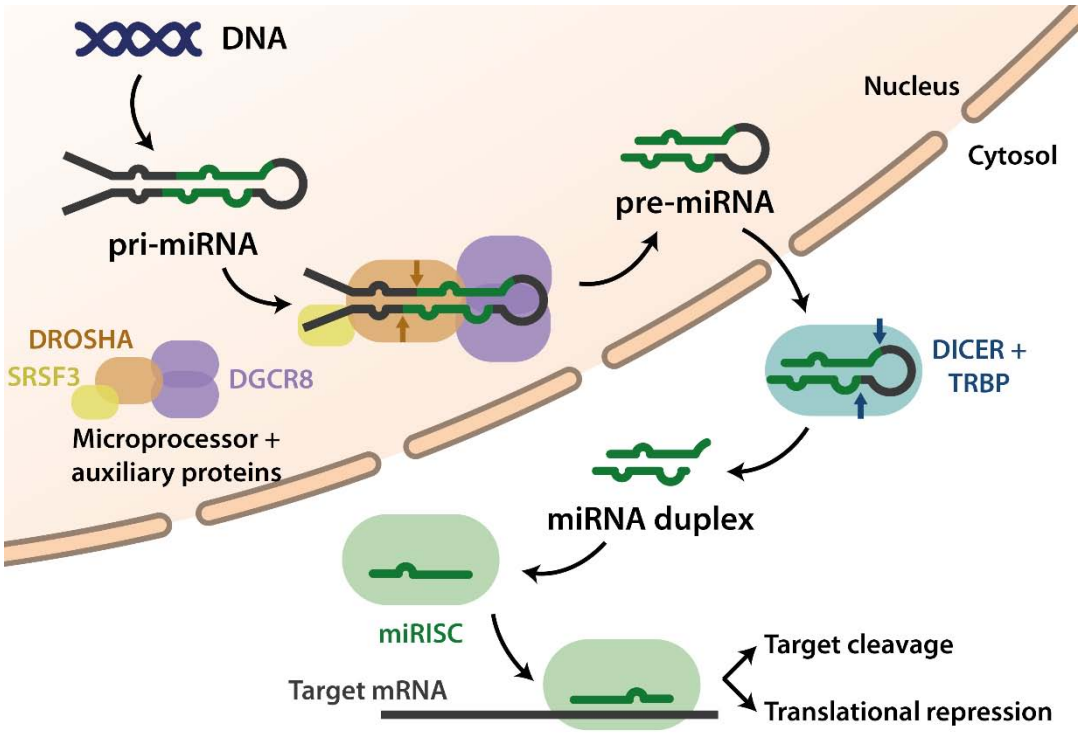


Figure 6. Canonical miRNA biogenesis pathway*. The primary transcripts of miRNA genes (pri-miRNAs) undergo several processing steps to generate mature miRNAs. See main text for details. Pre-miRNA: precursor miRNA; miRISC: RNA-induced silencing complex, loaded with a mature miRNA.*

1.7.2. Processing motifs in miRNAs

Seed sequences are critical for miRNA function, as proven by the fact that seed-target interactions are highly conserved in evolution (Bartel, 2009). Seed sequences are defined by their position in mature miRNAs, which is determined by where DROSHA and DICER cleave the pri- and pre-miRNAs, respectively (***Figure 6***). Therefore, DROSHA and DICER cleavage must be highly precise, and a shift of just 1 nt in a cleavage site can completely change the seed sequence and, with it, the targets of a mature miRNA. DICER determines its cleavage site by measuring ~22 nt from the basal end of the pre-miRNA, which is dictated by DROSHA (Park et al., 2011). Therefore, in practice, DROSHA is the main determinant of the sequences of mature miRNAs. Accordingly, DROSHA cleavage is tightly controlled (Auyeung et al., 2013; Fang and Bartel, 2015; Kim et al., 2021; Kwon et al., 2019; Roden et al., 2017).

There are multiple structural and positional features in pri-miRNAs that define the DROSHA cleavage site and its cleavage efficiency (***Figure 7***):

- The **basal junction** is the dsRNA-ssRNA junction at the basal end of the stem (i.e., the opposite end from the loop).
- The **apical junction** is the dsRNA-ssRNA junction at the apical end of the stem (i.e., where the loop starts).
- The **lower stem** is the part of the stem that is located between the basal junction and the DROSHA cleavage sites.
- The **basal UG motif** is a UG sequence at the basal junction of the 5p arm.
- The **mismatched GHG motif** (mGHG) is a complex motif located in the lower stem. Despite its name, which is kept for historical reasons, it need not be a GHG sequence and it is not defined by an exact sequence. Instead, it is defined by the sequences of two trinucleotides that usually face each other at a specific position in the lower stem, one in each arm. Their composition affects DROSHA cleavage site determination and cleavage efficiency. Currently, mGHG motifs are best defined by an "mGHG score", which is based on experimental data from (Fang and Bartel, 2015). The mGHG scores range from 0 to 100 and quantify how efficiently DROSHA cleaves a pri-miRNA that harbors each possible pair of trinucleotides at the mGHG position (Kwon et al., 2019).
- The **apical UGUG motif** is a UGU or GUG at the apical junction.
- The **downstream CNNC motif** is a CNNC sequence downstream of the 3' end of the stem-loop.

According to structural evidence, DROSHA recognizes the basal junction, the basal UG motif (if present), and the lower stem, with different affinities for different mGHG sequences (Partin et al., 2020). On the other hand, DGCR8 binds to the apical end of the stem, the apical loop, and the UGUG motif (if present). Together, DROSHA and DGCR8 form a "molecular ruler" that only accommodates stems that are ~35 bp long. Finally, the CNNC motif is recognized by the accessory protein SRSF3, but this is not always required for correct pri-miRNA processing (Kim et al., 2021).

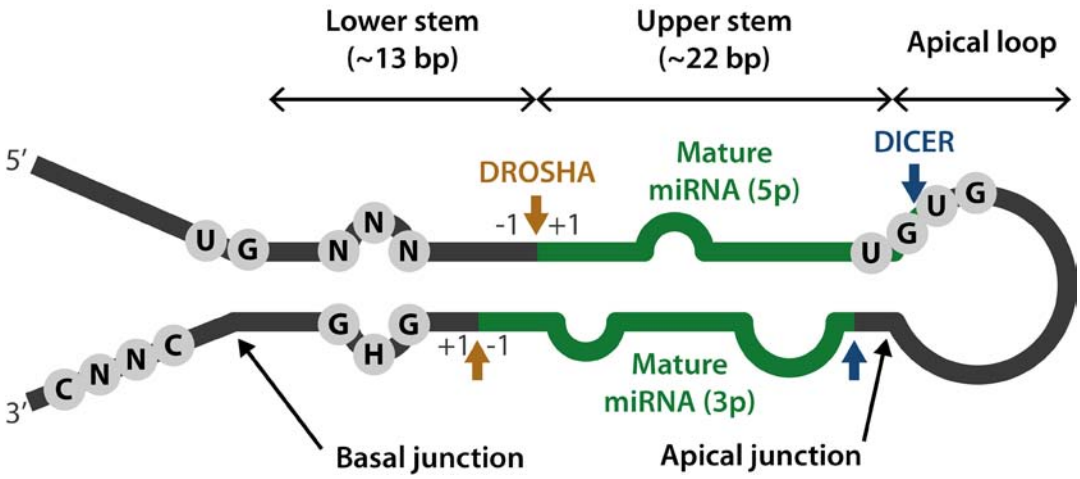


Figure 7. Structure and sequence features of pri-miRNAs. *Pri-miRNAs form hairpin structures whose stem is ~35 bp long. DROSHA and DICER cleavage sites are indicated with brown and blue arrows, respectively.*

Not all sequence motifs are present in all pri-miRNAs. In fact, most pri-miRNAs only have one or two motifs, and many pri-miRNAs have none (Auyeung et al., 2013). However, pri-miRNAs that lack all sequence motifs may not generate biologically active mature miRNAs because they are not cleaved efficiently by Microprocessor (Kim et al., 2021). According to current models, the minimum requirements for DROSHA to cleave a pri-miRNA are as follows (Kim et al., 2021; Roden et al., 2017):

- A hairpin-like structure whose stem is ~33-39 bp long.
- A stable lower stem (< 4 mismatches).
- At least one sequence motif.

If DROSHA is to cleave a pri-miRNA, the exact cleavage site is determined by one of the following features, by order of preference (Kwon et al., 2019):

- A **strong mGHG motif** (mGHG score ≥60), if present, establishes the DROSHA cleavage site at position +7 counting from the first mGHG nucleotide in the 5p arm.
- Otherwise, the **basal UG motif** and the **basal junction** define the DROSHA cleavage site. DROSHA cleaves at 14 nt after the U of the UG motif, or at 13 nt after the basal junction. Most times, the UG motif is exactly at the basal junction and the two rules are equivalent. However, occasionally the UG motif is not exactly at the basal junction. In such

cases, DROSHA may follow both rules simultaneously, generating heterogeneous products.

Finally, the following features improve DROSHA cleavage efficiency (Kim et al., 2021; Kwon et al., 2019):

- The **basal UG, apical UGUG**, and **downstream CNNC** motifs help Microprocessor bind to the pri-miRNA in the correct orientation and increase cleavage efficiency.
- **mGHG motifs** with higher mGHG scores lead to higher cleavage efficiencies.

As hinted above, motifs can be defined by "positional" criteria (i.e., position of the motif relative to DROSHA cleavage sites) or by "structural" criteria (i.e., position of the motif relative to structural features) (***Table 2*** and ***Figure 7***). Although both definitions are usually equivalent because the positions of the structural features are conserved, there are numerous exceptions (Kwon et al., 2019). In such cases, which criterion should prevail? Several arguments can be made in favor of each criterion:

- The positional definition of motifs is strongly supported by evolutionary conservation data (Auyeung et al., 2013) and by positional enrichment of motifs in pri-miRNAs that are correctly processed compared to those that are not (Kim et al., 2021).
- However, mechanistically, Microprocessor measures physical distances, and the truly conserved feature is the physical distance, not the nucleotide distance, between DROSHA cleavage sites and sequence motifs. The relationship between physical distance and nucleotide distance depends on the presence of mismatches, bulges, and loops. As a consequence, some motifs and structural features are not exactly at their expected base pair position, but they are still functional (Kwon et al., 2019). In addition, basal junctions and UG motifs can be slightly shifted from their optimal position if a strong mGHG motif is present (Kwon et al., 2019).

Table 2. Key DROSHA recognition features in pri-miRNAs as defined by positional and by structural criteria.

Feature	Positional definition	Structural definition
Basal junction	-13 nt from 5p DCS.	dsRNA-ssRNA junction at basal end.
Apical junction	+22 nt from 5p DCS.	dsRNA-ssRNA junction at apical end.
Basal UG	-14 nt from 5p DCS.	UG at basal junction.
mGHG	-7...-5 nt from 5p DCS, and opposite nt in 3p arm.	Same as positional.
Apical UGUG	UGU/GUG at +21...+25 nt from 5p DCS.	UGU/GUG at apical junction.
Downstream CNNC	+16...+18 nt from 3p DCS.	+5...+6 (up to +3...+11) from 3p basal junction.

mGHG: "mismatched GHG". DCS: DROSHA cleavage site. Distances are expressed in nucleotides (nt) and they express the range where the first nucleotide of the motif should start. Sources: (Auyeung et al., 2013; Kim et al., 2021; Kwon et al., 2019; Roden et al., 2017). See also ***Figure 7****.*

1.7.3. MicroRNA alterations in cancer

MicroRNAs are key regulators of gene expression that often target cancer-related mRNAs (Arenas et al., 2022; Goodall and Wickramasinghe, 2021). A single miRNA typically has a large number of targets, and these can be enriched in either oncogenes or tumor suppressor genes. When a miRNA preferentially targets oncogenes, it acts as a tumor suppressor; when it preferentially targets tumor suppressor genes, it acts as an "oncomiR" (cancer-promoting miRNA). As a result, expression of miRNAs is sometimes altered in cancer, so that oncomiRs are upregulated and tumor suppressor miRNAs are downregulated (Arenas et al., 2022; Goodall and Wickramasinghe, 2021).

Alterations in miRNA expression in cancer can have several degrees of complexity. In some cases, a single miRNA, such as miR-21 or miR-155, is so strongly oncogenic that its overexpression alone can drive tumorigenesis in mice (Costinean et al., 2009; Medina et al., 2010). In addition, some miRNAs, such as the miR-34 family, the let-7 family, and the miR-15/16 cluster, are strong tumor suppressors that become recurrently inactivated in various cancer types (Arenas et al., 2022; Goodall and Wickramasinghe, 2021). However, most frequently, when a single type of miRNA binds to a target mRNA, the effect on mRNA expression is mild (Bracken et al., 2016). Instead, multiple alterations in complex miRNA-mRNA regulatory networks can converge in a large effect on a single pathway (Bracken et al., 2016).

To date, most research on the role of miRNAs in cancer has focused on alterations in miRNA expression and their association with genomic copy number alterations, promoter hypermethylation, or alterations in proteins involved in miRNA biogenesis (Arenas et al., 2022; Bracken et al., 2016; Goodall and Wickramasinghe, 2021). However, point mutations that affect miRNA targeting or biogenesis remain underexplored (Urbanek-Trzeciak et al., 2020). Currently, the best known miRNA that accumulates driver point mutations in cancer is *MIR142*, a tumor suppressor in B-cell non-Hodgkin lymphomas (Kwanhian et al., 2012; Rheinbay et al., 2020). In particular, *MIR142* recurrently harbors loss-of-function mutations that affect its seed (and, hence, its targeting) or that impair the biogenesis of mature miR-142 (Kwanhian et al., 2012).

1.8. Splicing in cancer

1.8.1. Conserved sequences determine splicing

Most protein-coding genes, as well as many lncRNA genes, contain introns. Introns are removed from primary transcripts by the spliceosome ribonucleoprotein complex (Sibley et al., 2016). The spliceosome accurately recognizes exon-intron junctions, cleaves off introns, and joins exons together. This accuracy is largely thanks to specific sequences at exon-intron boundaries (***Figure 8***). In particular, the most constrained nucleotides are the first and last two intronic nucleotides, which in >98.5% of the cases in humans are GT and AG, respectively (Sibley et al., 2016). In the context of variant annotation, splice sites are usually defined as these first and last two intronic nucleotides, whereas splice regions are defined as a broader region that contains splice sites and nearby intronic nucleotides (McCarthy et al., 2014). In this text, we will use these definitions of splice sites and splice regions.

Despite their importance, GT/AG sequences are insufficient to define exon-intron junctions. Indeed, other intronic and exonic sequences help determine splice donor and acceptor sites (***Figure 8***). Regarding splice donor sites, the last exonic position has a strong preference for G, the third intronic position has a strong preference for A or G, and the fifth intronic position has a strong preference for G (Sibley et al., 2016). As for splice acceptor sites, two loosely defined intronic sequences known as the branch point and the polypyrimidine tract are required (De Conti et al., 2013). Furthermore, additional sequences deep within exons and introns may act as splicing enhancers and silencers (De Conti et al., 2013). Overall, the full picture of *cis*-acting elements that determine splicing is more complex than just the splice donors and acceptors.

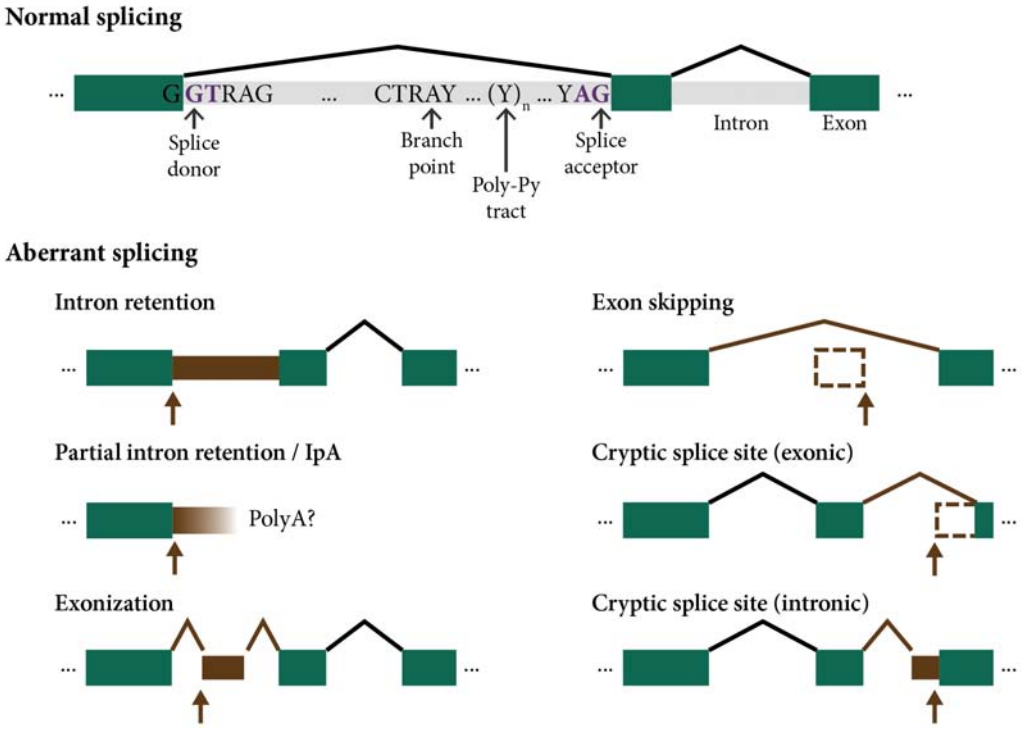


***Figure 8. Normal and aberrant splicing**. **Above:** sequence features required for correct splicing. Exons are displayed in green and introns are displayed in light grey. Black lines represent the sites that are joined together by the spliceosome. The most constrained nucleotides in the splice donor site and the splice acceptor site are highlighted in purple. Poly-Py: polypyrimidine; R: A or G; Y: C or T. **Below:** examples of splicing aberrations caused by hypothetical variants (brown arrows) in a hypothetical transcript. The alterations are highlighted in brown. IpA: intronic polyadenylation.*

1.8.2. Functional effects of variants at splicing sequences

Correct splicing is crucial for generating functional gene products. In cancer, somatic variants can disrupt splice sites or other sequences that determine splicing efficiency in *cis*, or they can create novel splice sites (Shiraishi et al., 2018). Approximately 60% of variants at splice sites and 34% of variants at splice regions have a detectable effect on RNA splicing patterns in *cis* (Calabrese et al., 2020).

The most frequent effects of splice site variants on RNA include the following (Shiraishi et al., 2018) (***Figure 8***):

- **Intron retention**. The full mutated intron is retained.
- **Exon skipping**. The affected exon is not included in the mature transcript, and instead its two neighboring exons are joined together.
- **"Partial intron retention" and intronic polyadenylation**. A mutated splice donor is expressed along with the first few intronic nucleotides, but the full intron is not retained (Andrades et al., 2022). Instead, sequencing coverage drops to zero or near zero at some point within the intron. This may be caused by intronic polyadenylation (Zhao et al., 2021b). According to this model, the variant at the splice donor initially causes intron retention, but a polyadenylation signal within the intron causes the transcript to get terminated prematurely. Alternatively, this coverage pattern may reflect that sequencing coverage decreases along introns, or it may reflect transcription-coupled processes (Sibley et al., 2016). Therefore, in this text, we prefer using the term "partial intron retention" to refer to these aberrant coverage patterns, and we only use "intronic polyadenylation" when we find additional evidence supporting such a process. In particular, if most sequencing reads spanning the splice junction are mutant and the intron contains polyadenylation signals, intronic polyadenylation is a likely cause.
- **Usage of cryptic splice sites**. Alternative splice donor or acceptor sites are used. If the cryptic splice site is located within an exon, the mature transcript will contain a deletion; if it is located within an intron, the mature transcript will contain an insertion.
- **Creation of novel exons (exonization)**. An intronic sequence may be recognized as an exon by the spliceosome, especially when a variant creates a novel intronic splice site (Calabrese et al., 2020).
- **Usage of alternative isoforms**. This is a special case of any of the effects mentioned above (Andrades et al., 2022). It occurs when the transcript generated by the splice site variant is another canonical isoform of the gene. For example, one isoform of a gene may include an exon and another one may skip it. If a splice site variant causes skipping of that exon, the resulting transcript will still be canonical, not aberrant.

1.8.3. Splice site variants in cancer

Variants at splice sites of PCGs are the best studied non-coding variants in cancer because they can be detected relatively easily, they are sometimes recurrent, and they usually cause major aberrations in gene products. For example, splice site variants can cause alterations in RNA processing that shift the reading frame in CDSs, remove of parts of a gene product, or terminate transcripts prematurely (Andrades et al., 2022; Shiraishi et al., 2018). Sometimes, the aberrant transcripts are degraded by nonsense-mediated decay, which can further impair gene function (Jung et al., 2015). As a consequence, splice site variants are most frequent in tumor suppressor genes, such as *TP53* (Shiraishi et al., 2018). For example, in lung adenocarcinoma, the tumor suppressor genes *TP53* and *SMARCA4* are recurrently inactivated by splice site variants (Bouaoun et al., 2016; Peinado et al., 2022).

Alternatively, splice-altering variants can affect oncogenes. These cases are relatively rare because they must maintain the oncogenic activity of the protein while disrupting domains that promote its degradation or inactivation. Nevertheless, these alterations are extremely important at a clinical level. For example, in ~3-4% of lung adenocarcinoma tumors, splice-altering variants cause the exon 14 of the *MET* oncogene to be skipped while retaining the reading frame (Frampton et al., 2015). The aberrant MET protein lacks a degradation domain and, therefore, it has increased oncogenic activity. Importantly, splice site mutant MET is targeted by capmatinib and tepotinib, two specific inhibitors that have been approved for the treatment of metastatic non-small cell lung cancer (Mathieu et al., 2022).

Splicing can also be altered in *trans*. In particular, RNA and protein components of the spliceosome, including *U2AF1*, *SF3B1*, *SRSF2*, and the *U1* small nuclear RNA, are recurrently mutated in multiple cancers (Seiler et al., 2018; Shuai et al., 2019; Suzuki et al., 2019). The mutations cause transcriptome-wide splicing alterations that affect oncogenes and tumor suppressor genes. In general, the effects on each individual gene are thought to be mild, but the combined effects on all genes may be quantitatively relevant. Nevertheless, further research is needed to determine the how these transcriptome-wide splicing alterations promote oncogenesis.

1.9. Promoters in cancer

Promoters are regulatory DNA sequences in which transcription of genes is initiated. Promoters contain sequence motifs that are specifically recognized by various proteins, including RNA polymerases and transcription factors, which mediate gene transcription (Messeguer et al., 2002). Therefore, variants in promoter sequences may affect gene expression by reducing or increasing the binding affinity of transcription factors (Calabrese et al., 2020).

A major landmark in the study of non-coding driver mutations in cancer was the discovery of recurrent point mutations in the *TERT* promoter (Fredriksson et al., 2014; Horn et al., 2013; Huang et al., 2013). The *TERT* gene encodes the reverse transcriptase subunit of the telomerase complex, which maintains telomere length and prevents cell senescence. Recurrent mutations in the *TERT* promoter create binding sites for ETS transcription factors, increasing TERT expression. *TERT* promoter mutations are among the most recurrent driver mutations in human cancer and they have been described in over a dozen cancer types (Fredriksson et al., 2014; Rheinbay et al., 2020).

Promoters other than *TERT* harbor cancer driver mutations, but never at such high recurrence or across so many cancer types. For example, driver mutations in the promoters of *PLEKHS1*, *SDHD*, *TFPI2*, and *FOXA1* affect transcription factor binding and/or gene expression in *cis* (Fujimoto et al., 2016; Melton et al., 2015; Rheinbay et al., 2017). Another well-studied case is the promoter of *WDR74* (Khurana et al., 2013; Nik-Zainal et al., 2016; Weinhold et al., 2014), but its reported variants may not be real because the region is prone to mapping artifacts (Rheinbay et al., 2020). Recurrent variants in other promoters have been reported, but their functional significance has not been explored in detail (Weinhold et al., 2014).

Promoters of lncRNAs may also accumulate somatic variants, but they have rarely been studied. In multiple cancer types, the promoters of *RMRP* and *NEAT1* harbor somatic variants that affect transcription factor binding and alter expression levels (Rheinbay et al., 2020; Rheinbay et al., 2017). However, both loci show patterns of non-specific mutational processes and the *RMRP* locus is prone to mapping artifacts (Rheinbay et al., 2020).

1.10. UTRs in cancer

Both 5'- and 3'-UTRs are involved in the post-transcriptional regulation of gene expression, containing elements that can affect the stability, the translation efficiency, and the subcellular localization of mRNAs (Gruber and Zavolan, 2019; Schuster and Hsieh, 2019). As a consequence, UTRs may be altered in cancer.

The most studied UTR alterations in cancer are those in 3'-UTRs. 3'-UTRs often contain miRNA binding sites and other regulatory sequences that shorten the half-life of mRNAs (Mayr and Bartel, 2009). As a consequence, shortening or mutating 3'-UTRs of certain oncogenes can be beneficial to cancer cells. The mechanisms by which 3'-UTRs are altered in cancer include:

- **Alternative polyadenylation**. Generally, 3'-UTRs are shortened at a transcriptome-wide level in proliferating cells and in malignant cells, even in the absence of mutations in *cis* (Mayr and Bartel, 2009). 3'-UTRs may be shortened because proliferating and malignant cells tend to use alternative polyadenylation sites located earlier within the transcript (Mayr and Bartel, 2009). Although this phenomenon may increase the protein levels of certain oncogenes, some researchers have questioned whether it can promote oncogenesis when considered at a transcriptome-wide level, as shorter 3'-UTRs do not always increase the half-life of the mRNA and the global effect may be weak (Gruber and Zavolan, 2019). In fact, it is still unclear whether shortened 3'-UTRs are a cause or a consequence of malignant transformation (Gruber and Zavolan, 2019).
- **Alternative splicing**. Certain introns and exons in UTRs can be either included or excluded in the mature mRNA, generating UTRs of varying lengths (Chan et al., 2022). Together, alternative splicing and alternative polyadenylation constitute the two major mechanisms by which 3'-UTRs are generally shortened in cancer. Inhibition of 3'-UTR splicing decreases the expression of oncogenes and has an anti-tumor effect, supporting a causative role of 3'-UTR splicing in oncogenesis (Chan et al., 2022).

- **Point mutations and deletions**. For example, in lymphoma, the 3'-UTR of the oncogene *CCND1* recurrently suffers deletions and point mutations that induce its premature cleavage and polyadenylation (Wiestner et al., 2007). In addition, pan-cancer driver point mutations in the 3'-UTR of *TOB1* are associated with a decrease in *TOB1* mRNA expression, whereas those in the 3'-UTR of *NFKBIZ* in lymphoma are associated with an increase of its mRNA (Rheinbay et al., 2020). Furthermore, recurrent mutations in the 3'-UTR of *NOTCH1* in chronic lymphocytic leukemia cause a deletion in its CDS by promoting the usage of a cryptic splice site, removing a protein domain that mediates its degradation (Puente et al., 2015).
- **Translocations**. For example, in various cancer types, the *HMGA2* oncogene is targeted by chromosomal translocations that swap its 3'-UTR with that of various other genes, allowing it to escape repression by the let-7 family of tumor suppressor miRNAs (Mayr et al., 2007). Interestingly, the translocation partners that receive the *HMGA2* 3'-UTR are sometimes tumor suppressor genes, thus strengthening the oncogenic effect of the translocation.

Although 5'-UTRs, especially those of oncogenes, contain various regulatory elements, reports on 5'-UTR alterations in cancer are scarce (Schuster and Hsieh, 2019). For example, in prostate cancer, the 5'-UTR of *TMPRSS2* is recurrently fused with genes such as *ERG* and *ETV1*, causing their overexpression (Tomlins et al., 2005). Moreover, a recurrent germline point mutation in the 5'-UTR of the tumor suppressor gene *CDKN2A* decreases CDKN2A protein expression and predisposes to melanoma (Liu et al., 1999). In particular, the mutation creates an upstream translation start site that impairs translation of the actual *CDKN2A* open reading frame. Furthermore, pan-cancer driver point mutations in the 5'-UTR of *MTG2* decrease the expression of its mRNA (Rheinbay et al., 2020). Finally, it has been suggested that cancer cells may transcribe oncogenes from alternative downstream transcription start sites, shortening their 5'-UTRs and thereby removing regulatory elements, but evidence in real tumors is currently scarce (Schuster and Hsieh, 2019).

Chapter 2. Objectives

The general objective of this thesis was to evaluate the presence of novel cancer-promoting somatic mutations in the non-coding genome of various cancer cohorts, with special focus on lncRNAs, miRNAs, and intronic splice regions. The specific objectives were:

1. To develop computational pipelines to identify high-confidence somatic variants in targeted sequencing data of genomic DNA from cancer samples, with or without matched normal tissue samples.
2. To search for novel non-coding cancer drivers in lung adenocarcinoma datasets, with special focus on lncRNAs, miRNAs, and intronic splice regions.
3. To computationally predict whether candidate non-coding drivers and their variants are functional.
4. To develop a miRNA-centric pipeline for annotating variants in miRNA genes, with special focus on identifying variants that affect seed sequences and DROSHA processing motifs.
5. To search for novel cancer-promoting non-coding variants at intronic splice sites in external diffuse large B-cell lymphoma datasets, and to evaluate the impact of the identified recurrent splice site variants on RNA processing.

Chapter 3. Non-coding mutations in lung adenocarcinoma

This Chapter addresses ***Objectives 1-4***. Here, we describe our search for non-coding drivers in in-house targeted sequencing data of genomic DNA from lung adenocarcinoma samples. The sequencing was directed towards all human miRNA genes, exons of a selection of lncRNAs, and exons of a selection of PCGs. The analysis was complemented with data from external sources, especially from The Cancer Genome Atlas (TCGA) / Pan-Cancer Analysis of Whole Genomes (PCAWG). To contextualize our work, we begin this Chapter with an introduction on the clinical and molecular features of lung adenocarcinoma.

3.1. Background: lung adenocarcinoma

3.1.1. Epidemiology, classification, and clinical characteristics of lung cancer

Lung cancer is the deadliest and the second most diagnosed cancer worldwide. In 2018, 2.1 million people were diagnosed with lung cancer (13% of all cancers) and 1.8 million people died from it (19% of all cancer deaths) (Wild et al., 2020). Lung cancer is strongly associated with exposure to tobacco smoke, but up to ~15% of cases are not caused by it (Wild et al., 2020).

Lung cancer is broadly classified as small cell lung cancer (~15% of lung cancers) and non-small cell lung cancer (NSCLC, ~85% of lung cancers) (Barta et al., 2019; Wang et al., 2021) (***Figure 9***). Within NSCLC, the most common subtype is lung adenocarcinoma (LUAD), which constitutes up to ~60% of all NSCLC cases, followed by squamous cell carcinoma (~30%), and other minor subtypes (~10%). Although a more sophisticated histological classification of lung cancer has been adopted in recent years (Nicholson et al., 2022), for the purposes of this text we will use the simpler classification.

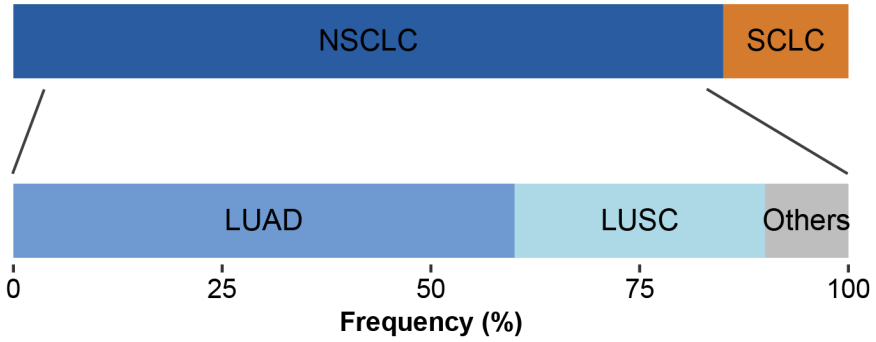


Figure 9. Frequencies of the main lung cancer subtypes. *NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma. Data source: (Wang et al., 2021).*

The 5-year survival rate of lung cancer patients is dismal, being only 10-20% (Wild et al., 2020). Still, novel screening techniques for early diagnosis and novel therapies have improved the outcome of some groups of lung cancer patients in the last years. For example, targeted therapies have been developed for patients who harbor specific molecular alterations, such as *EGFR* mutations, *ROS1* fusions, *ALK* translocations, and *MET* activation (Wang et al., 2021). In addition, immunotherapy may improve the survival of a subset of lung cancer patients. Nevertheless, fewer than 25% of NSCLC patients can be treated with a currently approved targeted therapy, and tumors eventually become resistant (Wang et al., 2021). Taken together, these facts highlight the urgent need for better clinical approaches against lung cancer, which may be achieved by a better understanding of the underlying molecular mechanisms of the disease.

LUAD is the most frequent subtype among all lung cancers and among never-smokers (Wild et al., 2020). Generally, LUAD develops in mucus-secreting cells in outer parts of the lung, especially in alveoli. A typical LUAD tumor is usually complex, consisting of a mixture of different types of cells with heterogeneous histological and genetic characteristics (Wild et al., 2020). Despite the complexity of LUAD tumors, they harbor recurrent driver mutations in a small collection of PCGs, some of which can be targeted by current or by developing therapies (Wang et al., 2021).

3.1.2. Driver genes in lung adenocarcinoma and the emerging role of the non-coding genome

3.1.2.1. The single-driver model and its limitations

A large number of driver alterations have been identified in LUAD, and all of them affect protein-coding genes (Bailey et al., 2018; Collisson et al., 2014). Currently, the prevailing model of LUAD driver alterations is a "single-driver" model (Skoulidis and Heymach, 2019). According to this model, each LUAD tumor has one "main" driver event, which may affect genes such as *KRAS*, *EGFR*, *BRAF*, *MET*, and others (***Figure 10***). These "main" driver events are mostly mutually exclusive and, therefore, they can be used to classify LUAD tumors at a molecular level. In addition, targeted therapies are available or under development for some of these driver events. However, the single-driver model has several issues (Skoulidis and Heymach, 2019):

- It does not account for other "secondary" drivers that may co-occur and interact with the "main" drivers. In fact, a typical tumor has, on average, 5 driver mutations (Campbell et al., 2020). As a consequence, groups of LUAD tumors that share the same "main" driver can be highly heterogeneous from a phenotypical and clinical perspective (Skoulidis and Heymach, 2019). This heterogeneity may be dictated by "secondary" driver alterations in genes such as *TP53*, *STK11*, *CDKN2A*, or *CDKN2B*, whose relevance in LUAD is well-established.
- Approximately 1 in every 4 LUAD tumors has none of the "main" driver alterations, leaving them unclassified. Although these tumors may have alterations in "secondary" drivers, they may not be enough to explain the tumor phenotype by themselves.

The shortcomings of current models of LUAD drivers suggest that the spectrum of driver mutations in LUAD tumors has not been fully elucidated yet. However, researchers are struggling to complete the puzzle. One possible reason why the catalog of LUAD drivers is incomplete is that most research has focused on coding sequences, which constitute ~1.1% of the human genome (Nurk et al., 2022). Therefore, some of the "missing" LUAD drivers may reside within the non-coding genome, which remains underexplored.

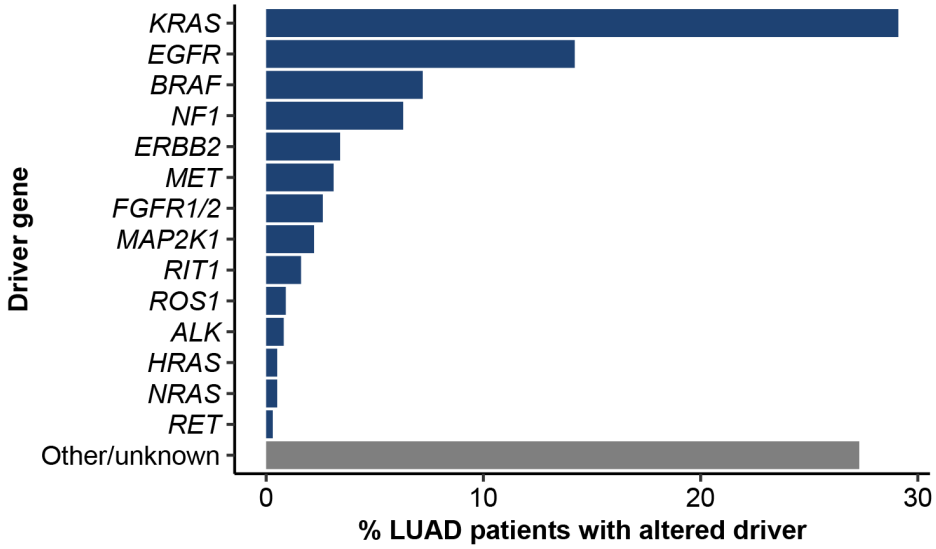


Figure 10. Frequency of "main" driver alterations in lung adenocarcinoma (LUAD). *The statistics include point mutations, genomic rearrangements, amplifications, and deletions. Data from (Skoulidis and Heymach, 2019).*

3.1.2.2. Recent efforts to identify non-coding LUAD drivers

As detailed in ***Section 1.5.2***, most research on LUAD driver mutations has ignored non-coding DNA for various technical and conceptual reasons. As a consequence, most studies have only sequenced exons of PCGs, either by whole-exome sequencing (WES) or by targeted sequencing of pre-selected sets of PCGs that are likely to be implicated in LUAD (Collisson et al., 2014). However, as research on non-coding DNA gained popularity and the required experimental techniques became more feasible, major efforts were made to perform WGS in LUAD and in other cancers.

To date, the two largest pan-cancer WGS projects are those of the PCAWG Consortium and the Hartwig Medical Foundation. PCAWG, which is an extension of the TCGA project, analyzed 2658 whole genomes of primary tumors from various tissues in search for non-coding drivers (Campbell et al., 2020; Rheinbay et al., 2020). On the other hand, the Hartwig Medical Foundation led similar efforts in metastatic tumors (Priestley et al., 2019). Furthermore, databases such as COSMIC compile non-coding variants from multiple studies (Tate et al., 2019). However, the number of LUAD tumors was limited both in PCAWG (N = 38 LUAD primary tumors) and in the

Hartwig study (N = 143 metastatic NSCLC tumors, subtypes not specified) (Campbell et al., 2020; Priestley et al., 2019; Rheinbay et al., 2020). To our knowledge, there are no LUAD-specific reports of non-coding drivers because relatively few LUAD whole genomes have been published.

To detect mutations in the non-coding genome, an alternative to WGS is targeted sequencing. In this technique, a set of pre-defined regions of interest are captured and sequenced. Therefore, targeted sequencing is more manageable than WGS, but it requires a prior definition of target regions. For example, miRNA genes and lncRNA exons can be of interest because they are implicated in the regulation of gene expression and, therefore, they might harbor driver mutations (***Sections 1.6-1.7***). Furthermore, targeted sequencing also captures the first ~100-200 nt that flank the regions of interest, albeit at lower coverage. As a result, targeted sequencing of exons also allows for the identification of variants in intronic splice regions and in proximal promoters.

In this work, we have analyzed a collection of 70 LUAD primary tumor samples, 27 matched normal adjacent tissue samples, and 37 LUAD cell lines by targeted sequencing of all human miRNA genes, exons of a pre-selected set of cancer-related lncRNAs, and exons of a selection of PCGs in search for novel non-coding LUAD drivers, expanding previous cohort sizes. The main focus of our work has been variants in exons of miRNAs and lncRNAs as well as variants in intronic splice regions. Moreover, we have analyzed variants in proximal promoters and UTRs. Furthermore, we have complemented our results by reanalyzing WGS data from TCGA-LUAD focusing on our sequencing targets of interest. Finally, we have tested the putative biological role of the top driver candidates using publicly available external data.

3.2. Materials and methods

3.2.1. Software and online tools

Table 3. Software and computational tools used in our work.

Software	Version	URL / citation
ANNOVAR	2017-07-17	https://annovar.openbioinformatics.org
BWA	0.7.13-r1126	(Li and Durbin, 2009)
CADD	1.6	https://cadd.gs.washington.edu
Cutadapt	-	(Martin, 2011)
FastQC	0.11.5	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
FATHMM-MKL	2.3	https://fathmm.biocompute.org.uk/
GATK	≥3.8.10	https://gatk.broadinstitute.org
gdc-client	1.3.0	https://github.com/NCI-GDC/gdc-client
IGV	2.3.94	https://software.broadinstitute.org/software/igv
MAJIQ	2.2-e25c4ac	(Vaquero-Garcia et al., 2016)
miRDB	-	http://www.mirdb.org/custom.html
MuTect2	4.0.3.0	https://gatk.broadinstitute.org
OncoDriveCLUSTL	1.1.4	(Arnedo-Pac et al., 2019)
OncoDriveFML	2.3.0	(Mularoni et al., 2016)
Picard	2.9.1	https://broadinstitute.github.io/picard/
PROMO	3.0.2	http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
Qualimap	2.2.1	(Okonechnikov et al., 2015)
R	≥ 3.5.2	https://www.r-project.org/
RNAsnp	1.2	https://rth.dk/resources/rnasnp/software.php
samtools (+ bcftools & htslib)	≥ 1.7	http://www.htslib.org/
Strelka2	2.9.10	(Kim et al., 2018)
TargetScan	5.2	http://www.targetscan.org/vert_50/seedmatch.html
VarScan2	2.4.3	(Koboldt et al., 2012)

Analyses that required high computational resources were performed in either of the following high-performance computing clusters:

- ALHAMBRA (University of Granada), https://alhambra.ugr.es.
- UBELIX (University of Bern), https://ubelix.unibe.ch.
- Cluster of the Cancer Research Center of Salamanca (CSIC/USAL).

3.2.2. External resources

Table 4. External resources used in our work.

Resource	Version	URL / reference
CADD	v1.5	https://cadd.gs.washington.edu/
Cancer Gene Census	COSMIC v90	https://cancer.sanger.ac.uk/census
Cancer LncRNA Census	v2	(Vancura et al., 2021)
COSMIC non-coding variants	v90	https://cancer.sanger.ac.uk/cosmic/download
dbSNP	150	https://www.ncbi.nlm.nih.gov/snp
ExAc	v3	https://exac.broadinstitute.org
GENCODE	v29	https://www.gencodegenes.org
gnomAD	3.0	https://gnomad.broadinstitute.org
Human genome	hg38	ftp://ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.fasta.gz
Human genome	hg19	ftp://ftp.broadinstitute.org/pub/seq/references/Homo_sapiens_assembly19.fasta
Known indels	-	ftp://ftp.broadinstitute.org/bundle/hg38/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz
miRBase	21	https://www.mirbase.org
MirGeneDB	2.1	https://mirgenedb.org/
phyloP (100-way)	-	https://genome-euro.ucsc.edu
RepeatMasker	-	https://genome-euro.ucsc.edu
TarBase	v8	https://dianalab.e-ce.uth.gr/html/diana/web/index.php

The annotation from GENCODE v29 is equivalent to Ensembl v94/v95. The hg38 human genome was used unless otherwise specified. Annotations and variants were converted between hg19 and hg38 human genome versions using the R package `liftOver`. Converting our sequencing targets from hg38 to hg19 resulted in the loss of 4 out of 18 000 targets (*MIR6859-2*, *MIR1234*, *MIR4477A*, and *LINC00850*). To compare sets of genes, Ensembl gene identifiers were used unless otherwise specified.

3.2.3. Sample acquisition

We acquired samples from 70 LUAD primary tumors and their matched normal adjacent tissues, as well as 37 LUAD cell lines, as we recently described (Peinado et al., 2020; Peinado et al., 2022).

3.2.3.1. Patient samples

DNA and RNA from 70 LUAD tumors and their matched normal adjacent tissues were obtained from the Basque Biobank (www.biobancovasco.org) and they were processed following standard operating procedures. All samples had been acquired in the Basque Country region of Spain. The study was approved by the Research Ethics Committee of Granada (CEI Granada, Department of Health, Regional Government of Andalusia, Spain) and by the Basque Foundation for Health Innovation and Research (Spain). Signed informed consent, following the procedures of the Declaration of Helsinki and institutional and national guidelines, was obtained from all participants.

LUAD patients were diagnosed from August 2008 to January 2016 and an independent experienced pathologist confirmed all diagnoses via pathological examinations. The inclusion criteria were 1) histological diagnosis of lung adenocarcinoma, 2) availability of demographic and clinical data, 3) availability of DNA and RNA samples, and 4) provision of signed informed consent. Clinical information is summarized in ***Table 5***. Treatment was highly heterogeneous: it could involve any combination of surgery, radiotherapy, and chemotherapy with or without neoadjuvant. Purity of all samples was estimated to be ≥70% according to an independent pathologist.

Table 5. Summarized clinical characteristics of our LUAD cohort.

Clinical variable	**Value**	
Sex	Male	50 (71%)
	Female	20 (29%)
Age at diagnosis (years)	Median [range]	62 [46-80]
Stage (T)	T1	15 (21%)
	T2	40 (57%)
	T3/T4	9 (13%)
	Not available	6 (9%)
Stage (N)	N0	35 (50%)
	N1	10 (14%)
	N2	7 (10%)
	Not available	18 (26%)
Stage (M)	M0	14 (20%)
	M1	1 (1%)
	Not available	55 (79%)
Smoking status	Current smoker	31 (44%)
	Ex-smoker	30 (43%)
	Non-smoker	7 (10%)
	Not available	2 (3%)
Vital status	Alive	42 (60%)
	Deceased	28 (40%)
Relapse	No	43 (61%)
	Yes	27 (39%)
Overall survival (months)	Median [range]	61.5 [2-135]
Disease-free survival (months)	Median [range]	49.5 [0-92]

3.2.3.2. Cell lines

Thirty-seven LUAD cell lines were grown under standard conditions (37°C, 5% carbon dioxide) in DMEM or RPMI 1640 medium supplemented with glutamine, 10% fetal bovine serum and 1% penicillin/streptomycin/ amphotericin. The characteristics of the patients from whom the cell lines were retrieved are summarized in ***Table 6***.

Table 6. Summarized clinical characteristics of the patients from whom our LUAD cell lines were derived.

Clinical variable	**Value**	
Sex	Male	21 (57%)
	Female	14 (38%)
	Not available	2 (5%)
Age at diagnosis (years)	≤45	6 (16%)
	>45	23 (62%)
	Not available	8 (22%)
Disease subtype	Adenocarcinoma / carcinoma	32 (86%)
	Bronchoalveolar carcinoma	4 (11%)
	Papillary adenocarcinoma	1 (3%)
Tissue	Lung	16 (43%)
	Metastatic site	17 (46%)
	Not available	4 (11%)
Smoking status	Smoker	15 (41%)
	Non-smoker	7 (18%)
	Not available	15 (41%)

Sources: American Type Culture Collection (ATCC) and PubMed.

3.2.4. Gene capture and targeted DNA sequencing

Gene capture and targeted sequencing of genomic DNA were performed for the full collection of LUAD cell lines and primary tumors, as well as for 27 of the 70 matched normal samples, as previously described (Peinado et al., 2020; Peinado et al., 2022).

3.2.4.1. Selection of targets for gene capture

Gene capture was performed using a custom NimbleGen SeqCap EZ Choice Library (Roche, Inc., Madison, WI, USA). The baits for the gene capture were designed using the NimbleDesign software (Roche, v4.0). Selection of targets was performed within the limitations of the commercial kit (up to 7 Mb). At the time of the design of the experiment, Ensembl v79 was used as a reference annotation. For downstream analyses, the annotation was later updated to GENCODE v29 (Ensembl v95) (***Section 3.2.9***).

The following targets were included in the design:

- All human miRNAs from miRBase 21 (n = 1881).
- Exons from a curated list of putative cancer-related lncRNAs (n = 908). The list was obtained as follows:
    - Curated putative lung cancer-related lncRNAs profiled by the LncPath™ Human Cancer Array (Arraystar Inc.).
    - Disease-related lncRNAs from the LncRNADisease database (Chen et al., 2013).
    - Experimentally characterized lncRNAs from the lncRNAdb database (Amaral et al., 2010).
    - Manual curation of the scientific literature (PubMed search: 'lncRNA AND "lung cancer"').
- Exons from a list of putative cancer-related PCGs (n = 1307). The list was obtained as follows:
    - Curated putative lung cancer-related PCGs profiled by the LncPath™ Human Cancer Array (Arraystar Inc.).
    - Manual curation of the scientific literature.

The total number of targets was 18 000. Gene capture does not only capture the intended targets, but also the first nucleotides of flanking regions. For this reason, for downstream computational analyses, the target coordinates were padded by 200 nt at 5' and 3' (unless specified otherwise). This allowed us to analyze promoters and splice regions.

3.2.4.2. Gene capture and sequencing protocols

Gene capture was carried out following the instructions of the NimbleGen SeqCap EZ Library SR User's Guide v3.0 (Roche, Inc.) from pooled libraries prepared using the TruSeq DNA Sample Preparation Kits (Illumina, Inc., San Diego, CA, USA). Briefly, 300 ng of genomic DNA were fragmented using a Covaris S2 sonicator yielding 180-220 bp fragments. After end repair and adapter ligation, the adapter-ligated fragments were amplified by PCR (9 cycles). The PCR fragments were purified and the fragments of correct size were selected. DNA was denatured and hybridized against biotinylated probes, which were then captured using streptavidin-bound magnetic beads. The DNA bound to the beads was isolated and amplified by PCR (14 cycles). The quality and the concentration of the DNA were evaluated using NanoDrop (Thermo Scientific) and BioAnalyzer (Agilent).

The paired-end sequencing was performed on a NextSeq 500 instrument (Illumina) using a NextSeq 500/550 Mid Output Kit (Illumina), 2x150 cycles.

3.2.5. External datasets

3.2.5.1. WGS data from TCGA-LUAD

WGS alignment files (BAM format, hg19 human genome) from TCGA-LUAD were retrieved from the Legacy Portal of Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/legacy-archive, version 12.0, accessed in June 2018). The dataset included 152 patients, most of which were discarded because they were from "low-pass" WGS (≤ 10X depth). In addition, only one tumor and one normal file per patient were kept. If multiple files were present for the same sample, the one with the largest size was kept. These criteria resulted in BAM files for 59 tumor-normal pairs.

The `gdc-client` command line tool was used to download the 118 BAM files. The BAM files were filtered using `samtools view` so that they only contained reads that spanned the padded sequencing targets from our design. Then, variant calling was performed in the same way as for our paired sequencing data, but using hg19 as a reference genome (***Section 3.2.7***). Variants were converted to hg38 prior to the annotation step.

The International Cancer Genome Consortium (ICGC) Data Portal (https://dcc.icgc.org) also provided annotated variants from an analysis of 38 of the 59 TCGA-LUAD WGS samples performed by the PCAWG Consortium. PCAWG combined multiple variant calling algorithms following a similar rationale to us (Campbell et al., 2020; Rheinbay et al., 2020) (***Section 3.2.7.3***). However, only one variant calling algorithm, Mutect2, was used by both PCAWG and us. We downloaded the variant calling format (VCF) files of the 38 LUAD patients from the ICGC Data Portal ("Data Repositories" tab) and then we filtered them so that they only contained variants within our regions of interest using `tabix` (`htslib`).

3.2.5.2. WES data from TCGA-LUAD

WES data from TCGA-LUAD were downloaded from the GDC Data Portal (version 31.0, https://portal.gdc.cancer.gov) (N = 582). Alignment files in the BAM format, restricted to our regions of interest, were downloaded for coverage analyses using the `gdc-client` tool and its "BAM slicing" feature. Annotated variant files in the MAF format were directly downloaded from the GDC Data Portal. For each sample, up to 4 files were available, each of them originating from a different variant calling pipeline. We retained the variants that were annotated as "PASS" in the "FILTER" column in at least 2 of the 4 pipelines. If a patient had more than one sequenced tumor sample, we combined all unique variants found across all samples.

The targets of the exome capture design used by TCGA-LUAD were retrieved from Agilent's eArray website (https://earray.chem.agilent.com), "SureSelect Human All Exon v4" protocol.

3.2.5.3. Gene expression datasets

Clinical, transcriptome, and miRNA expression data from TCGA-LUAD were downloaded using the R package `TCGAbiolinks`. MicroRNA expression data were downloaded as "reads per million" and transformed to $\log_2$ adding a pseudocount of 1. Transcriptome profiling data were downloaded as fragments per kilobase of exon per million mapped fragments normalized by upper quartile method (FPKM-UQ). Then, they were transformed to $\log_2$, adding a pseudocount of 1. If the same patient had expression data from multiple tumor or normal samples, their expression values were averaged. For splicing analyses, filtered RNA-Seq BAM files from TCGA-LUAD were downloaded by BAM slicing via the GDC application programming interface (https://docs.gdc.cancer.gov/API/Users_Guide/BAM_Slicing). The query regions were the full splice site mutant genes of interest, padded by 200 nt.

Transcriptome, proteome, and miRNA expression data from Gillette et al were downloaded from the supplementary material of the original publication (Gillette et al., 2020). MicroRNA expression data were available as $\log_2$-transcripts per million (TPM). Transcriptome expression data were available as $\log_2$-transformed FPKM. Unfiltered protein expression data were available as $\log_2$-transformed standardized TMT ratios, where the ratios were calculated by dividing the expression of each protein in each sample by the expression of the same protein in a common reference sample. The transcriptome expression dataset only contained PCGs, and therefore it could not be used to evaluate lncRNA expression.

Transcriptome and miRNA expression data of cell lines from the Cancer Cell Line Encyclopedia (CCLE) were downloaded from the DepMap portal (https://depmap.org/portal, data version 20181103). LUAD cell lines were selected based on the annotation from DepMap.

3.2.6. Power analysis

Power analysis for the detection of driver features was performed based on a binomial power model as previously described (Lawrence et al., 2014; Rheinbay et al., 2017). In this context, a driver feature is defined as a feature that is significantly mutated in a cohort over the BMR. Although the statistical model for this power analysis is simpler than the models of the driver discovery tools used in our work, it is an adequate approximation because all driver discovery tools search for enrichments of variants above a BMR. The most critical simplification made by our power analysis model is assuming that all features of the same type have the same BMR.

The statistical power to detect a feature as a driver is defined as the probability of finding an excess of variants in the feature above the BMR. For a given feature, let p_0 denote the probability that it has at least one variant in a patient under the background model. The observed signal (i.e., frequency of variants in the feature, p_1) can be defined as:

$$p_1 = p_0 + r(1 - m)$$

Where r is the excess frequency of variants (above background) in the cohort, and m is the probability that a real variant is not detected due to technical limitations or sampling issues (a usual value is $m = 0.1$). We formulated the following null and alternative hypotheses:

- H_0: "All variants in a given feature are due to the BMR ($p_1 = p_0$)."
- H_1: "A proportion of variants in a given feature do not originate from the BMR ($p_1 \neq p_0$)."

p_0 can be estimated from the mutation rate per nucleotide in the tumor (μ), the length of the feature (L), and optional correction factor f_g:

$$p_0 = 1 - (1 - \mu f_g)^L$$

For μ, we used $\mu = 10^{-5}$ (10 mutations / Mb) based on observations in our own data (see ***Section 3.3.4***). For L, we estimated the median length of each feature type that was later studied in the driver analysis (***Table 7***). Feature types were defined as detailed in ***Section 3.2.9.1***.

The mutation rate factor, f_g, is a correction factor that accounts for the fact that each type of feature may be mutated above or below the overall μ. We estimated f_g as the ratio between the 90th percentile of the mutation rate of each feature type across all patients and μ (***Table 7***).

Using the estimated values of p_0, we then calculated p_1 for different values of r within the range 0.01-0.2. Then, in our cohort of $N = 70$ primary tumors, we estimated the maximum number of patients that were expected to be mutated in a given gene under the null model (N_{crit}) using a binomial distribution $B(N, p_0)$ and a significance threshold $\alpha = 0.25/n$, where n was the number of features of that type (for multiple testing corrections; ***Table 7***). In other words, N_{crit} was the minimum value that satisfied that:

$$P(X \leq N_{crit}) \geq 1 - \alpha, \qquad X \sim B(N, p_0)$$

Finally, the power was the probability of observing at least $N_{crit} + 1$ mutated patients under a binomial distribution $B(N, p_1)$.

Table 7. Parameters used for power analysis in each feature type.

Feature	L	*n*	f_g
CDSs	1089	1307	2.48
UTRs	1222	1444	2.01
Promoters, PCGs	400	1299	2.21
Splice regions, PCGs	80	1690	2.44
Exons, lncRNAs	598	908	3.13
Promoters, lncRNAs	400	852	2.32
Splice regions, lncRNAs	20	840	5.24

L: median length of feature type; n: number of features; f_g: mutation rate factor; CDS: coding sequence; UTR: untranslated region; PCG: protein-coding gene; lncRNA: long non-coding RNA. Features were selected and filtered as described in ***Section 3.2.9.1.***

3.2.7. DNA-Seq data analysis

Driver discovery algorithms can output gravely biased results if the input set of variants contains germline variants or sequencing artifacts. To minimize these biases, we developed in-house DNA-Seq analysis pipelines for detecting high-confidence somatic variants in paired and unpaired samples. Our pipelines were published elsewhere together with an analysis of a subset of our data (Peinado et al., 2020; Peinado et al., 2022), and they are detailed below.

3.2.7.1. Quality control and preprocessing of raw data

Quality of the raw FASTQ files was checked using FastQC. Then, Cutadapt was used to eliminate adapter sequences:

```
cutadapt -b AGATCGGAAGAGC -B AGATCGGAAGAGC \
–q 20 –m 50 \
-o trimmed.1.fastq.gz -p trimmed.2.fastq.gz \
sample.R1.fastq.gz sample.R2.fastq.gz > report.txt
```

3.2.7.2. Alignment to human genome and BAM processing

Alignment to the hg38 human genome was performed using BWA-MEM:

```
bwa mem –M –t 4 hg38.fa trimmed.1.fastq.gz \
trimmed.2.fastq.gz > aln.sam
```

The generated SAM files were processed using Picard to: (i) sort by coordinate; (ii) convert to BAM format and index; (iii) mark PCR duplicates; and (iv) check the quality of the BAM files. Furthermore, indel realignment and base quality score recalibration were performed using GATK and a source of known indels. Further BAM quality statistics were obtained using Qualimap.

3.2.7.3. Variant calling

Paired variant calling

An in-house pipeline for high-confidence somatic variant calling was developed by combining three state-of-the-art tools. The pipeline was applied

to all tumor samples for which matched normal sequencing data was also available (27 tumor-normal samples from our cohort, and 59 samples from TCGA-LUAD). For a variant to be called as a "high-confidence" somatic variant, it had to be detected by at least two out of three somatic variant calling tools: VarScan2, Strelka2, and MuTect2. Such consensus-based methodologies improve the sensitivity and the specificity of variant calling compared to applying only one tool (Campbell et al., 2020). Details on the command line parameters for each tool are provided below.

VarScan2

1. Variant calling:

```
samtools mpileup -E -q 1 -f $REF_PATH \
-l $TARGETS_FILE $NORMAL_BAM $TUMOR_BAM |\
java -Xmx6g -jar $VARSCAN_PATH somatic \
-mpileup $TUMOR_ID --output-vcf
```

2. Variant processing:

```
java -Xmx6g -jar $VARSCAN_PATH \
processSomatic ${TUMOR_ID}.snp.vcf

java -Xmx6g -jar $VARSCAN_PATH \
processSomatic ${TUMOR_ID}.indel.vcf
```

3. Obtaining the input for `bam-readcount`:

3a. For SNVs:

```
grep -v '^#' $VCF | \
awk '{print $1"\t"$2"\t"$2}' >\
readcounts_input/${NAME}.bed
```

3b. For indels:

```
awk 'BEGIN {OFS="\t"} {if (!/^#/) { isDel=(length($4) >
length($5)) ? 1 : 0; print $1,($2+isDel-
20),($2+isDel+20); }}' $VCF >\
readcounts_input/${NAME}.bed
```

4. False positive filter:

```
java -Xmx6g -jar $VARSCAN_PATH fpfilter \
$VCF <(bam-readcount -q 20 -b 13 -w 1 -f ${REF_PATH} \
-l readcounts_input/${NAME}.bed $TUMOR_BAM |\
awk 'BEGIN {OFS="\t"} {for(i=2; i<=NF; i++){$i =
toupper($i)}; print $0}') \
--output-file fpfilter/${NAME}.pass \
--filtered-file fpfilter/${NAME}.fail \
--max-mmqs-diff 100 --min-strandedness 0
```

Strelka2

The workflow was configured as:

```
configureStrelkaSomaticWorkflow.py \
--normalBam=${NORMAL_BAM} \
--tumorBam=${TUMOR_BAM} \
--referenceFasta=${REF_PATH} \
--targeted \
--callRegions=${TARGETS_FILE} \
--runDir=${WORK_DIR}
```

And then Strelka2 was run as:

```
runWorkflow.py -m local -j 32
```

MuTect2

`MuTect2` was run with the following parameters: `-mbq 13 --disable-read-filter MateOnSameContigOrNoMappedMateFilter`.

Then, `FilterMutectCalls` was run with default parameters.

Variant processing and merging

For each output of each variant caller, variants were normalized and left-aligned using:

```
bcftools norm –f ${REF_PATH} –m –
```

Then, the variants from the three callers were merged, so that only those detected by at least two of the three callers were kept, using in-house R scripts.

Variant annotation and filtering

Variants were annotated using ANNOVAR with the following databases:

- ensGene (Ensembl gene annotation).
- 1000g2015aug_all (polymorphisms in the general population, 1000 Genomes Project).
- exac03 (polymorphisms in the general population, ExAc Project).
- avsnp150 (polymorphisms in the general population, dbSNP Project).
- gnomad30_genome (polymorphisms in the general population, gnomAD Project)

Then, variants meeting at least one of the following criteria were filtered out:

- ≥0.1% frequency in 1000 Genomes, ExAc, or gnomAD.
- Variant affects a "Low_complexity" or "Simple_repeat" region from RepeatMasker (data downloaded from UCSC Genome Browser).
- Indels of length greater than 1. We found that the majority of the detected indels longer than 1 nt were artifacts and could impair downstream analyses.

Unpaired variant calling

Unpaired variant calling was performed on our full collection of 37 cell lines and 70 primary tumors. Among the 70 primary tumors, 27 had matched normal samples and therefore they were also analyzed by the paired pipeline, which allowed us to compare the performance of both pipelines.

Variant calling was performed using `bcftools`. Then, extensive filtering was performed to remove low-quality variants:

- Variants with a variant calling quality score < 20 were filtered out.
- Variants in regions with fewer than 8 total reads, or supported by fewer than 5 mutant reads, were filtered out in tumor samples.
- Variants with a low allele frequency (<20%) were flagged but not filtered out.

The command was:

```
bcftools mpileup -f ${REF_PATH} -R ${TARGETS_FILE} \
--threads 32 -q 1 -Q 13 -a 'FORMAT/AD' \
-Ou ${SAMPLE}.bam |\
bcftools call -vmO u --threads 32 |\
bcftools filter -e \
"%QUAL<20 | ((FMT/AD[0:0]+FMT/AD[0:1])<8 &
FMT/AD[0:1]<5)" \
-s "LowQual" -m +x -O u |\
bcftools filter –e \
"FMT/AD[0:1]/(FMT/AD[0:0]+FMT/AD[0:1])<0.2"\
-s "LowFreq" -m + -O u |\
bcftools sort -O z > ${SAMPLE}.vcf.gz
```

For normal samples, the `((FMT/AD[0:0]+FMT/AD[0:1])<8 & FMT/AD[0:1]<5)` filter was removed. This allowed us to remove more germline variants from tumor samples (see below).

Variants were merged and left-aligned using:

```
bcftools merge -F x -m none |\
bcftools norm -f ${REF_PATH} -m - |\
bcftools norm -d none
```

Then, variants were annotated using ANNOVAR in the same way as for the paired analysis. Finally, extensive filters were applied to remove as many germline or false-positive variants. In addition to the filters used for the paired analysis, variants detected in at least one normal sample were filtered out.

3.2.8. Comparisons between variant files

To compare VCF files from different analyses or cohorts, first they were restricted to the unpadded sequencing targets using `tabix`. Then, pairs of VCF files were compared using `bcftools stats`. Variants reported in only one of the two files were extracted using `bcftools isec`.

3.2.9. Driver discovery analyses

Driver discovery was performed using two tools which, to our knowledge, are the state-of-the-art tools for driver discovery in non-coding regions in targeted sequencing data (Rheinbay et al., 2020). Their methods for identifying positive selection differ:

- **OncoDriveFML** searches for features that are enriched in variants predicted to have a high functional impact based on some external score (Mularoni et al., 2016).
- **OncoDriveCLUSTL** searches for clustered variants within features (Arnedo-Pac et al., 2019). Variant clusters are scored based on the number of variants that they contain and the shape of the smoothed distribution of variant density.

Further details on how both tools work were provided in ***Section 1.5.4***. Because the tools rely on fundamentally different criteria to identify positive selection, they were considered complementary to one another.

Driver discovery was performed separately on each type of genomic feature: CDSs, non-coding RNA exons, promoters, UTRs, and splice sites. This was necessary because the distribution of variants in each trinucleotide context and the functional impact scores were expected to differ between types of genomic features (Arnedo-Pac et al., 2019; Mularoni et al., 2016). Therefore, before driver discovery, our sequencing targets were split and reannotated as described below.

3.2.9.1. Reannotation of targets

For driver discovery analyses, our sequencing targets were split into the different types of features, accounting for overlaps between features of different types based on criteria from the PCAWG Consortium (Campbell et al., 2020). During this process, feature annotation was updated to GENCODE v29. Each feature type was defined as follows (***Figure 11***):

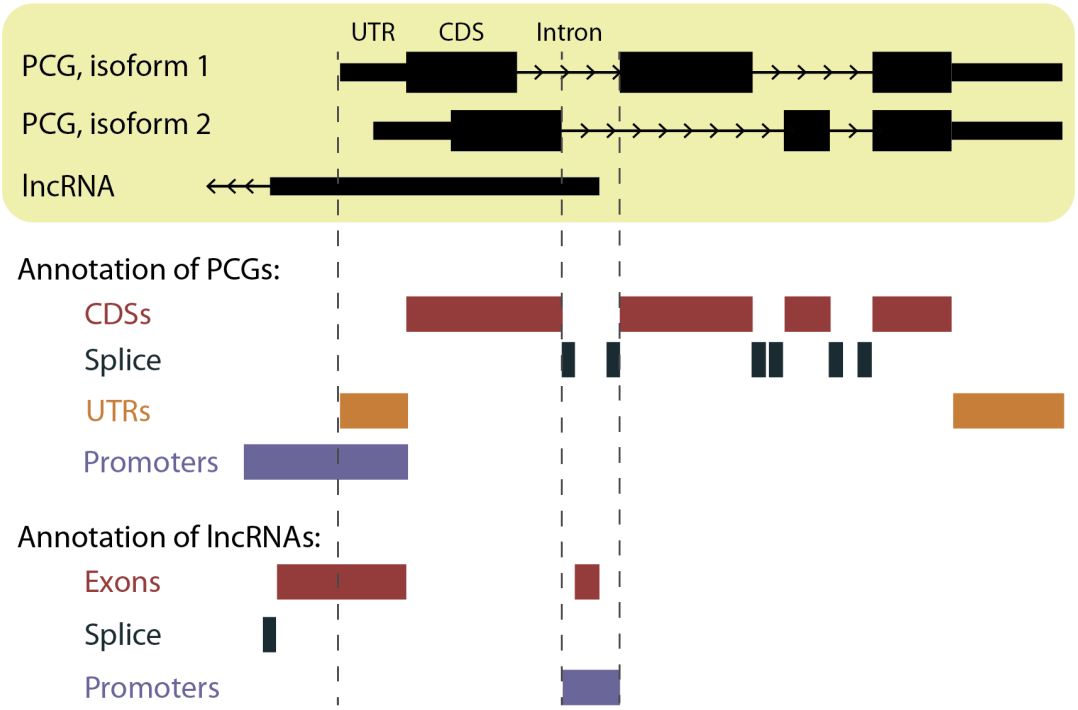


***Figure 11. Reannotation of sequencing targets**. See main text for a description of the criteria. PCG: protein coding gene; lncRNA: long non-coding RNA; UTR: untranslated region; CDS: coding sequence.*

- CDSs were always assigned the highest priority.
- For UTRs, the nucleotides that overlapped CDSs of other transcripts were removed.
- Splice regions were defined as the first 10 bp at both ends of each intron of GENCODE basic transcripts. Nucleotides that overlapped CDSs were removed.
- For lncRNA exons, nucleotides that overlapped CDSs, splice donor sites, or splice acceptor sites were removed. The following biotypes were considered as lncRNAs: "lincRNA", "antisense", "TEC", "processed_transcript", "non_coding", "sense_intronic", "3prime_overlapping_ncRNA", "bidirectional_promoter_lncRNA", "sense_overlapping", and "macro_lncRNA".
- For miRNA genes, miRBase 21 annotation was used. Nucleotides that overlapped CDSs, splice donors, or splice acceptors were removed.
- Promoters were defined as TSSs ± 200 bp, and then CDSs were subtracted.

Genes whose remaining features were (in total) shorter than 50 nt were filtered out. They originated either from spurious overlaps between our sequencing targets and genes that were not intended to be sequenced, or from non-coding features that mostly overlapped CDSs. Filtering these features out improved driver discovery because it reduced the number of hypotheses for multiple testing corrections.

3.2.9.2. OncoDriveFML

CADD scores were used to annotate the functional impact of variants. The following options were changed from the default configuration file template:

```
[genome]
build = 'hg38'

[score]
file = <path_to_cadd_file>
format = 'tabix'
chr_prefix = 'chr'
```

Then, OncoDriveFML was run as:

```
oncodrivefml \
-i <variants_file> \
-e <targets_file> \
-s targeted \
-o <out_dir> \
-c <config_file>
```

Other functional impact scores, including phyloP and RNAsnp, were also explored. However, few or no hits were found by these methods, even for CDSs.

3.2.9.3. OncoDriveCLUSTL

In a first attempt at running OncoDriveCLUSTL, we noticed that the tool may not handle multi-nucleotide variants (MNVs, two or more consecutive variants in the same sample) correctly because it seemed to consider them as independent (***Section 3.3.6***). Therefore, we modified our input to only keep the first nucleotide of MNVs, as if they were SNVs.

Optimal parameters for OncoDriveCLUSTL were found empirically based on the methodology described in the original publication (Arnedo-Pac et al., 2019). For each combination of parameters, OncoDriveCLUSTL was run in our paired dataset in CDSs. All possible combinations of the following parameter values were tested:

- `--smooth-window`: from 5 to 60, increments of 5.
- `--cluster-window`: from 5 to 60, increments of 5.
- `--simulation-window`: from 30 to 60, increments of 5.

Then, the Kolmogorov-Smirnov test was performed to evaluate the deviation of the non-significant p values ($p > 0.1$) from a uniform distribution. Next, the combinations of parameters that had the minimum Kolmogorov-Smirnov statistics (in absolute value) were selected, allowing for a 10% margin. Finally, out of these candidate combinations of parameters, the one that yielded the highest overlap of drivers with genes from the Cancer Gene Census was selected (Sondka et al., 2018).

OncoDriveCLUSTL was run with optimal parameters as follows:

```
oncodriveclustl \
-i <variants_file>
-r <targets_file> \
-o <out_dir> \
-sim region_restricted \
--genome hg38 \
--concatenate \
--smooth-window 45 \
--cluster-window 15 \
--simulation-window 40 \
--qqplot
```

By default, OncoDriveCLUSTL filters out clusters and features harboring fewer than 2 mutations in the cohort.

3.2.10. Curation of candidate drivers

3.2.10.1. Analyses performed on all feature types

The following features were studied in all candidate drivers regardless of their type of genomic feature:

- **BAM files were explored on Integrative Genomics Viewer (IGV)** to rule out mapping and sequencing artifacts and to visualize the distribution of variant alleles.
- **Genomic regions were explored in the UCSC Genome Browser** to study their genomic characteristics, the distribution of the variants, and the presence of nearby or overlapping genes.
- **Presence of the candidate in external collections of cancer-related PCGs or lncRNAs** was assessed from the Cancer Gene Census (CGC) and the Cancer LncRNA Census (CLC), respectively (Tate et al., 2019; Vancura et al., 2021).
- **Mappability** of the nucleotide positions affected by the variants was assessed by Umap (Karimzadeh et al., 2018).
- **Presence of recurrent somatic copy number alterations** across multiple cancer types was determined from a previous reanalysis of TCGA (Athie et al., 2020).
- **Regulatory elements** (promoters and enhancers) were retrieved from GeneHancer (Fishilevich et al., 2017).
- **Conservation** of the mutated nucleotide across 100 vertebrates was assessed using phyloP (Pollard et al., 2010). Scores from phyloP measure the rate of evolution of the nucleotide compared to a background model of neutral evolution. Higher positive values mean slower evolution (higher conservation).
- **Presence of the variant in a pan-cancer collection of non-coding somatic variants** from COSMIC (Tate et al., 2019).
- **Functional impact prediction by FATHMM-MKL**. FATHMM-MKL predicts the deleteriousness of non-coding variants by combining multiple features (Shihab et al., 2015). Scores range from 0 to 1. Higher scores mean higher deleteriousness. Variants were considered

deleterious if score > 0.7. The script to annotate variants using FATHMM-MKL was retrieved from https://github.com/HAShihab/fathmm-MKL and the database of pre-computed FATHMM-MKL scores was downloaded from http://fathmm.biocompute.org.uk/database/fathmm-MKL_Current_zerobased.tab.gz.

- **Functional impact prediction by CADD**. CADD also combines multiple features to predict the functional impact of variants (Kircher et al., 2014). Scores are expressed in Phred scale. Higher scores mean higher deleteriousness.

3.2.10.2. Analyses for specific feature types

lncRNA exons

- **Expression**. Expression data in normal tissues was retrieved from the Genome Tissue Expression (GTEx) Consortium (de Goede et al., 2021) and from normal samples in TCGA-LUAD. Expression data in tumor tissues was retrieved from TCGA-LUAD.
- **Survival**. Survival of TCGA-LUAD patients based on the expression levels of the candidate lncRNAs was also explored (***Section 3.2.13***).
- **Structural impact of variants**. The impact of SNVs on RNA structure was predicted using RNAsnp (Sabarinathan et al., 2013). Nucleotide sequences and annotations of the lncRNAs affected by the variants were retrieved from Ensembl v95. Then, the relative positions of the variants within the mature lncRNAs were inferred. Indels were not included in the analysis because they are not accepted as input by RNAsnp. Next, the `RNAsnp` command line tool was run in "mode 1".
- **Disruption of miRNA binding sites** was evaluated using miRcode (Jeggari et al., 2012) and miRDB (http://mirdb.org/mirdb/custom.html).

To obtain background distributions for RNAsnp, FATHMM-MKL, CADD, and phyloP scores in lncRNA exons, 10 000 SNVs were randomly generated within our target lncRNA exons. Both the positions and the nucleotide changes were randomly selected.

miRNAs

Due to their complexity, the methods for analyzing miRNA variants are detailed in a separate section (***Section 3.2.11***). In brief, miRNA variants were reannotated using an in-house pipeline to determine:

- Whether they affected miRNA seeds, mature miRNAs, pre-miRNAs, or pri-miRNAs.
- Whether they created or disrupted DROSHA processing motifs.

Splice regions

The effect of variants in splice regions on RNA splicing was evaluated in the TCGA-LUAD dataset using matched RNA-Seq data. The MAJIQ software was used as previously described (Andrades et al., 2022). MAJIQ estimates local splicing variations, i.e., it analyzes each splice junction separately and it does not report full transcripts. First, a database of known and novel local splicing variations was built using `majiq build` and GENCODE v29 annotation in GFF3 format, filtered so that it only contained our genes of interest:

```
majiq build -j 4 \
-c <config_file> \
-o <output_dir> \
--minreads 3 --minpos 2 --min-denovo 3 \
--irnbins 0.1 --min-experiments 1 \
<gencode_v29.gff3>
```

The configuration file specified the location of the RNA-Seq BAM files, the version of the human genome (hg38), and the strandedness of the RNA-Seq data (None).

Next, the degree of usage of each splice junction in the mutant samples was estimated by:

```
majiq psi <input_files> -o <output_dir> \
-n psi –minreads 3 –minpos 2 --min-experiments 1
```

The splicing alterations were visualized using `voila` and IGV. By design, `majiq` is unable to detect "partial intron retention" events, and therefore these events were only detected using IGV.

Promoters

- **Association of promoter variants and gene expression**. For hits in cell lines, expression data from the CCLE were analyzed. The expression data were restricted to 36 LUAD cell lines from our experiments (LC319 was missing from CCLE). For primary tumors, transcriptome profiling data from TCGA-LUAD were used, restricted to the 59 tumors included in our analysis. Association between promoter variants and changes in the expression of nearby genes was determined using Student's t-tests on the $\log_2$-transformed TPMs. For bidirectional promoters, the genes in both directions were studied.
- **Changes in transcription factor binding sites**. Sequence-based prediction of transcription factor binding sites was performed using PROMO (v3.0.2) (Messeguer et al., 2002).

UTRs

- **Structural impact**. The structural impact of variants in UTRs was assessed using RNAsnp in "mode 1". The input RNA sequences were retrieved using the genomic ranges defined in our reannotation of targets (***Section 3.2.9.1***). Ranges shorter than 200 nt were not analyzed. Indels were not analyzed because they are not accepted by RNAsnp.
- **Correlation with expression**. Changes in mRNA expression in *cis* were assessed in CCLE data. TCGA-LUAD data were not used because we found no UTR hits in the dataset.
- **miRNA binding**. The miRNAs that may bind to the wild type and mutant 3'-UTRs were predicted using miRcode (Jeggari et al., 2012) and miRDB (Chen and Wang, 2019). Input sequences for miRDB were obtained by querying the hg38 human genome using the ranges of the targets that we defined for driver discovery.

3.2.11. miRNA-centric reannotation of variants

Although ANNOVAR annotation was useful for initial data exploration, it was insufficient for an in-depth analysis of miRNA variants. Therefore, reannotation of miRNA variants was performed to include the following information:

- Whether each variant affected the pri-miRNA, pre-miRNA, mature miRNA, or seed.
- Additional information from external resources.
- Whether each variant disrupted or created DROSHA processing motifs.

3.2.11.1. Mapping variants to miRNA genes

All variants, regardless of how they were annotated by ANNOVAR, were intersected with a miRNA gene annotation derived from miRBase (version 21). The miRBase annotation file contained the genomic coordinates of 1881 miRNA genes and most of their mature miRNAs, although for 949 genes the 5p or the 3p mature miRNAs were missing due to lack of experimental support. MicroRNA gene sequences were padded by 200 nt upstream and downstream, and positions that overlapped CDSs from GENCODE were removed. Seed sequences were defined as nucleotides 2-7 of the mature miRNAs. Variants outside mature miRNAs but within miRNA genes were annotated as "pre-/pri-miRNAs", and those flanking miRNA genes were annotated as "pri-miRNA / intergenic". These ambiguities arose because miRNA gene annotation is not consistent: annotated miRNA genes contain, at minimum, the stem-loop sequence from the pre-miRNA, which is often (but not always) flanked by a variable number of nucleotides from the pri-miRNA. Therefore, most annotated miRNA genes do not represent the full primary transcripts (pri-miRNAs), which may or may not span more nucleotides outside of the annotated genes.

3.2.11.2. Annotation of miRNA variants using external resources

The variants that affected miRNA gene regions were further annotated using the same external resources as for lncRNAs (***Section 3.2.10.2***), except for RNAsnp, which is not recommended for sequences shorter than 200 nt (Sabarinathan et al., 2013).

3.2.11.3. Mapping miRNA variants to DROSHA motifs

We developed a novel pipeline to identify miRNA variants that disrupt or create DROSHA processing motifs (basal UG, mGHG, apical UGUG, and downstream CNNC). DROSHA motifs can be predicted based on "positional" or "structural" criteria, and both criteria can be informative (***Section 1.7***). Our pipeline, which implemented both criteria using in-house R scripts, was based on the following steps:

(i) Retrieve the sequences of the transcripts originating from miRNA genes, padded by 30 nt at 5' and 3'.
(ii) Define the positions of the 5p and 3p mature miRNAs within the sequences.
(iii) Predict the secondary structures of the sequences.
(iv) If the predicted structures are stem-loops, predict the basal and apical junctions of the pri-miRNA stems.
(v) Use the information from (ii) to predict the locations where motifs may be present based on "positional" criteria.
(vi) Use the information from (ii)-(iv) to predict the locations where motifs may be present based on "structural" criteria.
(vii) Intersect the positions of the variants with the positions defined in (v) and (vi), and use the positions and the sequences of the variants, as well as the regions from (v) and (vi), to determine whether each variant creates or disrupts a motif.

Below we detail each step.

Retrieval of pri-miRNA sequences

Usually, annotated miRNA genes do not contain the full pri-miRNA and, therefore, they may not contain the full stem and important motifs such as the downstream CNNC. Therefore, the coordinates of miRNA genes were padded by 30 nt at both ends and the sequences of these regions were retrieved using the `BSgenome.Hsapiens.UCSC.hg38` R package. The padding length of 30 nt was selected after extensive testing and it is supported by previous reports (Roden et al., 2017).

Definition of the positions of 5p and 3p mature miRNAs

In miRBase 21, only 932/1881 miRNAs had fully annotated 5p and 3p mature miRNAs. To increase the number of fully annotated miRNAs, miRBase annotation was complemented with that from MirGeneDB v2.0 (Fromm et al., 2020). Only canonical miRNAs from MirGeneDB were used. This increased the number of fully annotated miRNA genes to 1012.

Although previous reports have predicted the position of "missing" mature miRNAs based on the positions of the "non-missing" ones by assuming that DROSHA produces a 2-nt staggered cut (Kim et al., 2021; Urbanek-Trzeciak et al., 2020), we deemed these predictions to be too error-prone based on our observations. In particular, we observed many examples of pri-miRNAs whose distance between the 5p and 3p DROSHA cleavage sites was different from 2 nt. Moreover, our 1012 fully annotated pri-miRNAs already contained 275/295 (93%) of the human high-confidence pri-miRNAs defined by miRBase, and only 20/869 (2%) of the incomplete pri-miRNAs were high-confidence. Therefore, we preferred to work with the 1012 miRNAs that had a complete annotation rather than to add error-prone annotations for the remaining, mostly low-confidence miRNAs.

Secondary structure predictions

Secondary structures of the padded miRNA gene sequences were predicted using `RNAfold` with the `--no-LP` option (Lorenz et al., 2011). Output ".ct" files were used for downstream analyses.

Prediction of pri-miRNA stems

The method of prediction of pri-miRNA stems was largely based on previous reports (Roden et al., 2017), with modifications that improved performance (***Figure 12***):

1. **Filter out "non-hairpin-like" structures**. If <50% of the nucleotides from the 5p and 3p mature miRNAs are paired to each other, discard the structure.
2. **Roughly define the 5p and 3p arms**. An approximate position of the apical loop and the 5p and 3p arms was necessary at this point to define which direction was "basal" and which direction was "apical". As a rough approximation, the apical loop was defined as the stretch of unpaired nucleotides most equidistant to the two DICER cleavage sites. The 5p arm was defined as all nucleotides upstream of the apical loop, and the 3p arm was defined as all nucleotides downstream of the apical loop.
3. **Define the apical junction**.
   3.1. For each arm of the hairpin, if the DICER cleavage site is unpaired, find the first paired nucleotide towards the basal direction. Otherwise, keep the position of the DICER cleavage site.
   3.2. From the two positions defined in the previous step, keep the most apical one.
   3.3. Starting at the position defined in the previous step, search for the first mismatch affecting ≥2 nt (counting both arms) in the apical direction. Define the apical end of the stem as the last paired position before this mismatch.
4. **Define the basal junction**. From the apical junction, start moving towards the basal direction. Stop at the basal end, defined by either of these conditions:
   4.1. A stretch of ≥12 unpaired nucleotides (counting both arms) if the stem length up to this point is <30 nucleotides.
   4.2. A stretch of ≥6 unpaired nucleotides (counting both arms) or an unstable lower stem if the stem length up to this point is ≥30 nucleotides. Unstable lower stems were defined as: the next two nucleotides in either arm are unpaired, and no more than 2 of the next

6 nucleotides in the affected arm are paired. Stable lower stems are critical for DROSHA processing (Kim et al., 2021).

4.3. The beginning of a multi-branched junction: a point in which one arm pairs to itself.

4.4. Otherwise, the last paired nucleotide towards the basal end of the sequence.

The length of the stem was defined as the minimum distance between the apical and the basal ends of the stem.

5. **Final filtering**. Filter out stems that do not contain the DROSHA cleavage sites within them.

Stems could be successfully predicted for 842/1012 pri-miRNAs.

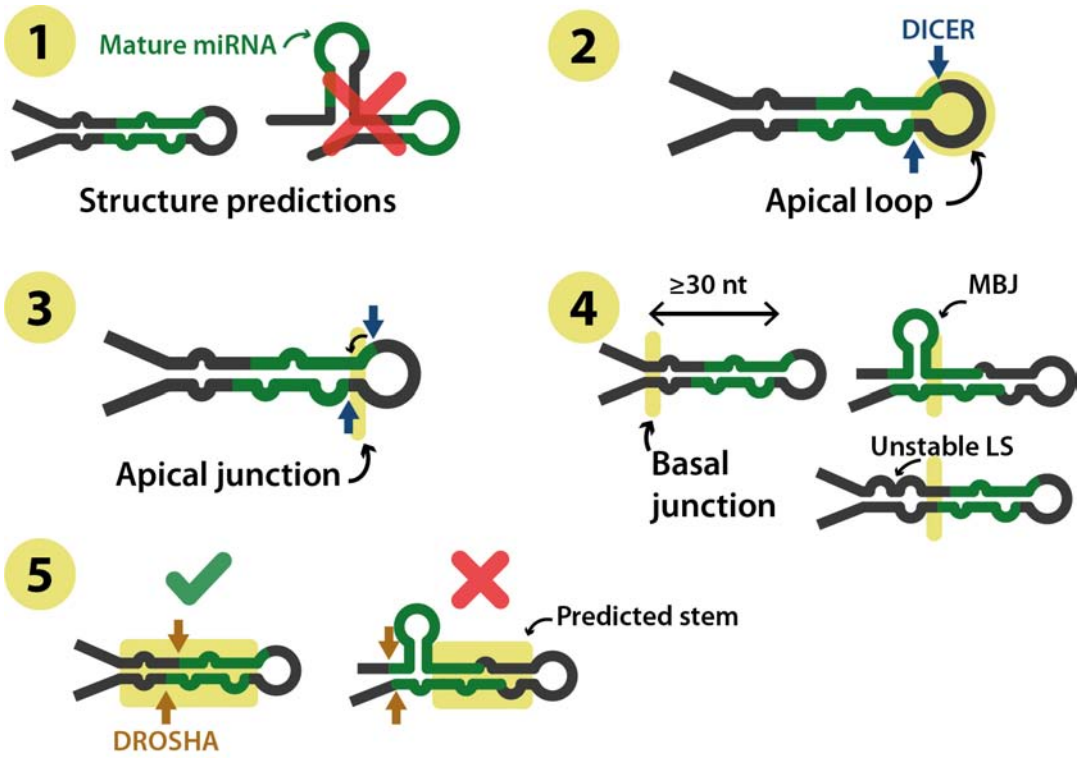


Figure 12. Workflow for predicting pri-miRNA stems. *The workflow consists of five main steps (see main text for details): 1) Removal of non-stem-loop structures. 2) Location of the apical loop. 3) Location of the apical junction. 4) Location of the basal junction. 5) Removal of stems that do not contain the DROSHA cleavage sites. MBJ: multi-branched junction; LS: lower stem.*

Prediction of motif positions

The positions where DROSHA motifs may be were retrieved using the criteria from ***Table 8***. The criteria were mostly based on previous reports (***Table 2***), with slight modifications for the structural criteria, which have been less researched to date and which we thoroughly assessed and optimized (***Section 3.3.7.2***). The sequences at the defined positions were also retrieved.

Table 8. Key DROSHA recognition features in pri-miRNAs as defined by positional and by structural criteria according to our method.

Feature	Positional definition	Structural definition
Basal UG	-14 nt from 5p DCS.	UG at basal junction.
mGHG	-7...-5 nt from 5p DCS, and opposite nt in 3p arm.	Same as positional.
Apical UGUG	UGU/GUG at +21...+25 nt from 5p DCS.	UGU/GUG at -2...+1 from apical junction.
Downstream CNNC	+16...+18 nt from 3p DCS.	+5...+9 from 3p basal junction.

*mGHG: "mismatched GHG". DCS: DROSHA cleavage site. Distances are expressed in nucleotides (nt) and they express the range where the first nucleotide of the motif may start. See also **Figure 7** and **Table 2**.*

Mapping of variants to motifs

We intersected the positions of the variants with the positions where DROSHA processing motifs could be. Then, for basal UG, apical UGUG, and downstream CNNC motifs, we determined if the motif was present in the wild type pri-miRNA. If it was present, we assessed if the variant disrupted the motif. If it was not present, we predicted if the variant created a motif.

For mGHG motifs, we retrieved the normalized mGHG scores for each possible pair of trinucleotides from the Supplementary Table S1 of (Kwon et al., 2019). We assigned scores as follows:

1. For each pri-miRNA whose hairpin prediction was successful, retrieve the nucleotides at positions -7...-5 from the 5p DROSHA cleavage site.
2. If none of the selected nucleotides are paired to the 3p arm, stop.
3. Otherwise, retrieve the three nucleotides of the 3p arm that should be "opposite" of the selected nucleotides (regardless of whether they are paired or not).
4. Stop if the selected nucleotides are not facing each other. This can occur if:
    a. The nth nucleotide in the 5p arm is paired to a nucleotide other than the nth nucleotide in the 3p arm, or vice versa.
    b. There is a bulge in one of the arms.
5. Otherwise, at this stage we had two trinucleotides facing each other in opposite arms, from which the mGHG score was retrieved.

In this way, we assigned mGHG scores to wild type and mGHG-mutant pri-miRNAs, predicting the impact of the variants on DROSHA processing efficiency. For analyses that required a binary classification of the mGHG motif (present vs. absent), we considered a mGHG motif to be present if score ≥38 as previously defined (Kim et al., 2021).

3.2.12. Prediction of miRNA targets

Two online resources were used to predict the targets of wild type and mutant miRNAs based on their sequence:

- **TargetScan**: release 5.2 is the latest one that allows for custom input sequences (https://www.targetscan.org/vert_50/seedmatch.html). It only uses the seed sequence for its predictions.
- **miRDB**: it uses machine learning to predict miRNA targets based on the sequence of the full mature miRNA (Chen and Wang, 2019). To select targets, we used the recommended threshold of score ≥80.

The sequence-based predictions were complemented by the following analyses, which could only be performed for wild type miRNAs:

- **TarBase** (version 8.0): database of experimentally validated miRNA targets.

- **Correlation analyses**. Correlation between miRNA and target mRNA or protein expression was assessed in LUAD samples from two datasets: TCGA-LUAD (mRNA only) and Gillette et al (mRNA and protein). Correlation was quantified by Kendall's tau (τ). Statistical and biological significance were defined as FDR < 0.05 and τ < -0.2.

Because gene symbols change over time, Ensembl IDs were used to compare sets of genes. Conversion between gene symbols and Ensembl IDs was performed using the `biomaRt` R package. For each source of gene symbols, the closest archived version of Ensembl was searched for. Because TargetScan v5.2 uses an annotation that is no longer archived by Ensembl, ~5% of its predicted targets were lost during the conversion.

3.2.13. Survival analyses

Survival analyses based on gene expression were performed using the R packages `survival` and `survminer`. For Kaplan-Meier curves, patients were stratified in "high" and "low" expression subgroups using the median expression as a cutoff, and logrank p values were calculated. For univariate and multivariate Cox analyses, the exact gene expression values were used. Covariates for multivariate analyses were selected by backward elimination.

3.3. Results

3.3.1. Power analysis

To detect novel non-coding mutations in LUAD, we performed targeted sequencing of genomic DNA from an in-house collection of 70 LUAD primary tumors, 27 matched normal samples, and 37 LUAD cell lines. The sequencing was targeted towards all human miRNA genes from miRBase (release 21) and exons of 908 lncRNA genes and 1307 PCGs. First, we did a *post hoc* analysis to determine our statistical power for detecting drivers at different frequencies in our cohort of 70 LUAD primary tumors (***Figure 13***). We analyzed each type of genomic feature separately because they had different sequence properties and mutational patterns.

The highest statistical power was associated with splice regions and miRNAs (***Figure 13***). Even in our relatively small cohort, we predicted 80% power to detect driver mutations at frequencies ≥7% above background for splice regions in lncRNAs and ≥9% for miRNAs and for splice regions in PCGs. Due to their short lengths and high conservation, it is unlikely for splice regions and miRNAs to be highly mutated by random chance, and therefore an excess of mutations over the BMR can be detected more easily. The next highest power was associated with promoters of both PCGs and lncRNAs, for which we predicted 80% power to detect driver mutations at frequencies ≥12% above background. This high power was also in part because promoters were defined as ~400 nt regions, which were shorter than most CDSs and lncRNA exons. Next, for lncRNA exons we predicted 80% power to detect driver mutations at ≥14% frequency above background; and for UTRs and CDSs, at ≥15% above background. The higher predicted power to detect driver mutations in lncRNA exons seemed to be mostly due to the lower number of features, which led to a less restrictive multiple testing correction (***Table 7***). In summary, our power analysis suggested a relatively low power to detect an excess of mutations over the background in most features of interest.

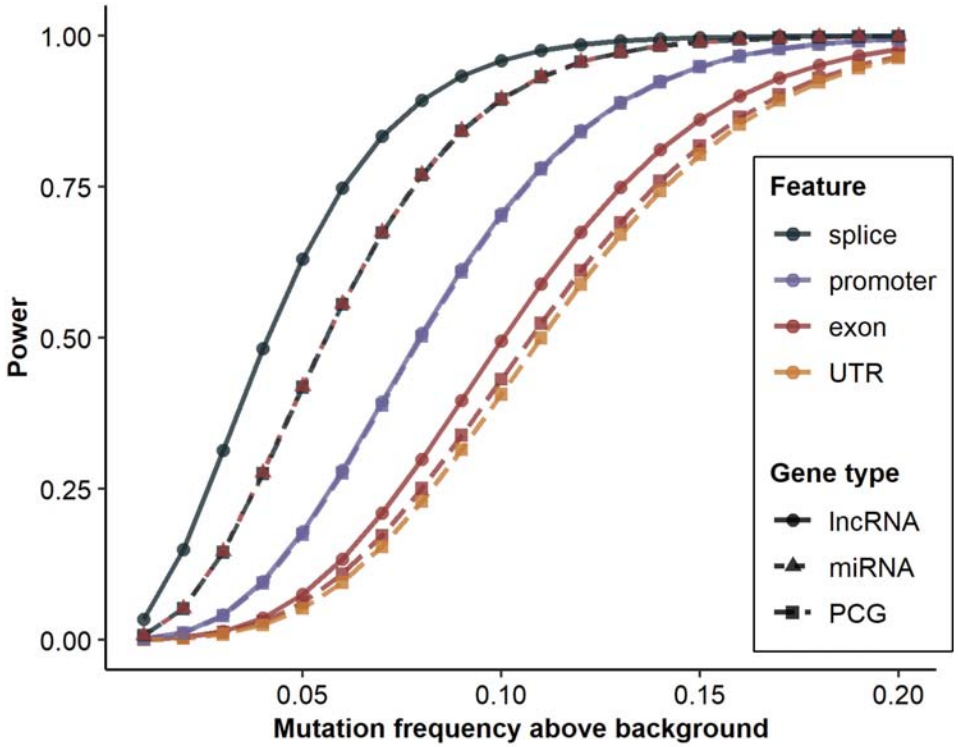


***Figure 13. Power analysis in our cohort of 70 LUAD primary tumors**. For each type of feature, we estimated the probability ("Power") of detecting mutations at different frequencies above the background. For protein-coding genes (PCGs), "exon" refers exclusively to coding sequences, excluding untranslated regions (UTRs). lncRNA: long non-coding RNA; miRNA: microRNA.*

3.3.2. Quality control of DNA-seq data

3.3.2.1. Raw FASTQ files

Quality of all raw FASTQ files was considered acceptable as assessed using FASTQC. Base detection quality was consistently high (>30 in Phred scale) along all nucleotide positions of the reads (***Figure 14A***) and across all physical positions in the sequencing cells (data not shown). In addition, average base detection quality was high across all reads (***Figure 14B***).

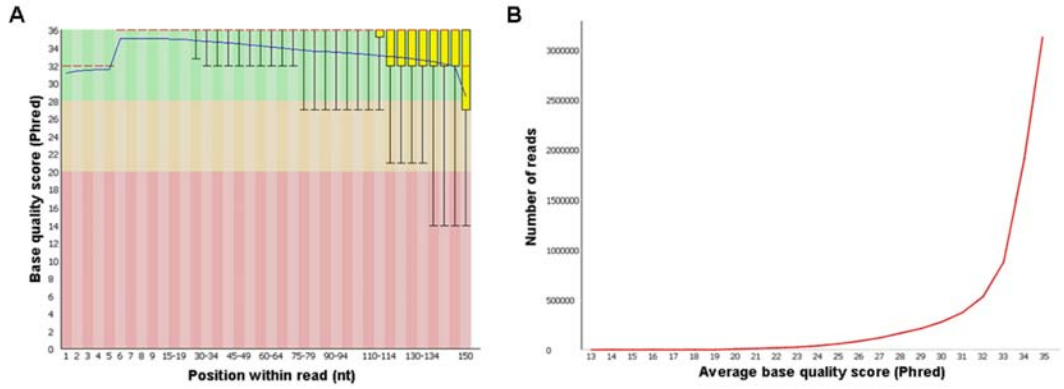


Figure 14. Quality metrics of a sequencing file from a representative tumor sample. ***A***. *Distribution of base quality scores at different positions along the reads.* ***B***. *Distribution of mean base quality scores of all reads. For a given p value, the Phred score is calculated as Phred = -10·log*$_{10}$*p.*

3.3.2.2. Alignment BAM files

The median number of reads per sample was ~13.0 million (IQR = 11.8-14.4 million) (***Figure 15A***). Across all samples, a median of >98% of the nucleotides targeted by our gene capture design were covered by at least one read. In addition, >92% of the target nucleotides were covered by at least 10 reads, >78% by at least 30 reads, and >63% by at least 50 reads (***Figure 15B-D***).

3.3.2.3. TCGA WGS data

The median number of reads per sample in our target regions was ~4.0 million (IQR = 3.4-4.8 million). The median percentage of target bases covered at ≥10X depth was 98.0% (IQR = 95.3%-99.1%); at ≥30X depth, 68.8% (IQR = 47.0%-82.1%); at ≥50X depth, 13.7% (IQR = 2.9%-40.1%).

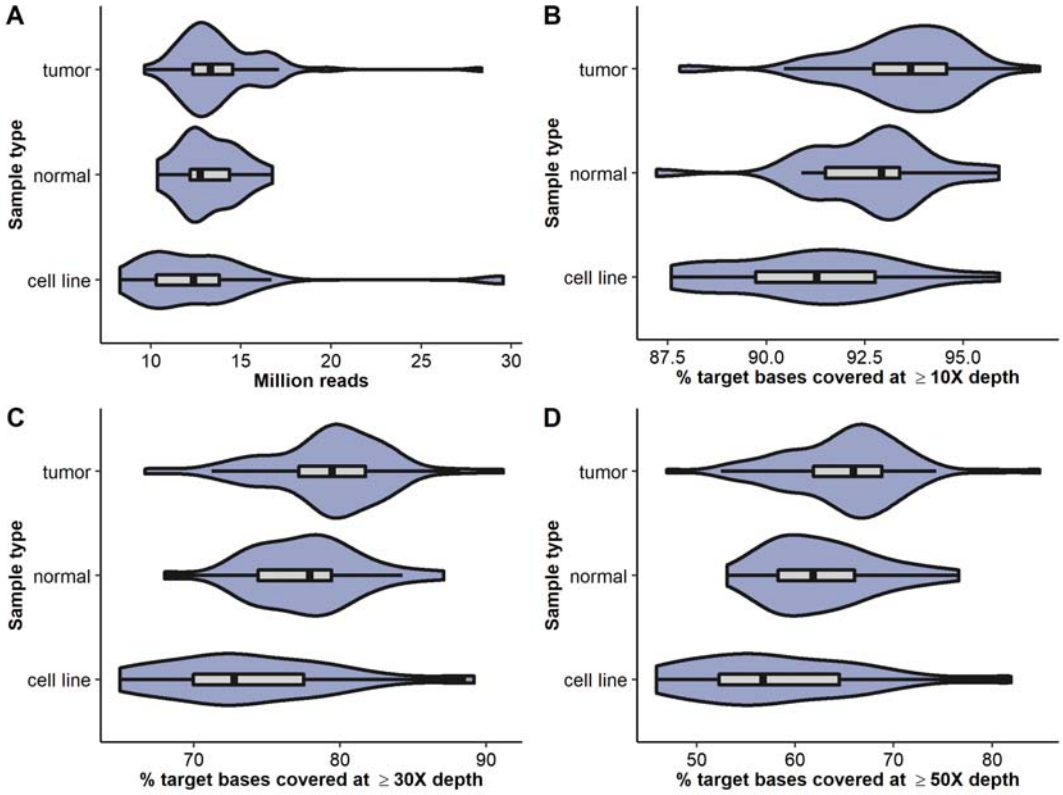


Figure 15. Quality metrics on the alignment BAM files. A. *Distribution of number of reads per sample.* ***B-D****. Percentage of target bases that were covered by at least 10, 30, or 50 reads, respectively.*

3.3.3. Evaluation of the variant calling pipelines

Objectively evaluating the performance of our variant calling pipelines in real data was challenging because there was no set of "ground truth" somatic variants. As an alternative, we compared the sets of variants detected by orthogonal methods in the same datasets, and then we thoroughly assessed the discrepancies. We restricted the analyses to the unpadded sequencing targets.

To evaluate our paired variant calling pipeline, we compared the variants detected by our pipeline with those detected by PCAWG in WGS data from TCGA-LUAD (N = 38). PCAWG used a "wisdom of the crowd" approach similar to our rationale, but their variant calling tools were different from ours, except for MuTect2 (Campbell et al., 2020; Rheinbay et al., 2020). The agreement between our pipeline and PCAWG's was high (84%) (***Figure 16A***).

We inspected a random selection of the discrepancies on the IGV software. Among the variants detected by our paired pipeline but not by PCAWG, ~33% seemed to be true positives at high depth (≥20% variant allele frequency, VAF), ~47% were subclonal (VAF < 20%), and ~20% were unclear or possibly germline (***Figure 16B***). On the other hand, among the variants detected by PCAWG but not by us, ~10% seemed to be true positives at ≥20% VAF and ~90% were subclonal. Overall, none of the two pipelines could be considered as "ground truth", but we deemed our pipeline to be acceptable.

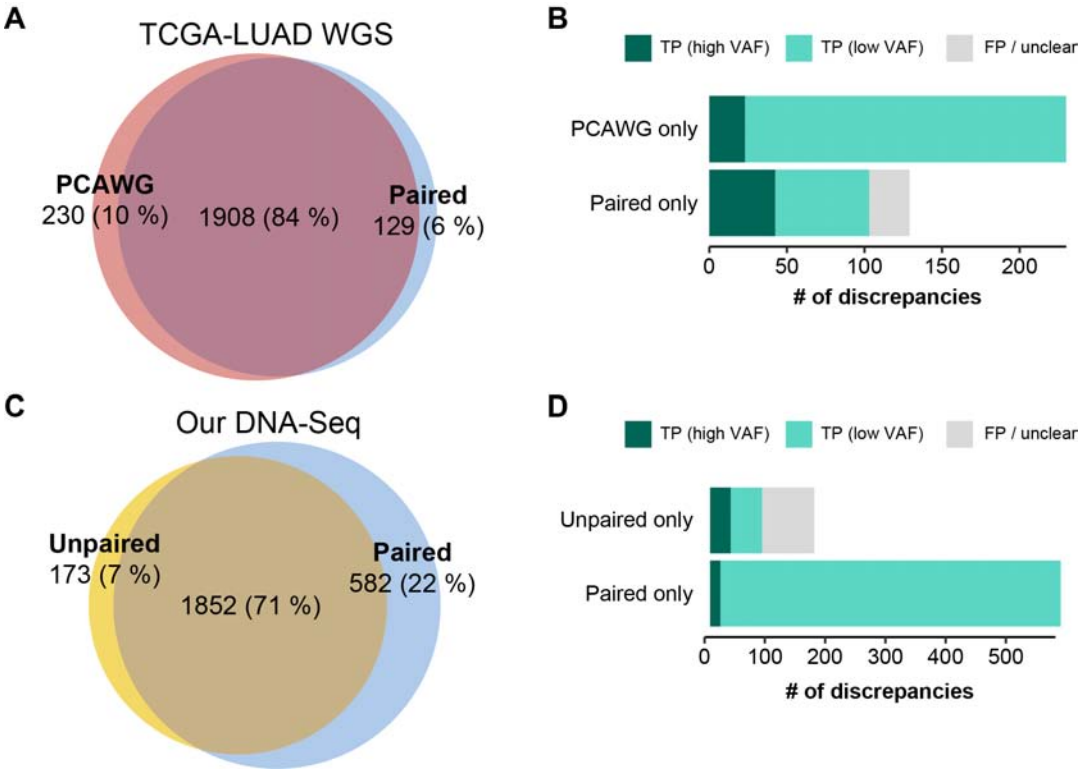


Figure 16. Number of variants detected by different somatic variant calling methods in the same datasets. A. *Comparison of our paired variant calling pipeline ("Paired") and the paired variant calling pipeline of the PCAWG Consortium in whole-genome sequencing (WGS) data from the TCGA-LUAD dataset (N = 38).* ***B.*** *Analysis of the discrepancies between the "Paired" and "PCAWG" pipelines in TCGA-LUAD data. TP: true positive; FP: false positive (germline variant or artifact); VAF: variant allele frequency. "High VAF" means VAF ≥ 20%; "Low VAF" means VAF < 20%.* ***C.*** *Comparison of our "Paired" pipeline and our "Unpaired" variant calling pipeline in our own paired DNA sequencing data (N = 27).* ***D.*** *Analysis of the discrepancies between the "Paired" and "Unpaired" pipelines in our paired DNA sequencing data.*

To evaluate our unpaired variant calling pipeline, we compared the variants detected by our unpaired and paired pipelines in our own paired DNA-Seq data (N = 27). The agreement was high (>70%) (***Figure 16C***). We used IGV to inspect a random selection of discrepancies. Among the variants detected by the unpaired pipeline but not by the paired pipeline, ~20% seemed to be true somatic variants at high depth, ~30% were subclonal, and ~50% seemed to be germline variants (***Figure 16D***). Therefore, the rate of germline variants in the unpaired analysis was ~4% ($0.5 \cdot 173/2025$), which we considered acceptable. On the other hand, among the variants detected by the paired pipeline but not by the unpaired pipeline, ~3% looked like true somatic variants at high depth and ~97% were subclonal.

To compare the performance of our paired and unpaired pipelines in additional external cohorts, we applied both of them to TCGA-LUAD WGS data. In a first approach, we only used 23 of the 59 normal samples to construct our germline variant resource for the unpaired pipeline to maintain a similar proportion to that of our cohort (27/70 normal samples). Here, the agreement between the paired and unpaired pipelines was only ~43%, mostly because the unpaired pipeline detected a large number of germline variants. In a second approach, we used all 59 normal samples to construct the germline resource for the unpaired analysis. As a result, the agreement rose to ~69% and most germline variants were successfully removed in the unpaired analysis. We concluded that using such a low number of normal samples to construct a germline variant resource for the unpaired pipeline was only successful in our cohort, which was relatively homogeneous (all patients were from the Basque Country region in Spain), and therefore a small proportion of normal samples could capture a high proportion of the germline variability of the population. However, the TCGA-LUAD cohort was more heterogeneous, as samples were acquired across many countries, and the subset of 23 normal samples failed to capture the germline variability of the cohort.

In conclusion, both our paired and unpaired variant calling pipelines were acceptable for the purposes for which they were used. Discrepancies between pipelines mostly involved subclonal variants, which were likely to be less relevant for driving cancer than high-VAF variants.

3.3.4. General variant statistics

The burden of variants per sample was mostly within the range of ~5-10 variants/Mb (***Figure 17*** and ***Table 9***). In primary tumors from our cohort, the number of variants per sample was higher in the unpaired analysis than in the paired analysis, possibly because the unpaired pipeline detected a small proportion of germline variants (***Section 3.3.3***). TCGA-LUAD WGS data had the lowest number of variants, which can be explained by the lower depth of the dataset compared to ours (***Section 3.3.2.3***). Finally, cell lines had the highest burden of variants, most of them surpassing 10 variants/Mb, which is a usual threshold to consider a tumor as hypermutated (Campbell et al., 2017).

Table 9. Burden of variants in our analyzed datasets.

Dataset	N	Median number of variants per Mb per sample (IQR)
Primary tumors, paired	27	5.38 (2.02-11.1)
Primary tumors, unpaired	70	8.67 (6.07-11.5)
Cell lines	37	16.8 (11.0-22.6)
TCGA-LUAD WGS	59	2.76 (1.35-6.05)

IQR: interquartile range. WGS: whole genome sequencing.

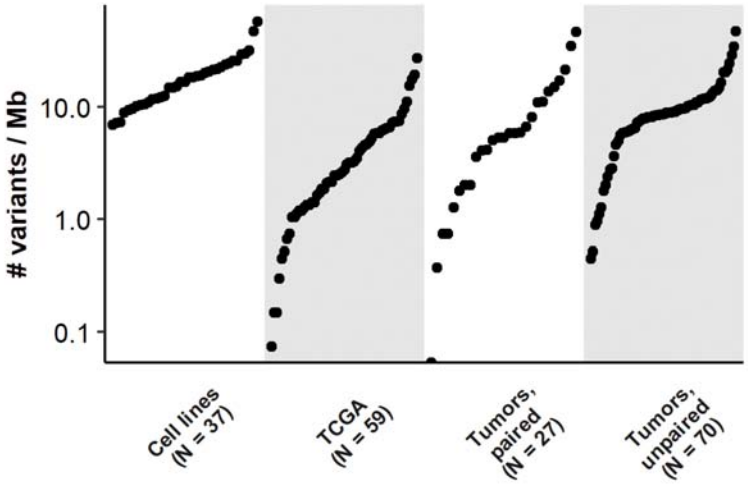


Figure 17. Variants per megabase (Mb) across the analyzed cohorts*. Each point represents one sample.*

The higher variant burden of cell lines compared to primary tumors was expected for various reasons:

- Cell lines are more genetically homogeneous than primary tumors, which facilitates variant calling.
- Removal of germline variants was less effective in cell lines than in the other datasets due to the lack of matched normal samples.
- Cell lines accumulate variants during their immortalization and cell culture (Domcke et al., 2013; Kim et al., 2017).
- Approximately half of our cell lines were metastatic. Metastatic tumors usually have higher burden of certain types of variants, such as MNVs and indels, when compared to primary tumors (Priestley et al., 2019). However, the frequency of SNVs in metastatic tumors is generally not significantly different from that of primary tumors, and therefore this may not have contributed to the higher variant burden of cell lines in a major way (Priestley et al., 2019).

Next, we classified the variants based on their biological effects (***Figure 18***). For this initial data exploration, we used the annotation reported by ANNOVAR. The largest number of variants was detected in PCGs, followed by ncRNAs and by intergenic regions. In PCGs, variants most frequently affected introns, followed by CDSs and by UTRs. Although we had only sequenced the first and last ~200 bp of each intron, introns had the highest mutation rate because their sequences are mostly under low constraint (Ulitsky, 2016). On the other hand, variants at splice sites of PCGs were rare because splice sites are short and highly conserved (Sibley et al., 2016). In contrast to PCGs, lncRNAs had more exonic than intronic variants, possibly because exonic nucleotides in lncRNAs are less constrained than those in CDSs and because some non-coding RNAs lack introns (Ulitsky, 2016). Finally, the reported intergenic variants were within 200 bp from the 5' and 3' ends of the targeted genes, as expected from our targeted sequencing design.

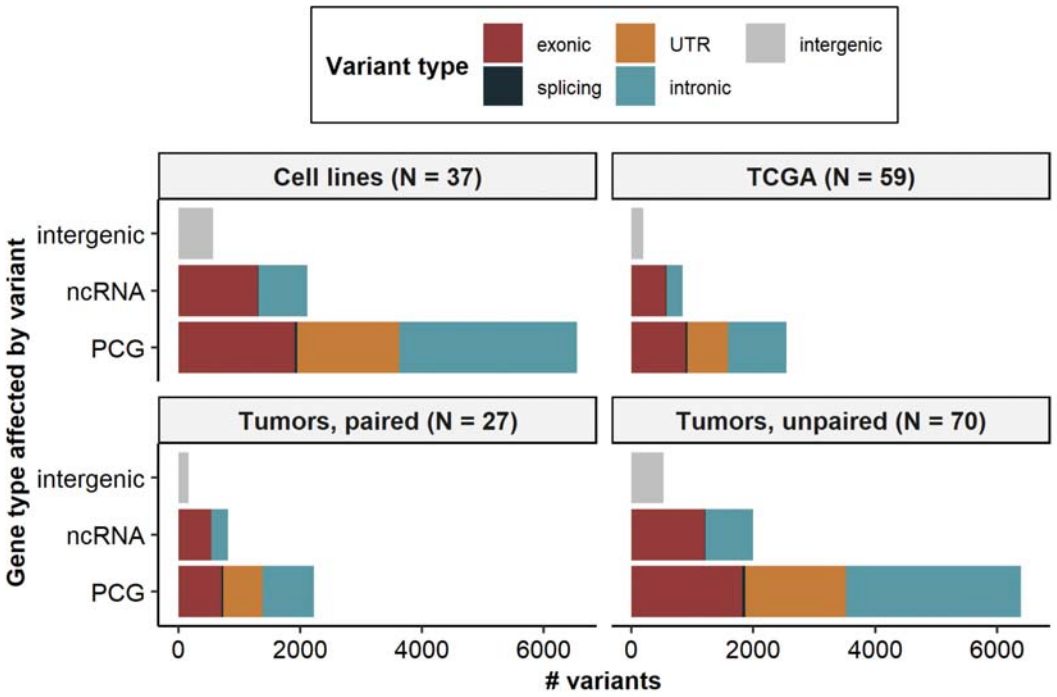


Figure 18. Distribution of variants in the analyzed cohorts according to the affected type of gene and gene region. *Variants were annotated using ANNOVAR. All types of non-coding RNAs (ncRNAs: miRNAs, lncRNAs, and others) are grouped in a single category. For protein-coding genes (PCGs), "exonic" refers to coding sequences, excluding untranslated regions (UTRs).*

3.3.5. Most lncRNA variants were passengers

Before performing driver discovery analyses, we wondered if variants in lncRNA exons, as a set, had high predicted impact. To this end, we compared the distribution of various structural and functional impact metrics in the observed variants across all datasets with that of randomly drawn variants from the target lncRNA exons. The observed variants did not have higher structural impact (RNAsnp), functional impact (FATHMM-MKL or CADD), or conservation (phyloP) scores than random variants (***Figure 19***). In fact, although differences between distributions were statistically significant for all scores but RNAsnp p values due to the large number of data points (Anderson-Darling tests, $p < 0.05$), average functional impact and conservation scores were actually lower for observed variants than for random variants, but the magnitude of the differences was small. No differences between datasets were observed by analyzing each dataset separately (data not shown).

Driver variants are sometimes recurrent across cohorts and cancer types. Therefore, we wondered if our variants in lncRNA exons had been reported in external pan-cancer WGS datasets from COSMIC (Tate et al., 2019). However, only 4/1224 (0.3%) variants in cell lines, 3/568 (0.5%) in paired primary tumors, 8/991 (0.8%) in unpaired primary tumors, and 10/548 (1.8%) in TCGA-LUAD were present in COSMIC.

In summary, the vast majority of lncRNA exon variants were likely passengers, in agreement with the notion that <1% of the variants in a tumor are drivers (Campbell et al., 2020). Therefore, we needed more sophisticated methods to find drivers, if present, among the overwhelming majority of passengers.

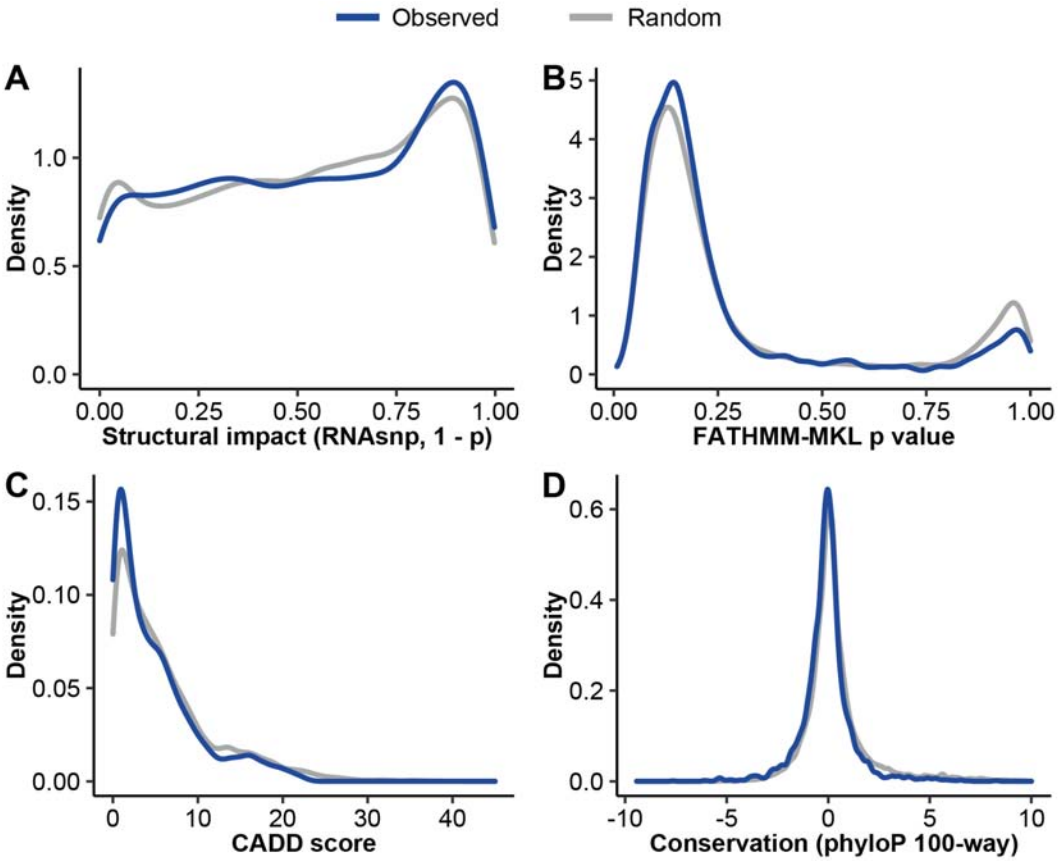


Figure 19. General features of lncRNA variants. *Variants in lncRNA exons were annotated by:* ***A****. Predicted structural impact (RNAsnp; p values were transformed as 1 - p);* ***B****. FATHMM-MKL;* ***C****. CADD; and* ***D****. Conservation (phyloP 100-way). Higher values mean higher impact or conservation. In blue, values for all variants across all analyzed cohorts (primary tumors, patients, and TCGA-LUAD WGS); in grey, values for 10 000 randomly generated variants within our target lncRNA exons.*

3.3.6. Driver discovery

To identify putative LUAD drivers in our datasets (cell lines, unpaired and paired primary tumors, and TCGA-LUAD WGS), we applied OncoDriveFML and OncoDriveCLUSTL to each dataset and to each type of genomic feature covered by our targeted sequencing: CDSs, lncRNA exons, miRNAs, promoters, splice regions, and UTRs. Whereas OncoDriveFML searches for enrichment in high functional impact scores (in our case, CADD scores), OncoDriveCLUSTL searches for clusters of variants (Arnedo-Pac et al., 2019; Mularoni et al., 2016). Here, we first provide a general overview of our results across all feature types. Then, we explore each result in each feature type in greater detail.

3.3.6.1. General performance of driver discovery tools

OncoDriveFML and OncoDriveCLUSTL performed quite differently in our data. At $q < 0.25$, OncoDriveFML detected a 54 drivers across all feature types and all cohorts (***Table 10***). In addition, the quantile-quantile plots of p values were acceptable for CDSs, with only slight inflation in some cohorts (***Figure 21***), but p values were deflated for lncRNA exons (***Figure 23***). On the other hand, in a first approach, OncoDriveCLUSTL predicted 353 drivers across all feature types and all cohorts and p values were heavily inflated (data not shown). About half of these predicted drivers ("hits") contained multi-nucleotide variants (MNVs), and we hypothesized that MNVs were being considered as independent events, biasing cluster detection. Indeed, when we modified our input to only keep the first mutated position of MNVs, the number of hits decreased to 185 (48% reduction) (***Table 11***). However, quantile-quantile plots of p values were still mostly inflated, suggesting that other factors were still causing a high false positive rate (***Figure 22*** and ***Figure 24***). Nevertheless, the hits from both OncoDriveFML and OncoDriveCLUSTL were significantly enriched in known cancer-related PCGs from the Cancer Gene Census (CGC) and in known cancer-related lncRNAs from the Cancer LncRNA Census (CLC) in, at least, some analyses, confirming that both methods were detecting genuine cancer genes (***Figure 20***). In the sections below, we detail our analyses of hits from different feature types.

Table 10. Number of hits in the OncoDriveFML analysis.

Feature	Cell lines (N = 37)	TCGA, WGS (N = 59)	Tumors, paired (N = 27)	Tumors, unpaired (N = 70)
PCG, CDS	5	11	6	4
PCG, splice	2	6	2	4
PCG, UTR	4	0	0	0
PCG, prom.	0	0	1	2
lncRNA, exon	0	0	0	0
lncRNA, splice	0	0	0	0
lncRNA, prom.	0	1	1	1
miRNA	3	0	0	1

WGS: whole-genome sequencing. PCG: protein-coding gene. CDS: coding sequence. UTR: untranslated region. Prom.: promoter. lncRNA: long non-coding RNA. miRNA: microRNA. Threshold: q < 0.25.

Table 11. Number of hits in the OncoDriveCLUSTL analysis.

Feature	Cell lines (N = 37)	TCGA, WGS (N = 59)	Tumors, paired (N = 27)	Tumors, unpaired (N = 70)
PCG, CDS	13	2	2	46
PCG, splice	1	0	0	0
PCG, UTR	5	0	0	63
PCG, prom.	2	0	0	8
lncRNA, exon	1	1	0	24
lncRNA, splice	0	0	0	0
lncRNA, prom.	5	0	0	9
miRNA	0	0	0	3

WGS: whole-genome sequencing. PCG: protein-coding gene. CDS: coding sequence. UTR: untranslated region. Prom.: promoter. lncRNA: long non-coding RNA. miRNA: microRNA. Threshold: analytical q < 0.25.

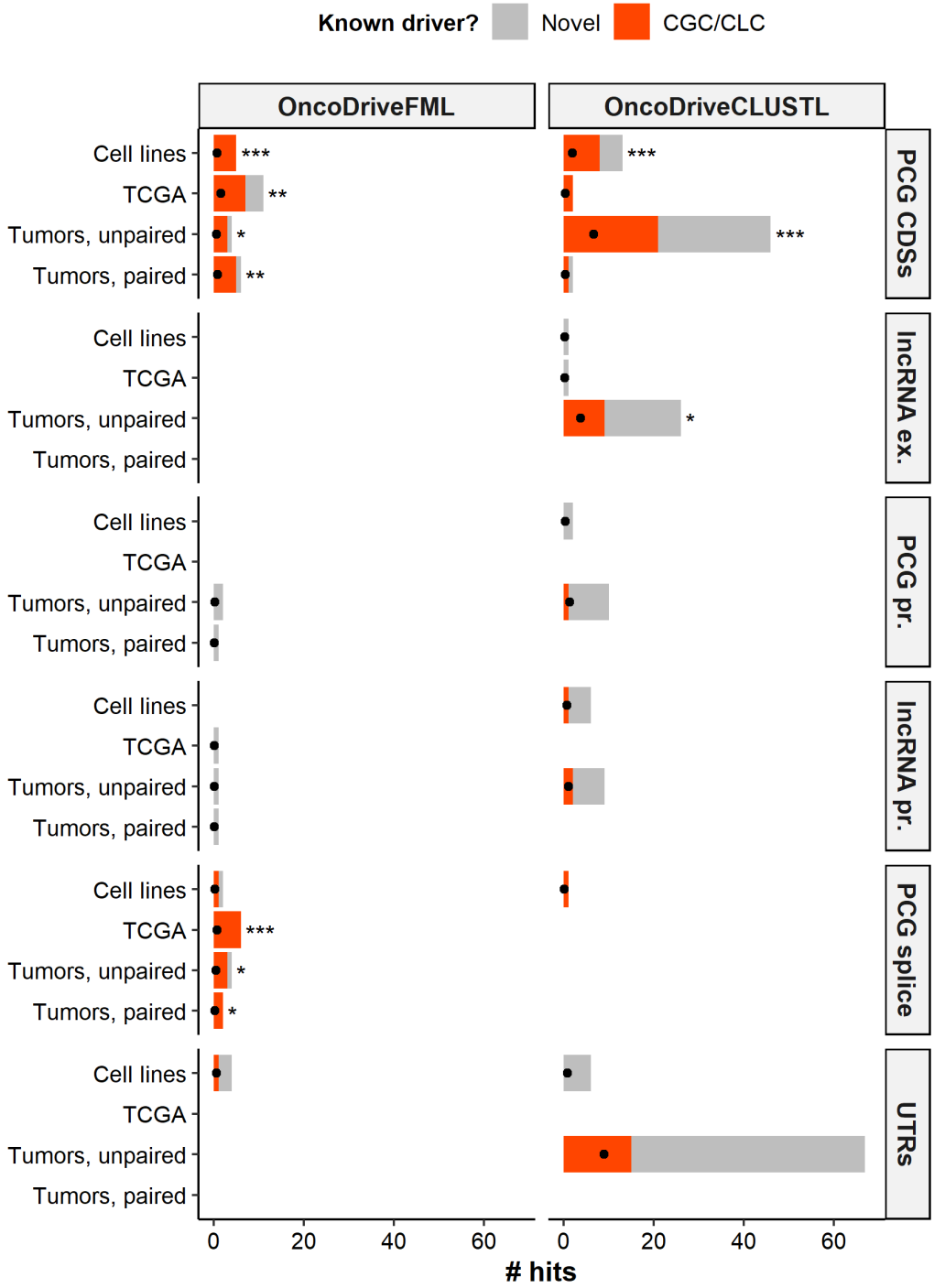


Figure 20. Overlap of driver hits with the Cancer Gene Census (CGC) or the Cancer LncRNA Census (CLC). *Black dots represent the expected number of CGC/CLC hits based on the proportion of CGC/CLC genes among our targets. The observed vs. expected numbers of CGC/CLC and non-CGC/CLC hits were compared using Fisher's exact tests and p values were adjusted to control the false discovery rate (*q < 0.05; **q < 0.01; ***q < 0.001). TCGA: The Cancer Genome Atlas. PCG: protein coding gene. CDS: coding sequence; lncRNA: long non-coding RNA; UTR: untranslated region; ex.: exon; pr.: promoter.*

3.3.6.2. Driver discovery in CDSs

To test OncoDriveFML and OncoDriveCLUSTL, we first applied them to CDSs of PCGs. As a first performance metric, we studied the quantile-quantile plots of p values. Both tools perform one statistical test for each mutated feature that passes their internal thresholds (***Section 3.2.9***). Assuming that most of the tests are expected to be non-significant, most p values should follow a uniform distribution. Therefore, in a quantile-quantile plot of observed vs. expected p values under a uniform distribution, most points should be along the diagonal, and the few significant findings should be noticeably above the diagonal. Indeed, this behavior was roughly observed for OncoDriveFML, although slight inflation of p values was detected in the paired datasets (***Figure 21***). On the other hand, OncoDriveCLUSTL consistently reported inflated p values, which caused a large number of significant findings (***Figure 22*** and ***Table 11***). Therefore, the results of OncoDriveCLUSTL were likely to contain a large proportion of false positives.

Importantly, the analyses shown here had already been optimized within the constraints of each tool. Regarding OncoDriveFML, we had tested it using functional impact scores from other sources, including phyloP (conservation), and FATHMM-MKL (machine learning-based). However, they either failed to detect any cancer genes or they reported similar findings to the analyses that used CADD scores (data not shown). Regarding OncoDriveCLUSTL, its parameters had been optimized as suggested in the original manuscript, and therefore other combinations of parameters performed even worse (Arnedo-Pac et al., 2019). Furthermore, input of OncoDriveCLUSTL had already been modified to remove MNVs, which we had identified as a major source of false positive findings.

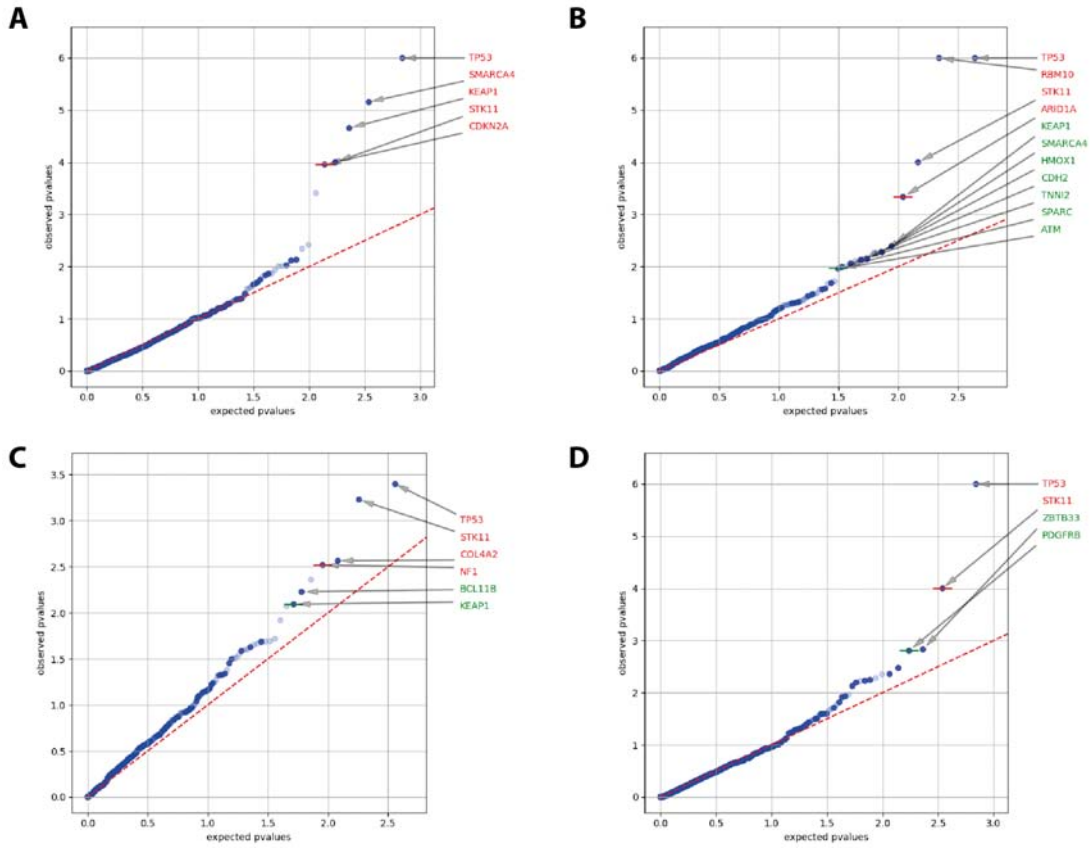


Figure 21. Driver discovery in coding sequences of protein-coding genes using OncoDriveFML. *Quantile-quantile plots of observed vs. expected $-log_{10}(p)$ values are shown. The red dashed line represents the theoretical uniform distribution. The top 10 results are highlighted. In red: $q < 0.01$. In green: $q < 0.25$. Cohorts:* ***A****. Cell lines (N = 37);* ***B.*** *TCGA-LUAD, WGS (N = 59);* ***C****. Primary tumors, paired (N = 27);* ***D****. Primary tumors, unpaired (N = 70).*

Next, we evaluated the overlap of our hits with the set of known cancer driver genes from the CGC. For OncoDriveFML, most of the hits were CGC genes (20/26, 77%) (***Figure 20***). These included well-known major LUAD driver genes, such as *TP53*, *STK11*, *ARID1A*, *SMARCA4*, *KEAP1*, *CDKN2A*, and *NF1* (Collisson et al., 2014). For OncoDriveCLUSTL, the overlap with the CGC was lower (32/63, 51%), which was consistent with a higher false positive rate. However, the clustering-based approach detected key LUAD drivers that were missed by OncoDriveFML, such as *KRAS* (Collisson et al., 2014). *KRAS* is an oncogene that recurrently harbors mutations at its twelfth codon, making this mutational pattern more easily detectable by clustering-based methods. Both driver discovery tools found a proportion of CGC genes that was higher than

expected by chance (Fisher's exact test, q < 0.05) in all cohorts except for OncoDriveCLUSTL in the paired cohorts. Overall, both OncoDriveFML and OncoDriveCLUSTL successfully detected known LUAD drivers in CDSs, but OncoDriveCLUSTL likely had a high false positive rate.

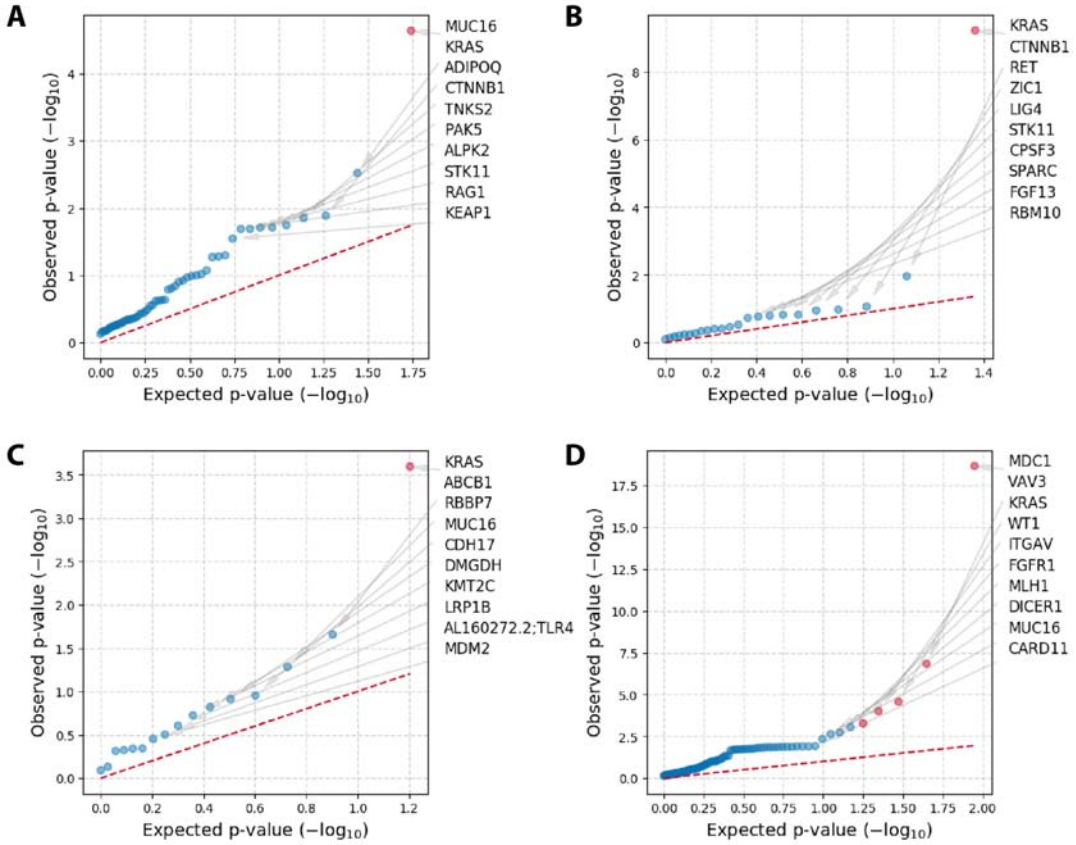


Figure 22. Driver discovery in coding sequences of protein-coding genes using OncoDriveCLUSTL. *Quantile-quantile plots of observed vs. expected $-\log_{10}(p)$ values are shown. The red dashed line represents the theoretical uniform distribution. The top 10 results are highlighted. In red: q < 0.01. Cohorts:* ***A****. Cell lines (N = 37);* ***B.*** *TCGA-LUAD, WGS (N = 59);* ***C****. Primary tumors, paired (N = 27);* ***D****. Primary tumors, unpaired (N = 70).*

3.3.6.3. Driver discovery in lncRNA exons

Selection of candidate drivers

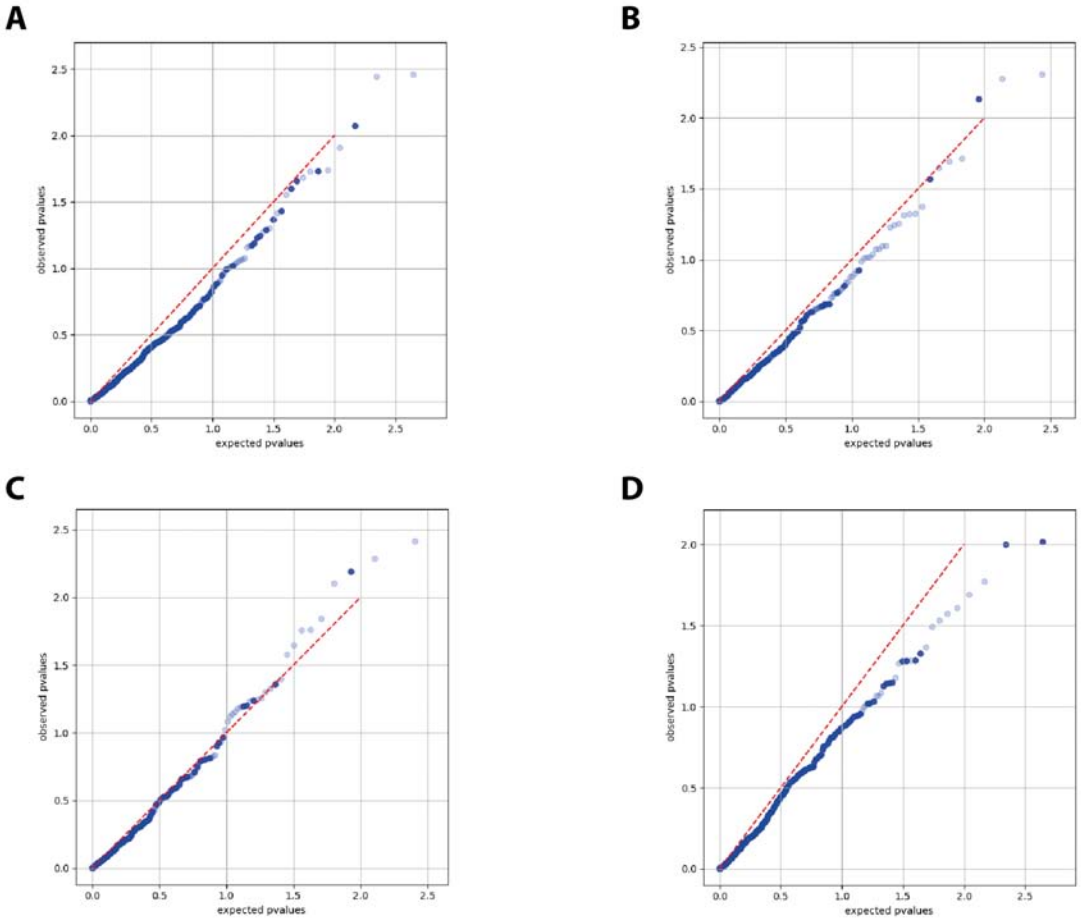


Figure 23. Driver discovery in lncRNA exons using OncoDriveFML. *Quantile-quantile plots of observed vs. expected –log10(p) values are shown. The red dashed line represents the theoretical uniform distribution. The top 10 results are highlighted. In red: q < 0.01. Cohorts:* ***A****. Cell lines (N = 37);* ***B****. TCGA-LUAD, WGS (N = 59);* ***C****. Primary tumors, paired (N = 27);* ***D****. Primary tumors, unpaired (N = 70).*

In lncRNA exons, OncoDriveFML did not detect any drivers in any of the datasets (***Figure 23***). However, in contrast to the analysis in CDSs, here p values looked deflated, especially in the unpaired analyses. This suggested an underperformance of OncoDriveFML in lncRNA exons. Moreover, the issue persisted when using other functional impact scores, including FATHMM-MKL, RNAsnp, and phyloP (data not shown). On the other hand, p values

from OncoDriveCLUSTL were inflated, suggesting a high false positive rate (***Figure 24***). Nevertheless, the hits from OncoDriveCLUSTL were significantly enriched in CLC lncRNAs in unpaired primary tumors (9/24, 38%) (Vancura et al., 2021) (***Figure 20***). The CLC hits were *HEIH*, *TUSC7*, *CAMTA1-DT*, *SOX2-OT*, *ZEB2-AS1*, *DHRS4-AS1*, *NEAT1*, *HOTAIR*, and *EGFR-AS1*.

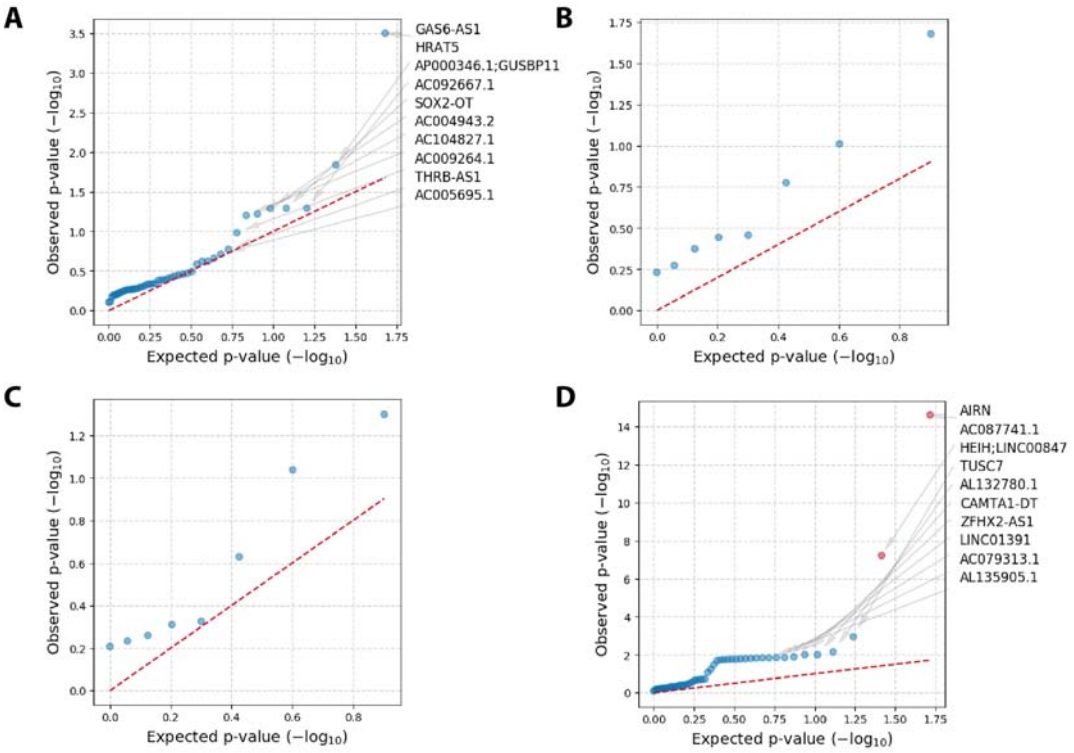


Figure 24. Driver discovery in lncRNA exons using OncoDriveCLUSTL. *Quantile-quantile plots of observed vs. expected –log$_{10}$(p) values are shown. The red dashed line represents the theoretical uniform distribution. The top 10 results are highlighted. In red: q < 0.01. Cohorts:* ***A****. Cell lines (N = 37);* ***B****. TCGA-LUAD, WGS (N = 59);* ***C****. Primary tumors, paired (N = 27);* ***D****. Primary tumors, unpaired (N = 70).*

To select candidate drivers for downstream analyses, considering the poor performance of both driver discovery tools, we explored the top 10 most significant hits and the CLC hits from OncoDriveCLUSTL in each cohort, keeping in mind that most of them were likely to be false positives. For each of the 18 analyzed hits, we explored its genomic region, the distribution of its variants, and the predicted functional impact of its variants (***Supplementary Table 1***). Only 3/18 (17%) of the hits had at least one variant with moderate-

to-high predicted functional impact: *TUSC7*, *SOX2-OT*, and *ZEB2-AS1* (CADD score ≥ 15 and FATHMM-MKL score ≥ 0.9). All of them were cancer-related lncRNAs according to the CLC. Next, we further explored each of the three lncRNAs to determine if their variants might be LUAD drivers.

Detailed analysis of candidate drivers

TUSC7

TUSC7 is an intergenic lncRNA with two annotated isoforms (***Supplementary Figure 1***). One primary tumor from our paired analysis had a high-impact somatic SNV affecting a conserved region in the last exon of both isoforms (chr3:116716439_T/A) (***Table 12***). The VAF was 31% which, accounting for tumor purity, suggested that the SNV was heterozygous or subclonal. The SNV did not affect any known regulatory elements. However, because most of the exonic sequence of *TUSC7* is not conserved, and intronic parts are highly conserved, it cannot be ruled out that the function of the *TUSC7* locus, if any, may be RNA-independent, and that it may contain as-yet uncharacterized regulatory DNA elements. In agreement with this, a small RNA gene (*RF01879*) is immediately downstream of the mutated region in *TUSC7*. Regulatory DNA elements are pervasively transcribed, and therefore *RF01879* may originate from such a type of element (Ibrahim et al., 2018).

Remarkably, external evidence suggests that *TUSC7* is a tumor suppressor locus in multiple cancer types. In particular, the *TUSC7* locus is recurrently deleted in three non-LUAD cohorts from TCGA: adrenocortical carcinoma, uterine carcinosarcoma, and uterine corpus endometrial carcinoma (Athie et al., 2020) (***Supplementary Figure 1***). *TUSC7* is the only gene within the recurrently deleted region (besides the overlapping *RF01879*), suggesting that the locus may be a bona fide tumor suppressor, at least in those cancer types. Furthermore, *TUSC7* has been reported as a TP53-regulated tumor suppressor lncRNA in various cancer cell lines, including the A549 LUAD cell line (Liu et al., 2013). Moreover, another study reported that *TUSC7* downregulation is associated with poor prognosis in TCGA-LUAD and that *TUSC7* has a tumor suppressor role in LUAD cell lines (Zhou et al., 2019). Both reports experimentally supported an RNA-dependent function of *TUSC7*.

Table 12. High-impact variants in candidate driver lncRNAs.

lncRNA	Variant	MKL	CADD	Struc	VAF
TUSC7	chr3:116716439_T/A	0.98	18.3	0.24	31%
SOX2-OT	chr3:181741757_CC/AG	0.99	17.6	0.05	29%
SOX2-OT	chr3:181741840_G/A	0.99	16.7	0.30	50%
ZEB2-AS1	chr2:144518562_A/T	0.92	17.4	0.20	19%
ZEB2-AS1	chr2:144519646_A/G	0.90	16.0	0.15	21%
ZEB2-AS1	chr2:144520222_A/G	1.00	22.0	0.10	55%
ZEB2-AS1	chr2:144520355_G/C	1.00	19.9	0.10	24%

Genomic coordinates use the hg38 reference genome. The following functional impact scores are reported: "MKL": FATHMM-MKL (range = 0-1, higher score means higher impact); CADD (Phred scale, higher score means higher impact); "Struc": RNAsnp p-value (lower score means higher impact). The variant allele frequency (VAF) in the affected sample is also included. For variants affecting multiple nucleotides or multiple transcripts, the most deleterious impact scores were selected.

To confirm the previously described roles of *TUSC7* in LUAD, first we determined the expression of *TUSC7* in external RNA sequencing (RNA-Seq) datasets. Median *TUSC7* expression in normal lung samples from the Genotype Tissue Expression (GTEx) project was 0 (de Goede et al., 2021). Median *TUSC7* expression was also 0 in tumor and matched normal lung samples from TCGA-LUAD (***Figure 25A***). Only 51/537 (9%) TCGA-LUAD samples had detectable expression of *TUSC7*. No survival analysis was performed because of the mostly undetectable expression of *TUSC7*. We noticed that the previous report from Zhou et al had excluded the ~91% of TCGA-LUAD samples that had undetectable *TUSC7* expression, thus gravely biasing the results of their survival analysis (Zhou et al., 2019).

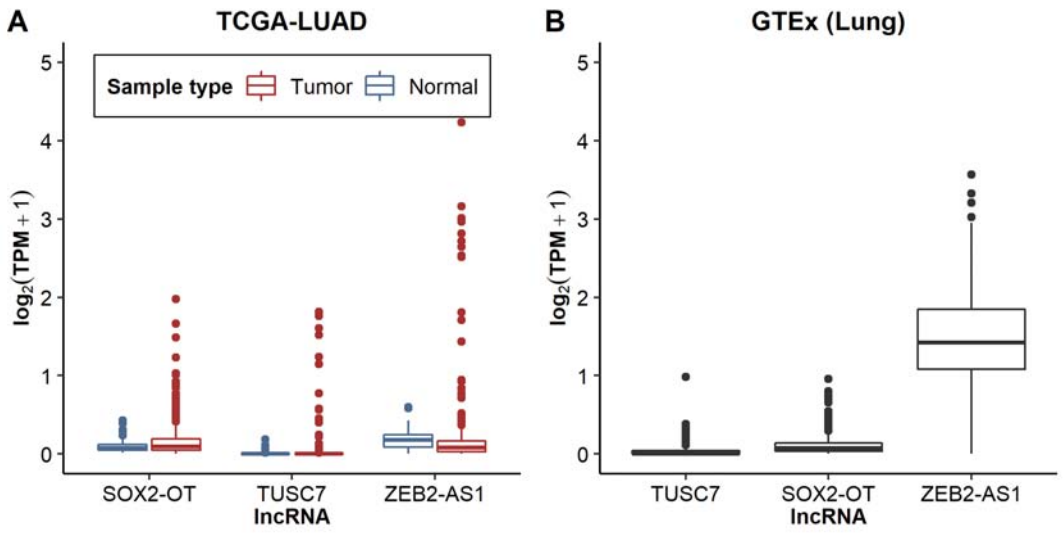


Figure 25. Expression of four candidate driver long non-coding RNAs (lncRNAs) *in: A. Lung adenocarcinoma and paired normal lung tissue samples from The Cancer Genome Atlas (TCGA-LUAD); **B**. Normal lung samples from the Genome Tissue Expression (GTEx) Project.* TPM: transcripts per million.Although *TUSC7* RNA expression was not detected in external RNA-Seq datasets, it was detected by previous studies in LUAD using quantitative polymerase chain reaction (qPCR) (Liu et al., 2013; Zhou et al., 2019). In a preliminary analysis from our group, we detected *TUSC7* RNA in our cohort of LUAD primary tumors and matched normal samples by qPCR, and we found that *TUSC7* is downregulated in LUAD primary tumors compared to normal samples, in agreement with previous reports (data not shown).

If *TUSC7* has low expression, it is unlikely to act as a ceRNA. Nevertheless, according to the online tool miRcode, the variant disrupted a binding site for the miR-27abc/27a-3p family of oncomiRs, a disruption that, if functional, might promote oncogenesis (Jeggari et al., 2012; Zhang et al., 2019).

In conclusion, although *TUSC7* may be a tumor suppressor locus in some cancers, it is unclear whether it has an RNA sequence-dependent role in LUAD. Critically, evidence of *TUSC7* expression in LUAD is inconsistent, and future work should clarify the discrepancies between RNA-Seq and qPCR and determine the number of copies per cell of *TUSC7* in our LUAD cohort.

SOX2-OT

SOX2-OT is a lncRNA that overlaps the *SOX2* PCG in the same strand (***Supplementary Figure 2*** and ***Supplementary Figure 3***). Over 20 *SOX2-OT* isoforms are annotated in GENCODE v29, and some of them span over 800 kb. Two exonic variants affecting *SOX2-OT* in our cohort of primary tumors had high functional impact (***Table 12***). Both affected a highly conserved ~500 bp region located ~26 kb downstream of *SOX2*. One of them was a dinucleotide variant (chr3:181741757_CC/AG) with VAF = 29%, suggesting that it was heterozygous or subclonal. The second one was a SNV (chr3: 181741840_G/A) with VAF = 50%, and therefore its zygosity was unclear. If the tumor sample had low contamination with normal DNA, the variant may be heterozygous. However, if contamination with normal tissue was high, the tumor was not diploid, or the variant was subclonal, the variant may have been present in more than 50% of the DNA molecules in the cells that harbor it. Finally, because we did not sequence the matched normal sample, the variant may have been germline heterozygous. Both variants only affected some *SOX2-OT* isoforms, as they were located near the 3' ends of only the longest isoforms. None of the variants affected any known regulatory elements or miRNA binding sites, and no recurrent somatic copy number alterations affected the region in TCGA. Interestingly, cap analysis of gene expression data suggested that there was a transcription start site ~250 bp upstream of the mutated region, and it could not be associated with any annotated genes (Andersson et al., 2014) (***Supplementary Figure 3***).

SOX2-OT had low median expression in normal lung (0.05 TPM both in GTEx and in TCGA-LUAD) and in LUAD (0.07 TPM in TCGA-LUAD) (***Figure 25***). Median expression of *SOX2* mRNA was also low in normal lung (0.6 TPM in GTEx; 5.6 TPM in TCGA-LUAD) and moderate-low in LUAD (10.7 TPM in TCGA-LUAD). *SOX2-OT* expression and *SOX2* mRNA expression had a moderate-strong non-linear correlation in LUAD and in normal lung samples (Kendall τ range = 0.5-0.62; ***Figure 26***). Despite being correlated to each other, high *SOX2-OT* expression was associated with a favorable prognosis in TCGA-LUAD (logrank p = 0.038, ***Figure 27A***), but *SOX2* mRNA expression was not (logrank p = 0.19, ***Figure 27B***).

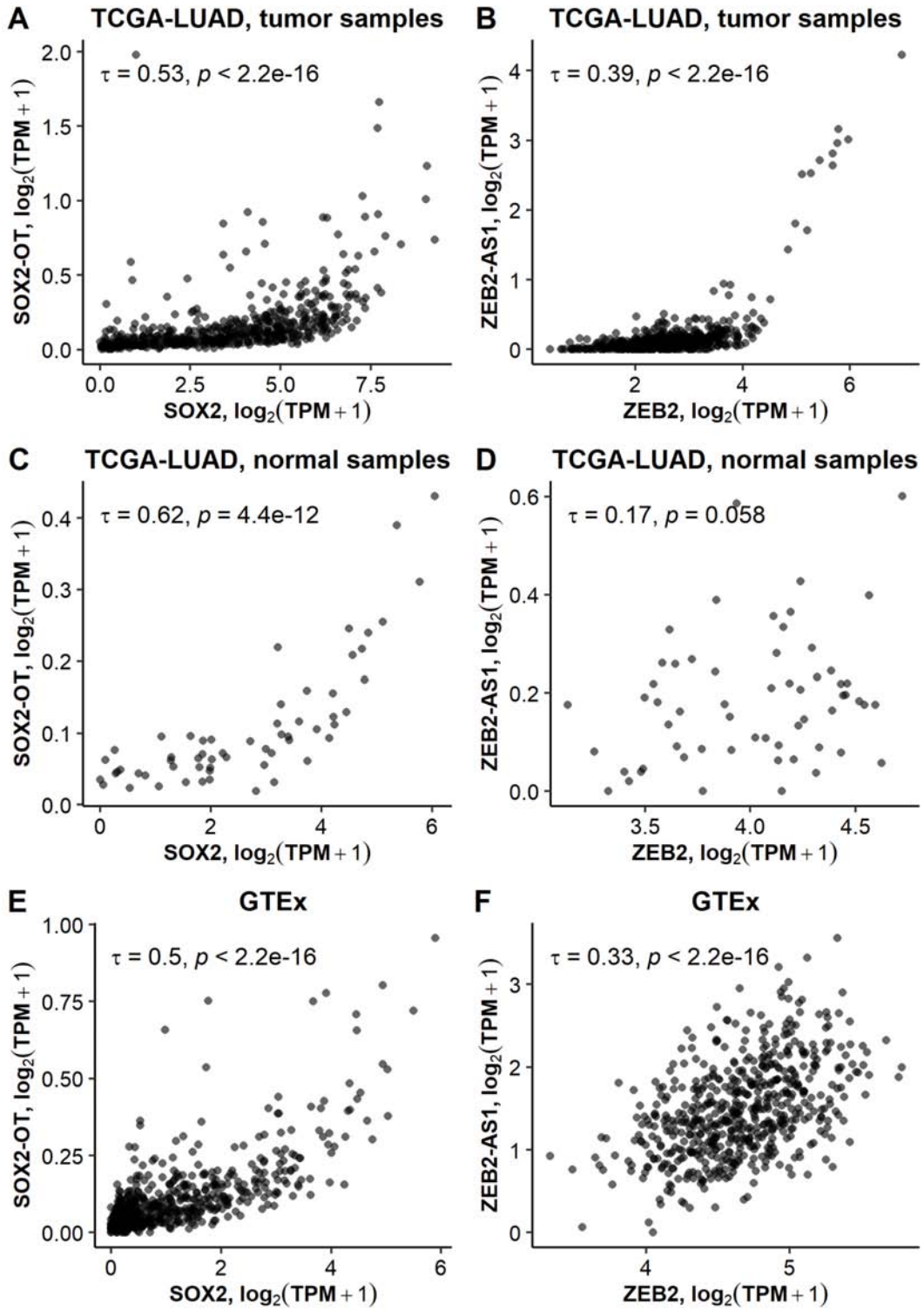


Figure 26. Correlation between expression of lncRNA hits and the mRNA of their overlapping PCGs. *Two correlations were explored: SOX2-OT/SOX2 (**A, C, E**) and ZEB2/ZEB2-AS1 (**B, D, F**). Three datasets were studied: **A, B**: lung adenocarcinoma (LUAD) samples from The Cancer Genome Atlas (TCGA); **C, D**: normal lung from TCGA-LUAD; **E, F**: normal lung from the Genome Tissue Expression (GTEx) Project. Correlations were estimated by Kendall's tau (τ).*

SOX2 is recurrently amplified and may act as an oncogene in small cell lung cancer (Rudin et al., 2012). In addition, previous reports have suggested that *SOX2-OT* is upregulated and has an oncogenic role in NSCLC (Chen et al., 2022; Hou et al., 2014). According to our reanalysis of TCGA-LUAD, *SOX2-OT* was significantly overexpressed in LUAD samples compared to normal lung samples, but the fold change was low (fold change = 1.04; t-test, $p = 1.5 \cdot 10^{-4}$; ***Figure 25***). In previous reports, fold changes were ~3 times higher than in ours (Hou et al., 2014). However, because basal expression of *SOX2-OT* is extremely low, small differences in mean *SOX2-OT* expression due to statistical noise can result in large differences in the fold change (Love et al., 2014). Moreover, whereas previous reports found an association between high *SOX2-OT* expression and poor prognosis in a different NSCLC cohort, we found the opposite trend in TCGA-LUAD, which contradicts its previously proposed oncogenic role (Hou et al., 2014) (***Figure 27A***). However, Hou et al combined various NSCLC subtypes whereas we focused exclusively on LUAD. Moreover, due to the low basal expression of *SOX2-OT*, assignment of patients to the "*SOX2-OT* high" and "*SOX2-OT* low" expression groups may be more dependent on statistical noise than on real biological differences.

Overall, the relevance of the *SOX2-OT* in LUAD is unclear, and its recurrent variants only affected the terminal nucleotides of some of its isoforms. Furthermore, the variants affected a ~500 bp highly conserved region that had a transcription start site immediately upstream, suggesting that the region may contain a currently unannotated transcript. Finally, the mechanistic and functional relationship between *SOX2-OT* and *SOX2* remains to be elucidated.

ZEB2-AS1

ZEB2-AS1 (previously known as *ZEB2* natural antisense transcript) is an antisense lncRNA that overlaps the *ZEB2* PCG (***Supplementary Figure 4***). *ZEB2-AS1* is located within a ~5 Mb region that is recurrently deleted in diffuse large B cell lymphoma (Athie et al., 2020). The only CGC gene within the recurrently deleted region is *ACVR2A*, which is a tumor suppressor gene.

Moderate-to-high-impact variants in *ZEB2-AS1* exons overlapped with the first intron, the 5'-UTR, and upstream regions of *ZEB2* (***Table 12***). The

variants did not affect any predicted miRNA binding sites. All VAFs but one were <25%, which was consistent with the variants being heterozygous and/or subclonal. The exception was a recurrent variant in two tumor samples (chr2:144520222_A/G) whose VAF was 55%. Because we had not sequenced the matched normal sample, we could not rule out that the variant was germline heterozygous. Furthermore, an intronic *ZEB2-AS1* variant was predicted to have a high functional impact similarly to the exonic variants (chr2:144519379_G/T: FATHMM-MKL score = 0.97, CADD score = 19.5), which suggested that the putative effect of the variants may be independent of the *ZEB2-AS1* RNA sequence.

In lung cancer, *ZEB2-AS1* has been reported as overexpressed and it has been suggested to be oncogenic (Guo et al., 2018). However, in TCGA-LUAD, we found no significant differences in *ZEB2-AS1* expression between LUAD and normal lung (t-test, p = 0.51) (***Figure 25***). Furthermore, high expression of *ZEB2-AS1* was associated with high overall survival in TCGA-LUAD (logrank p = 0.01; ***Figure 27C***), whereas *ZEB2* was borderline significant (logrank p = 0.08, ***Figure 27D***). These observations contradict the previously proposed oncogenic role of both genes, but they are consistent with the fact that the loci are deleted in some cancers (Athie et al., 2020). The discrepancies between TCGA-LUAD and Guo et al's report could not be evaluated further because Guo et al did not specify the lung cancer subtypes of their patients.

Previous reports have proposed a mechanistic relationship between *ZEB2* and *ZEB2-AS1*. In particular, *ZEB2-AS1* promotes ZEB2 translation by binding to its first splice donor site, causing intron retention that exposes an internal ribosomal entry site in the *ZEB2* mRNA (Beltran et al., 2008). The resulting increase in ZEB2 expression has been linked to epithelial-mesenchymal transition in various cancer cell lines (not including LUAD) (Beltran et al., 2008). In agreement with this, we found that expression of *ZEB2-AS1* and of *ZEB2* mRNA were weakly and positively correlated in LUAD and in normal lung (Kendall τ range = 0.17-0.39; ***Figure 26***). However, a significant correlation is not enough evidence to prove that *ZEB2-AS1* modulates *ZEB2*. In fact, the stoichiometry of *ZEB2* mRNA and *ZEB2-AS1* in lung samples casts doubts on a biologically relevant interaction between both RNAs. In primary

tumors from TCGA-LUAD, median expression of *ZEB2* was ~150x higher than that of *ZEB2-AS1*: 8.8 TPM vs. 0.06 TPM, respectively (***Figure 25***). *ZEB2* expression was also at least an order of magnitude higher than that of *ZEB2-AS1* in normal lung: 24.1 TPM vs. 1.67 TPM in GTEx, and 39.5 TPM vs. 0.13 TPM in normal samples from TCGA-LUAD. Therefore, in lung tissue, *ZEB2-AS1* may have too few molecules per cell to modulate *ZEB2 in vivo*.

In conclusion, it is unclear whether *ZEB2-AS1* has an RNA sequence-dependent function in LUAD and whether it acts via *ZEB2*. Experiments will be required to test the functions and the mechanisms of *ZEB2-AS1* in LUAD.

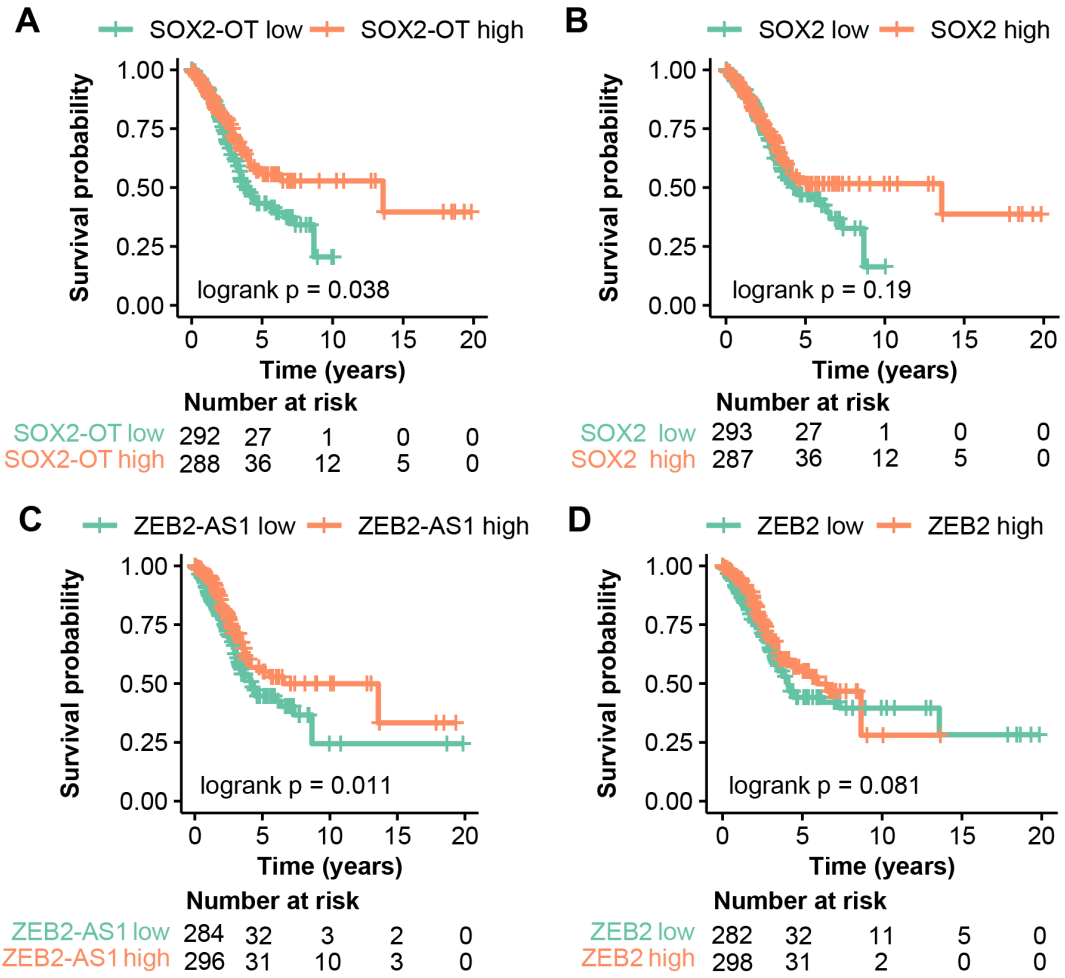


***Figure 27. Survival analyses on candidate driver long non-coding RNAs and their overlapping protein coding genes**: **A**. SOX2-OT; **B**. SOX2; **C**. ZEB2-AS1; **D**. ZEB2. Data from lung adenocarcinoma patients from The Cancer Genome Atlas (TCGA-LUAD). Patients were split in two expression groups using the median as a threshold. Logrank p values are shown.*

3.3.6.4. Driver discovery in miRNAs

OncoDriveFML predicted four driver miRNAs: *MIR3153* (in our primary tumors), *MIR4725*, *MIR4650-1,* and *MIR4712* (in our cell lines). In addition, OncoDriveCLUSTL detected 3 driver miRNAs: *MIR3153*, *MIR8064*, and *MIR3689F* (all of them in our primary tumors).

To further characterize miRNA variants, we developed an in-house pipeline to annotate, prioritize, and predict the functional impact of variants in a miRNA-centric manner. We reasoned that, because miRNAs are short and highly conserved, we may not be able to detect biologically active miRNA variants based on recurrence alone. Therefore, we did not only study the recurrent variants detected by the driver discovery tools, but also the non-recurrent ones. The results are detailed in ***Sections 3.3.7-3.3.8.3***.

3.3.6.5. Driver discovery in splice regions

In splice regions of PCGs, OncoDriveFML found 14 hits (***Table 10***) and OncoDriveCLUSTL found 1 (***Table 11***). No hits were detected in splice regions of lncRNAs. The PCG hits from OncoDriveFML were enriched in CGC genes (***Figure 20***). The hits affected 11 unique genes, as some of them were identified in more than one cohort. In particular, *TP53* was a hit in our primary tumors and in our cell lines. In addition, *STK11* was a hit in our paired and unpaired primary tumors and in TCGA-LUAD. Moreover, *RB1* was a hit in our paired and unpaired primary tumors. Other hits included *NF1*, *MET*, and *RBM10* (all three in TCGA-LUAD). All of these genes are known LUAD drivers that undergo recurrent splice site mutations in LUAD (Bailey et al., 2018; Collisson et al., 2014; Shiraishi et al., 2018).

Next, we focused on the hits from TCGA-LUAD because they had matched RNA-Seq alignment files, which allowed us to evaluate the impact of the splice site variants on mRNA processing. The TCGA-LUAD hits were *STK11*, *NF1*, *MET*, *RBM10*, *MUC16*, and *COL3A1*. We used the MAJIQ tool to identify alterations in splice junctions in mutant samples (Vaquero-Garcia et al., 2016), and then we confirmed the results by inspecting the RNA-Seq reads on IGV.

Exon skipping

Gene	Patient	Exon #	PSI
STK11	TCGA-64-1678	4	0.44
STK11	TCGA-55-6972	7	0.19
STK11	TCGA-05-4420	7	0.73
NF1	TCGA-78-7146	37	0.76
NF1	TCGA-44-2659	38	0.43
MET	TCGA-50-5930	11	0.44
MET	TCGA-50-6597	14	0.82
RBM10	TCGA-49-6742	2	0.41
RBM10	TCGA-49-4486	9	0.22

Intron retention

Gene	Patient	Intron #	PSI
STK11	TCGA-64-1678	3	NA
STK11	TCGA-78-7535	5	0.24
STK11	TCGA-55-6972	7	0.74

Cryptic splice site

Gene	Patient	Exon #	PSI
STK11	TCGA-78-7535	6	0.39

Partial intron retention

Gene	Patient	Intron #	PSI
MET	TCGA-50-5930	11	NA

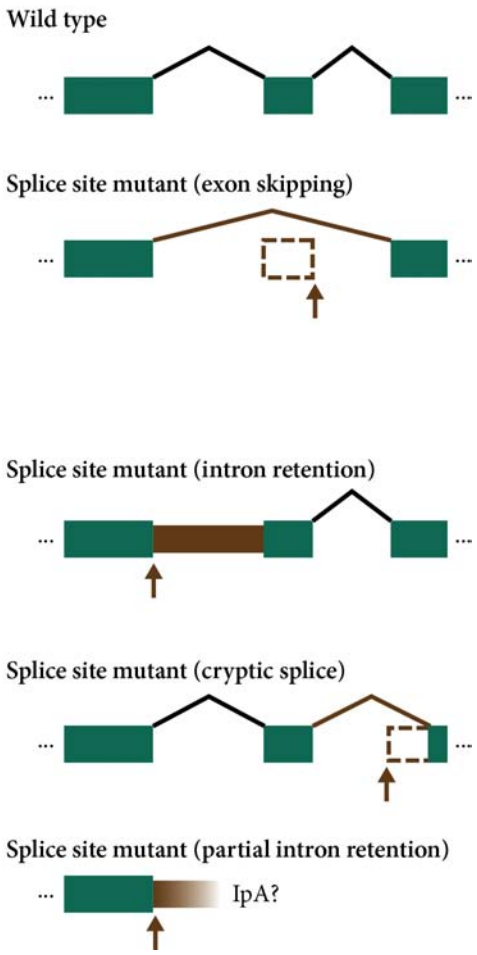


Figure 28. Impact of splice site mutations on RNA splicing in TCGA-LUAD. *Mutations were identified in whole-genome sequencing data, and their effect on RNA splicing was evaluated in matched RNA sequencing data. Splice site mutations are classified based on their effect on RNA splicing. For each effect, an example of how they may affect a hypothetical RNA is depicted. Brown arrows denote the position of the hypothetical mutation. For each aberration, the identifier of the affected patient and the position of the affected exon or intron is shown. Transcripts used as a reference for counting exons and introns were: ENST00000326873.11 (STK11), ENST00000356175.7 (NF1), ENST00000397752.7 (MET), and ENST00000377604.7 (RBM10). PSI: "percentage spliced in", proportion of transcripts that were estimated to undergo a splicing event. NA: not available (the aberration was not detected by MAJIQ, but it was detected on IGV). IpA: intronic polyadenylation.*

In *STK11*, four samples were mutated at splice sites in the TCGA-LUAD WGS dataset (***Supplementary Figure 5***). Using the transcript ENST00000326873.11 as a reference, the variants affected introns 3, 5, 6, and 7. All four variants affected mRNA processing (***Figure 28***). In particular, one at the splice acceptor site of intron 3 caused exon 4 skipping in some transcripts and intron 3 retention in others; one at the splice acceptor site of intron 5 caused usage of a cryptic splice site within exon 6 that led to a 35 nt deletion, as well as intron retention in other transcripts; one at the splice acceptor site of intron 6 caused intron 6 retention, as well as some minor skipping of exon 7 in other transcripts; and one at the splice donor site of intron 7 caused exon 7 skipping. *STK11* is a recurrently inactivated tumor suppressor gene in LUAD (Collisson et al., 2014). Remarkably, its kinase activity is mapped to more than half of the protein sequence (amino acids 49-309 out of 433, UniProt ID: Q15831), and therefore truncating variants at most of the *STK11* coding sequence are likely to disrupt its function.

In *NF1*, two somatic variants affected splice sites in TCGA-LUAD WGS data (***Supplementary Figure 6***). Using the transcript ENST00000356175.7 as a reference, the variants affected introns 37 and 38. The variants caused skipping of exons 37 and 38, respectively (***Figure 28***). *NF1* is a recurrently inactivated tumor suppressor gene in LUAD, and its mutations are not clustered in any specific parts of the gene (Collisson et al., 2014).

In *MET*, two variants affected splice regions (***Supplementary Figure 7***). Using transcript ENST00000397752.7 as a reference, one variant affected the splice donor site in intron 11, causing exon skipping as well as partial intron retention in other transcripts (***Figure 28***). Here, we define "partial intron retention" as an event in which some reads span the exon-intron junction, all or most of them harboring the splice site variant, but the whole intron is not retained: instead, coverage drops to zero or near zero within the intron. Such patterns have been attributed to intronic polyadenylation (Zhao et al., 2021b). In the case of our identified *MET* variant, the first ~300 nt in intron 11 were expressed, and then coverage dropped to nearly zero (***Supplementary Figure 11***). Although we did not find a canonical AAUAAA polyadenylation signal within the expressed intronic region, we found two instances of the second

most frequent polyadenylation signal in humans, AUUAAA (Beaudoing et al., 2000), supporting that the variant may have caused intronic polyadenylation (***Supplementary Figure 11***). On the other hand, one variant affected the third nucleotide of intron 14, causing skipping of exon 14 in all reads (***Figure 29***). Skipping of exon 14 of *MET* is a recurrent driver event in LUAD and it has special clinical relevance because two drugs, capmatinib and tepotinib, have been approved for the treatment of LUAD tumors that harbor these alterations (Frampton et al., 2015; Mathieu et al., 2022). Importantly, the variant discussed here was located 1 bp downstream from the canonical "GT" splice donor sequence, but it still affected splicing. Therefore, our results highlight that variants at splice regions beyond the canonical splice donor and acceptor sequences can cause clinically relevant aberrations in splicing.

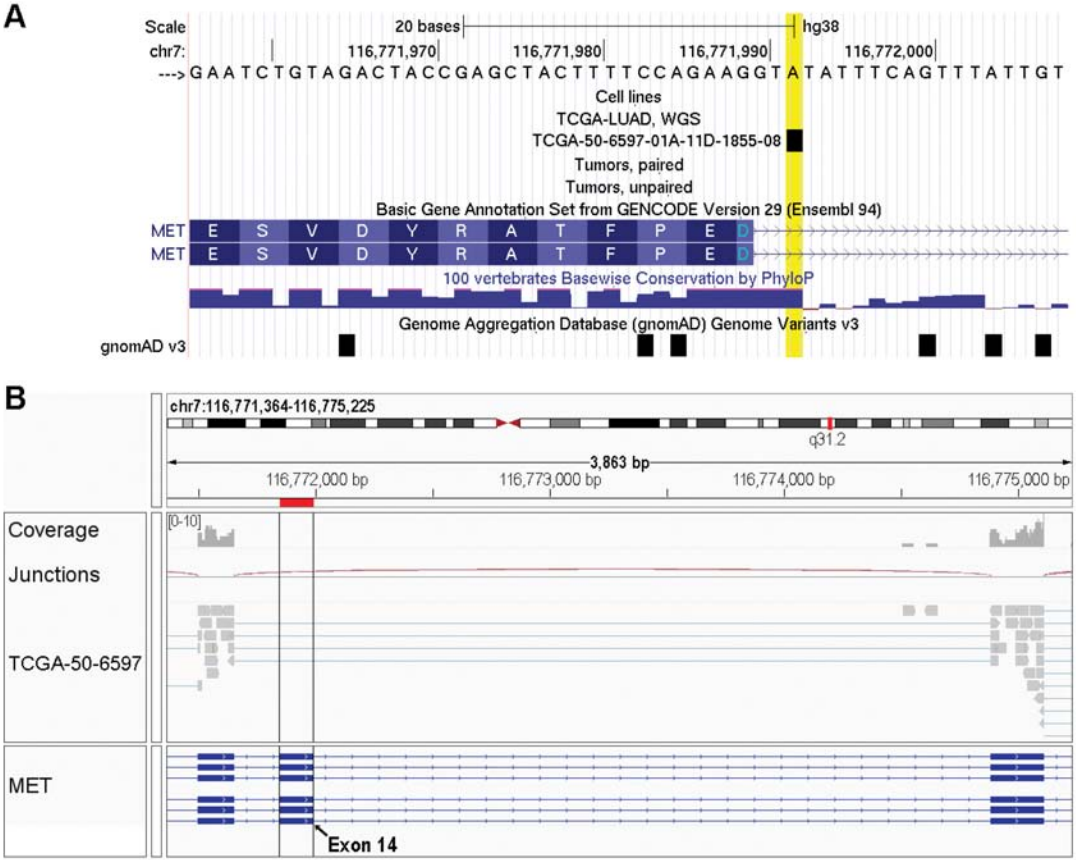


Figure 29. Skipping of MET exon 14 associated with a somatic variant in the third position of intron 14 in a TCGA-LUAD primary tumor.A*. Genomic position of the variant (highlighted in yellow). Figure generated in the UCSC Genome Browser.* ***B****. RNA sequencing data shows that exon 14 is skipped. Figure modified from the Integrative Genomics Viewer.*

Regarding *RBM10*, two WGS samples from TCGA-LUAD had somatic variants in splice regions (***Supplementary Figure 8***). Using the transcript ENST00000377604.7 as a reference, one variant affected the third nucleotide of intron 2 and another one affected the fifth nucleotide of intron 9. Both variants were associated with aberrations in mRNA splicing (***Figure 28***): the one at intron 2 caused exon 2 skipping, whereas the one at intron 9 caused exon 9 skipping (***Figure 30***). *RBM10* is recurrently inactivated in LUAD and its mutations do not cluster in any specific parts of its sequence (Collisson et al., 2014). Although *RBM10* has been consistently detected as one of the top LUAD driver genes, its role in LUAD is still unclear (Bailey et al., 2018; Collisson et al., 2014). RBM10 modulates splicing, mainly by promoting exon skipping, and inactivation of RBM10 may cause transcriptome-wide changes in splicing patterns (Seiler et al., 2018). However, further research is needed to fully elucidate how RBM10 is involved in tumorigenesis in LUAD.

MUC16 harbored three somatic variants in splice regions, one of which affected a splice site (***Supplementary Figure 9***). They affected introns 3, 22, and 78 of transcript ENST00000397910.8. However, we found no evidence of any of the variants causing splicing aberrations. *MUC16* harbored a large number of variants along the whole gene, which encodes the second largest human protein, being ~14 500 amino acids long (Lawrence et al., 2013). Previous studies have also detected a high mutation rate in *MUC16*, but it is usually considered a false positive hit from driver discovery methods that do not account for gene length or BMR heterogeneity (Lawrence et al., 2013). Although the tools employed in our work should have accounted for both confounding factors, our low sample size may have impaired the construction of an accurate background, and this may have caused *MUC16* to be detected as a false positive hit.

Finally, *COL3A1* had two variants in splice regions: one at the splice donor site of intron 2 (reference transcript: ENST00000304636.7), and one at the 9th last nucleotide of intron 24 (***Supplementary Figure 10***). However, neither of them were associated with detectable aberrations in mRNA splicing. This, together with the lack of evidence for driver mutations in *COL3A1* in external studies, suggested that it may be a false positive hit.

In conclusion, in the TCGA-LUAD dataset, we have found 4 PCGs that accumulate splice-altering driver mutations at their intronic splice regions: *STK11*, *NF1*, *MET*, and *RBM10*. All of them are known LUAD driver genes, and targeted therapies for *MET* exon 14 skipping are clinically available (Bailey et al., 2018; Collisson et al., 2014; Mathieu et al., 2022). Importantly, variants beyond the 2-nt splice donor and acceptor sites caused splicing aberrations, highlighting the importance of analyzing wider regions, especially the third and fifth intronic nucleotides (Shiraishi et al., 2018). In future work, it may be of interest to experimentally analyze the RNA of our primary tumors that harbored mutations in splice regions of genes such as *TP53* and *RB1* to confirm whether the variants caused splicing aberrations.

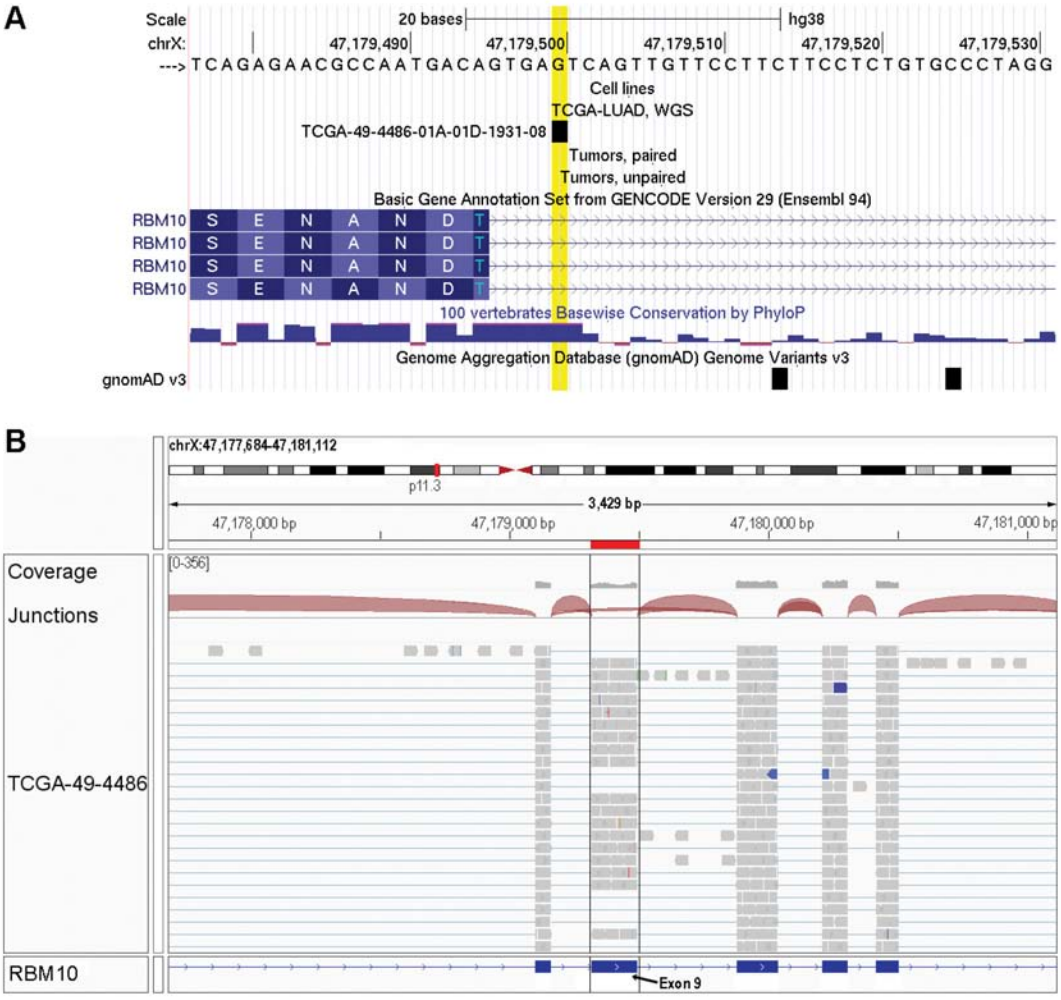


Figure 30. Skipping of RBM10 exon 9 associated with a somatic variant in the fifth position of intron 9 in a TCGA-LUAD primary tumor. ***A****. Genomic position of the variant (highlighted in yellow). Figure generated in the UCSC Genome Browser.* ***B****. RNA sequencing data shows that exon 9 is skipped. Figure modified from the Integrative Genomics Viewer.*

3.3.6.6. Driver discovery in promoters

OncoDriveFML reported 3 hits in PCG promoters and 3 hits in lncRNA promoters (***Table 10***). On the other hand, OncoDriveCLUSTL found 10 hits in PCG promoters and 14 hits in lncRNA promoters (***Table 11***). Only the *PLAUR* promoter was detected by both tools (in our unpaired primary tumors). No OncoDriveFML hits were in the CGC or CLC (***Figure 20***). Regarding the OncoDriveCLUSTL hits, only one promoter modulated a CGC PCG: *PIK3CA* (in primary tumors). In addition, two modulated CLC lncRNAs: *FENDRR* (in cell lines and in primary tumors) and *MYCNUT* (in primary tumors).

Three promoter variants had high predicted impact (***Table 13***). We tested if they affected transcription factor binding using PROMO (Messeguer et al., 2002). First, in *PRMT5*, chr14:22929753_G>T was predicted to disrupt a GR-alpha binding site and introduce a FOXP3 and a GR-beta binding site. It also overlapped with a splice site of the lncRNA *PRMT5-DT*, which may explain its high impact (***Supplementary Figure 12***). Although the variant affected a repetitive region, no mapping issues were observed. Second, also in *PRMT5*, chr14:22929717_G>A was predicted to disrupt a TFII-I binding site and introduce an HNF-1B binding site. Finally, in *NKX3-2*, chr4:13544981_C>A was predicted to disrupt an RXR-alpha binding site and introduce one for USF2. No artifacts were detected (***Supplementary Figure 13***).

Table 13. High-impact variants in candidate driver promoters.

Dataset	Variant	Genes	MKL	CADD	VAF
Tumors, p. + unp.	chr14:22929717_G/A	*PRMT5*, *PRMT5-DT*	1.0	22.6	15%
Tumors, unpaired	chr14:22929753_G/T	*PRMT5*, *PRMT5-DT*	1.0	32.0	29%
Tumors, unpaired	chr4:13544981_C/A	*NKX3-2*, *LINC01096*	1.0	20.3	35%

Genomic coordinates use the hg38 reference genome. The following functional impact scores are reported: "MKL": FATHMM-MKL (range = 0-1, higher score means higher impact); CADD (Phred scale, higher score means higher impact). VAF: variant allele frequency; "p. + unp.": paired and unpaired analyses.

To further test the functional impact of variants in all promoter hits, we determined whether they were associated with expression changes in *cis*. For bidirectional promoters, we studied the genes in both directions. For the hits in cell lines, we used transcriptome expression data from the Cancer Cell Line Encyclopedia (CCLE). We found no significant associations between variants in any of the promoter hits and changes in gene expression (***Figure 31***). However, the association between variants in the *PDCD6* promoter and *PDCD6* mRNA expression was borderline significant (t-test, p = 0.08). The *PDCD6* promoter had one variant, chr5:271495-271495_G>A, in two cell lines, H1437 and H2122. The variant was predicted to disrupt a p53 binding site (Messeguer et al., 2002). However, the position had very low evolutionary conservation (phyloP score = -2.9) and the variant had moderate-low predicted functional impact (CADD score = 5.8; FATHMM-MKL score = 0.86). Our only hit in TCGA-LUAD, *AP001107.2*, was not expressed in any of the 59 samples from TCGA-LUAD. Finally, we could not test the hits in our primary tumors because we lacked gene expression data from them. Overall, we could not find any strong hits among promoters using available data.

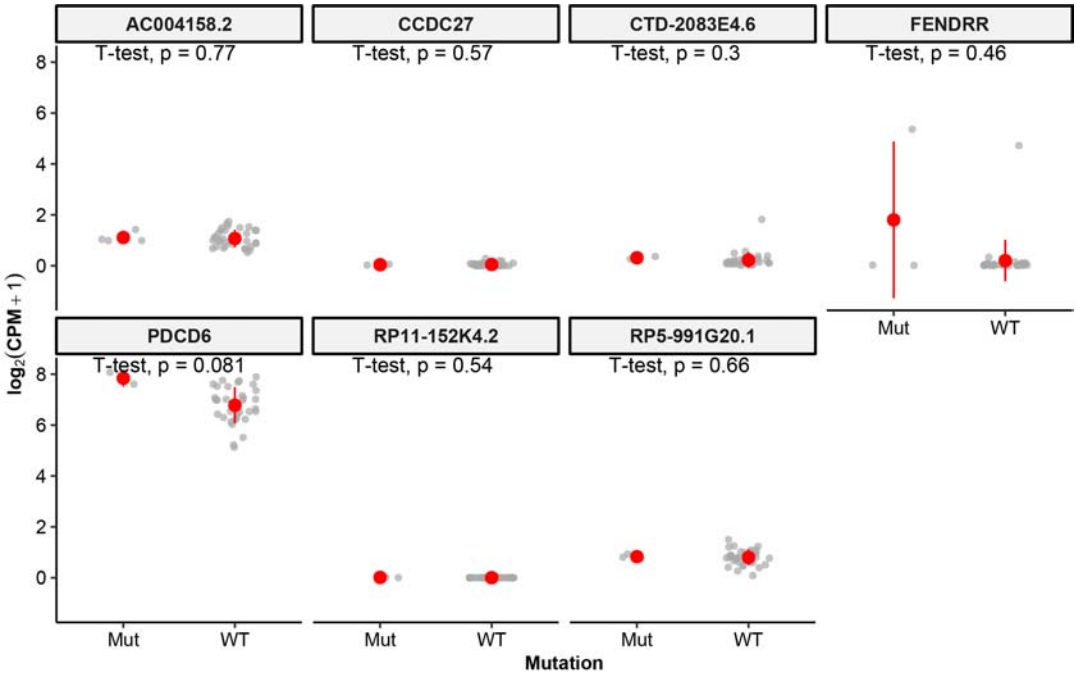


Figure 31. Association between variants in promoter hits and gene expression in cell lines. *Transcriptome expression data were retrieved from the Cancer Cell Line Encyclopedia (CCLE) for our LUAD cell lines. Red points and lines show the mean and the standard deviation. "Mut": mutant; "WT": wild type.*

3.3.6.7. Driver discovery in UTRs

While OncoDriveFML only detected 4 driver UTRs, all of them in cell lines (***Table 10***), OncoDriveCLUSTL reported 68 driver UTRs, 63 of which were in unpaired primary tumors (***Table 11***). Again, OncoDriveFML performed acceptably and OncoDriveCLUSTL had highly inflated p values (***Figure 32***). Out of the 72 total hits, 16 (22%) were CGC genes, but the enrichment of CGC genes was not significant (***Figure 20***). The CGC hits included *PPM1D*, *SMARCE1*, *LRP1B*, and *FGFR1*, among others.

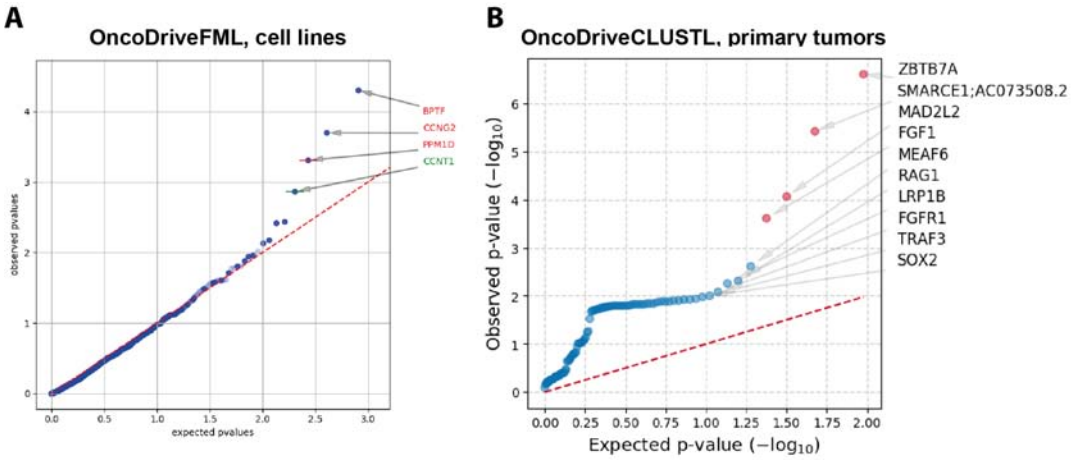


***Figure 32. Representative quantile-quantile plots of observed vs. expected p values from driver analyses in untranslated regions of protein-coding genes**: **A**. OncoDriveFML in our cell lines. **B**. OncoDriveCLUSTL in our unpaired dataset of primary tumors. The red dashed line represents the theoretical uniform distribution. The top 10 results are highlighted. In red: q < 0.01.*

We further evaluated the 4 hits from OncoDriveFML, the top 5 hits from each OncoDriveCLUSTL analysis, and the CGC hits. Setting the thresholds RNAsnp p < 0.1, CADD score ≥ 15, and FATHMM-MKL score ≥ 0.9, three variants in our UTR hits had high predicted functional and structural impact (***Table 14***). One of them, chr12:48716738_G/A, affected the 5'-UTR of *CCNT1* in the A549 cell line (***Supplementary Figure 14***). Its VAF was 100%. The affected nucleotide is part of the 5'-UTR of only one out of three isoforms, whose 5'-UTR is longer than the rest. Nevertheless, the variant may affect the promoter of the other isoforms. The second variant, chr3:181712310_G/C, affected the 5'-UTR of *SOX2* in two primary tumors (VAFs = 38% and 45%)

(***Supplementary Figure 15***). Finally, chr3:181714021_G>T affected the 3'-UTR of *SOX2* in a primary tumor and its VAF was 48% (***Supplementary Figure 15***).

Table 14. High-impact variants in candidate driver untranslated regions (UTRs).

Dataset	Variant	Gene (UTR)	MKL	CADD	Struc
Cell lines	chr12:48716738_G/A	*CCNT1* (5'-UTR)	0.97	21.3	0.02
Tumors, unpaired	chr3:181712310_G/C	*SOX2* (5'-UTR)	0.91	18.1	0.06
Tumors, unpaired	chr3:181714021_G/T	*SOX2* (3'-UTR)	0.94	16.4	0.05

Genomic coordinates use the hg38 reference genome. The following functional impact scores are reported: "MKL": FATHMM-MKL (range = 0-1, higher score means higher impact); CADD (Phred scale, higher score means higher impact); "Struc": RNAsnp p value (lower score means higher impact).

Next, we determined if the SNVs in the UTRs were associated with changes in gene expression in *cis*. We were only able to perform this analysis in cell lines, which had available transcriptome profiling data from the CCLE, as no hits were found in TCGA-LUAD. Only 3'-UTR variants in *SP1* were associated with lower mRNA expression, but the significance was lost after p value correction to control the false discovery rate (unadjusted p = 0.044, q = 0.44) (***Figure 33***). One 3'-UTR variant in *SP1* was recurrent in H322 and in HCC4006 (chr12:53416182_G/A) and it was predicted to have a moderate-high functional impact (CADD score = 17.9; FATHMM-MKL score = 0.92) but low structural impact (RNAsnp p = 0.29).

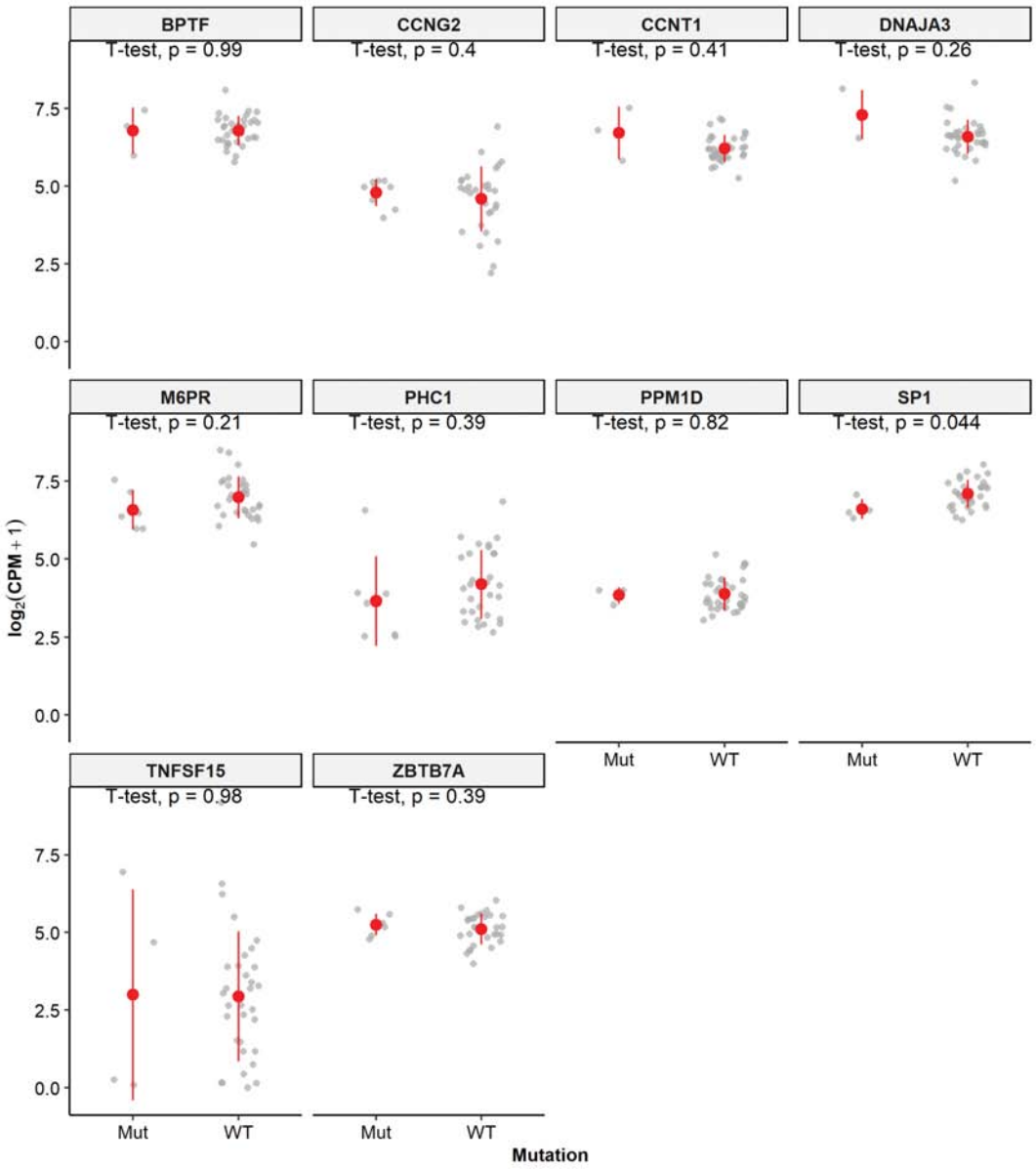


Figure 33. Association between variants in UTR hits in LUAD cell lines and mRNA expression according to CCLE data*. The 36 LUAD cell lines from our study that had expression data in CCLE are plotted. Mut: mutant; WT: wild type; CPM: counts per million. The p values shown here are not corrected for multiple hypothesis testing.*

MicroRNAs typically downregulate gene expression by binding to complementary sequences in 3'-UTRs (Bartel, 2004, 2009). Therefore, we predicted whether any of the high-impact SNVs in 3'-UTRs created or disrupted miRNA binding motifs. We tested the 3'-UTR variant in *SOX2*, which was identified in the functional impact analysis, and 3'-UTR variants in

SP1, which was identified in the expression analysis. For *SP1*, we also included a recurrent variant identified in two of our primary tumor samples (chr12:53414765_C/T). Although we were unable to evaluate whether the variant affected *SP1* expression, it was predicted to have a high functional impact (CADD score = 19.8; FATHMM-MKL score = 0.98). Using miRcode, we found that none of the tested variants affected any miRNA binding sites (Jeggari et al., 2012). To confirm our negative findings, we used the "Custom Prediction" tool from miRDB (http://mirdb.org/mirdb/custom.html). According to miRDB, the largest change in predicted miRNA binding was induced by the chr12:53414765_C/T variant in *SP1*, which changed the binding score of miR-1277-5p from 70 to 80. However, despite being the largest predicted change, it was still a relatively small difference, and median miR-1277-5p was <1 TPM in external LUAD miRNA sequencing (miRNA-Seq) datasets from TCGA and from Gillette et al (Gillette et al., 2020). As a result, the variant may not have a meaningful impact on *SP1* levels. Overall, we failed to identify any meaningful changes in predicted miRNA binding caused by the high-impact 3'-UTR variants.

In summary, different criteria to prioritize the hits in UTRs yielded different top candidates, but none of them had clear biological impact according to available data. However, we were unable to evaluate expression changes in our primary tumors due to lack of data. Finally, other criteria not explored here, such as disruption or creation of protein binding motifs, may help explain the effect of the high-impact UTR variants.

3.3.7. Annotation of miRNA motifs

To determine whether miRNA variants disrupted or created DROSHA processing motifs, we implemented two novel methods for annotating DROSHA motifs using two different criteria: distance from DROSHA cleavage sites ("positional" method) and distance from structural features ("structural" method) (***Section 3.2.11.3***). Before annotating the variants, we tested whether our methods successfully predicted pri-miRNA stems and DROSHA processing motifs.

3.3.7.1. Performance of the stem prediction workflow

The structural method for predicting DROSHA processing motifs, as well as the prediction of mGHG motifs, required the structures of pri-miRNA stem-loops. Using our method, we successfully predicted 842/1881 stems from miRBase 21 miRNAs. The remaining stems could not be predicted for two main reasons:

- **Lack of a complete annotation of the 5p and 3p miRNAs (932 failures)**. When either the 5p or the 3p mature miRNAs are missing from the annotation, they can be predicted by assuming that DROSHA produces a cut staggered by 2 nt (Kim et al., 2021; Urbanek-Trzeciak et al., 2020). However, we chose not to follow this approach because we found numerous exceptions to this rule in miRBase, and because most (275/295, 93%) human high-confidence miRNAs (as defined by miRBase) had complete annotations in either miRBase or MirGeneDB.
- **RNAfold did not predict a stem-loop structure that contained one mature miRNA in each arm (107 failures)**. Wrong predictions could arise because: (i) the input was incorrect; or (ii) the algorithm failed. Currently, it is not possible to retrieve complete pri-miRNA sequences from gene annotations. Instead, miRNA genes must be padded by an arbitrary number of nucleotides (in our case, 30 nt) in hopes of spanning the full stem-loop of the pri-miRNA (Kim et al., 2021). We tried padding miRNA genes by different lengths, but results were similar or worse than those obtained by our final approach. We also tried using mfold for secondary structure predictions (Zuker, 2003), but results were similar to those of RNAfold. Therefore, we deemed our RNAfold approach as acceptable within current limitations.

After optimizing secondary structure predictions, we used them to predict pri-miRNA stem-loops. To test our stem-loop predictions, first we examined the distribution of stem lengths. When calculated as the shortest length among the two arms, optimal stem lengths are within the range of 33-39 nt (Roden et al., 2017). Indeed, 468/842 (56%) of our predicted stems were within the optimal

range (***Figure 34A***). Our predictions were biased towards short stems, as 216/842 (26%) were shorter than 33 nt, whereas 158/842 (19%) were longer than 39 nt. To determine the source of this bias, we examined the secondary structure predictions and our stem annotations in a selection of stems below, within, and above the optimal length range. While most stem predictions outside the optimal range looked correct, some were shorter than expected because RNAfold had predicted a multi-branched junction within the stem. These secondary structures were likely to be incorrect, and they could especially impact the predictions of basal junctions. This was likely to cause structural motif predictions to be less reliable than positional predictions.

An additional test for our stem prediction method was the distribution of distances between the 5p DROSHA cleavage sites and the predicted basal junctions (***Figure 34B***). The optimal distance in our predictions was 13 nt, which agrees with previous reports (Auyeung et al., 2013; Kim et al., 2021). However, a significant number of basal junctions were 1-2 nt away from the optimum. Such deviations from the optimal distance have also been described (Kwon et al., 2019), and they can be attributed to either pri-miRNAs that harbor a strong mGHG motif at the -7 position, stems with an unusual number of mismatches and bulges, or incorrect predictions. Overall, we concluded that our stem predictions were mostly correct.

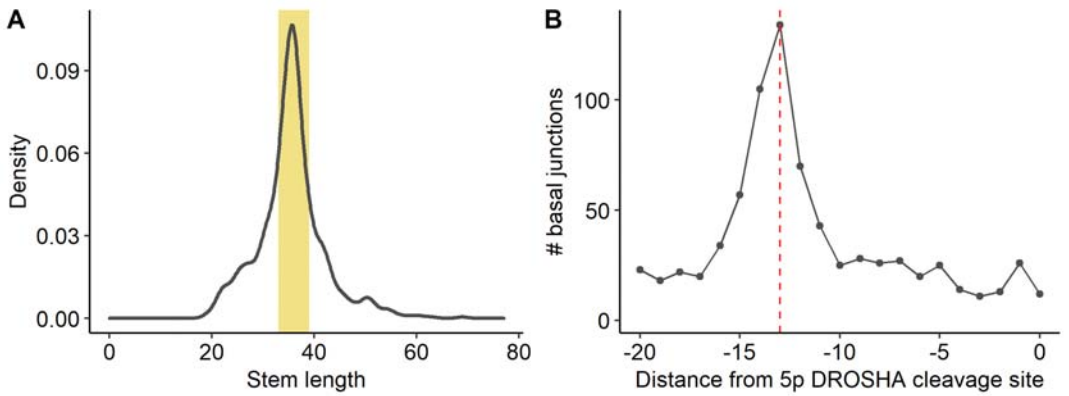


Figure 34. Quality tests of our stem prediction method. A. *Length distribution of predicted stems. Lengths are expressed in nucleotides (nt). In yellow, optimal range of 33-39 nt proposed by (Roden et al., 2017).* ***B****. Distances between basal junctions and 5p DROSHA cleavage sites. The optimal distance of -13 nt is highlighted with a red dashed line (Auyeung et al., 2013).*

3.3.7.2. Optimal ranges for DROSHA motifs

To confirm whether the positions of DROSHA motifs reported in the bibliography (***Table 2***) matched our data, we studied the occurrence of each sequence motif as a function of its distance to DROSHA cleavage sites (for positional predictions) or to structural features (for structural predictions). As expected, we found a peak of UG sequences at position -14 from 5p DROSHA cleavage sites (***Figure 35A***) and at basal junctions (***Figure 35B***) (Auyeung et al., 2013). We also detected a peak of UGU/GUG sequences within the range of 21-25 nt from 5p DROSHA cleavage sites (***Figure 35C***) and at apical junctions (***Figure 35D***), as previously described (Auyeung et al., 2013). Furthermore, we observed a high proportion of CNNC sequences at 16-18 nt from 3p DROSHA cleavage sites (***Figure 35E***) and at 5-6 nt from basal junctions in the 3p arm (***Figure 35F***), as previously reported (Auyeung et al., 2013; Roden et al., 2017). Finally, we found a strong increase in mGHG scores at position -7 from 5p DROSHA cleavage sites (***Figure 36***), in agreement with previous reports (Kwon et al., 2019).

Whereas positional definitions of motifs have been thoroughly studied, structural definitions are mostly based on indirect evidence. For example, because basal UG motifs are enriched at position -14 from 5p DROSHA cleavage sites and basal junctions are enriched at position -13, UG motifs should be at basal junctions (Auyeung et al., 2013). However, direct evidence linking DROSHA cleavage motifs and structural features is scarce (Kwon et al., 2019; Roden et al., 2017). In this context, we aimed to optimize the structural definitions of motifs based on our own observations (***Table 8***). Basal UG motifs were indeed enriched at basal junctions, although some may be 1-2 nt upstream (***Figure 35B***). On the other hand, apical UGUG motifs were frequent within the range of -2...+1 from the apical junction, which suggested that they may not have to be located strictly at apical junctions (***Figure 35D***). Finally, for downstream CNNC motifs, Roden et al defined a "strict" range of 5-6 nt from the 3p basal junction and a "permissive" range of 3-11 nt (Roden et al., 2017). Based on our observations, we used a "middle ground" range of 5-9 nt (***Figure 35F***).

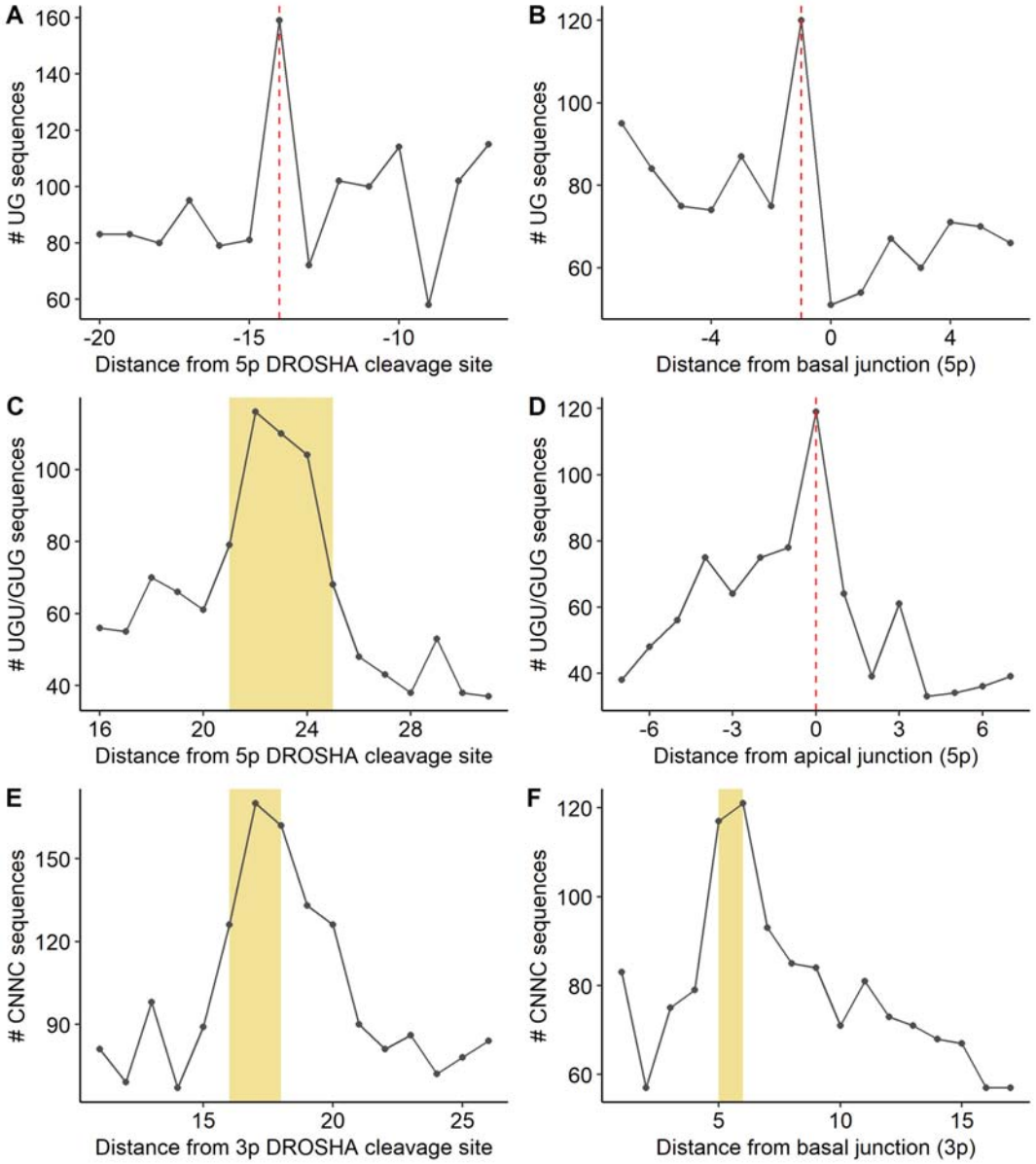


Figure 35. Occurrence of UG, UGUG, or CNNC motifs as a function of the distance from structural features or DROSHA cleavage sites. *The analysis was performed in the 842 pri-miRNAs whose stem-loop structures could be predicted.* ***A****. Occurrence of UG sequences as a function of the distance from 5p DROSHA cleavage sites. The red dashed line marks the known optimal position of the basal UG motif (Auyeung et al., 2013).* ***B****. Occurrence of UG sequences as a function of the distance from the basal junction. The red dashed line marks the known optimal position of the basal UG motif (Auyeung et al., 2013).* ***C****. Occurrence of UGU or GUG sequences as a function of the distance from the 5p DROSHA cleavage site. The yellow rectangle marks the optimal range of the apical UGUG motif (Auyeung et al., 2013).* ***D****. Occurrence of UGU or GUG*

sequences as a function of the distance from the apical junction. The red dashed line marks the apical junction. ***E****. Occurrence of CNNC sequences as a function of the distance from the 3p DROSHA cleavage site. The yellow rectangle marks the optimal range of the downstream CNNC motif (Auyeung et al., 2013).* ***F****. Occurrence of CNNC sequences as a function of the distance from the basal junction. The yellow rectangle marks the optimal range of the downstream CNNC motif (Roden et al., 2017).*

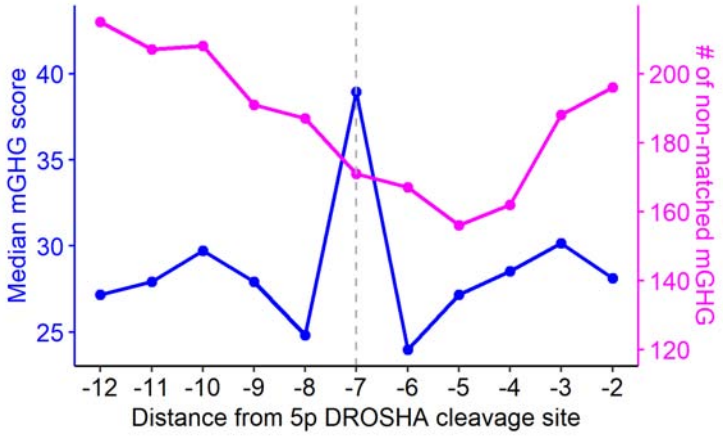


Figure 36. Median mGHG score across human pri-miRNAs as a function of the distance (in nucleotides) to the 5p DROSHA cleavage site*. The mGHG scores were normalized between 0 and 100 (Kwon et al., 2019). In pink: number of pri-miRNAs to which no mGHG score could be assigned, which could happen because: 1) none of the nucleotides at the GHG position in the 5p arm were paired with nucleotides in the 3p arm; or 2) the two trinucleotides at the mGHG position were not facing each other. n = 842 successfully predicted pri-miRNA hairpins.*

3.3.7.3. DROSHA motif predictions

After defining the optimal ranges for DROSHA motifs according to both the positional and the structural methods, we used both methods to predict DROSHA motifs in all pri-miRNAs and then we compared the predictions. The highest agreement between the structural and positional methods was reached for apical UGUG motifs (Jaccard similarity index = 0.73; ***Figure 37***), followed by downstream CNNC motifs (Jaccard similarity index = 0.53). The lowest agreement was reached for basal UG motifs (Jaccard similarity index =

0.31), which could be explained by the stricter definition of basal UG motifs compared to the other motifs. We thoroughly evaluated the discrepancies between both methods by examining the predicted positions of the motifs, the annotations of mature miRNAs, and the RNAfold predictions. Generally, the structural method was less reliable than the positional method because it depended on secondary structure and stem predictions that sometimes failed, as detailed in previous sections. However, despite its limitations, the structural method complemented the positional method in cases in which plausible motifs were not exactly located at their positional optima (Kwon et al., 2019). For these reasons, we used both methods for annotating variants.

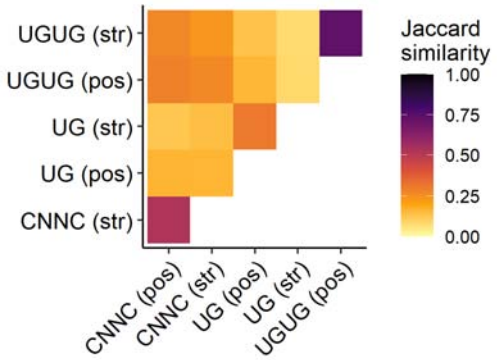


Figure 37. Similarity between DROSHA motif prediction methods*. The Jaccard similarity index was used to measure the agreement between the sets of miRNAs predicted to have each motif according to structural (str) and positional (pos) methods.*

We also determined the co-occurrence of different combinations of motifs across all analyzed pri-miRNAs (***Figure 38A-B***). The most frequent motifs were downstream CNNC motifs, followed by mGHG, apical UGUG, and basal UG. Roughly ~21% of the pri-miRNAs lacked all DROSHA motifs. Most frequently, motifs occurred either alone (~35-39%) or in combination with only one other motif (~30-31%). Co-occurrence of all four motifs was rare, affecting only ~1-2% of all pri-miRNAs.

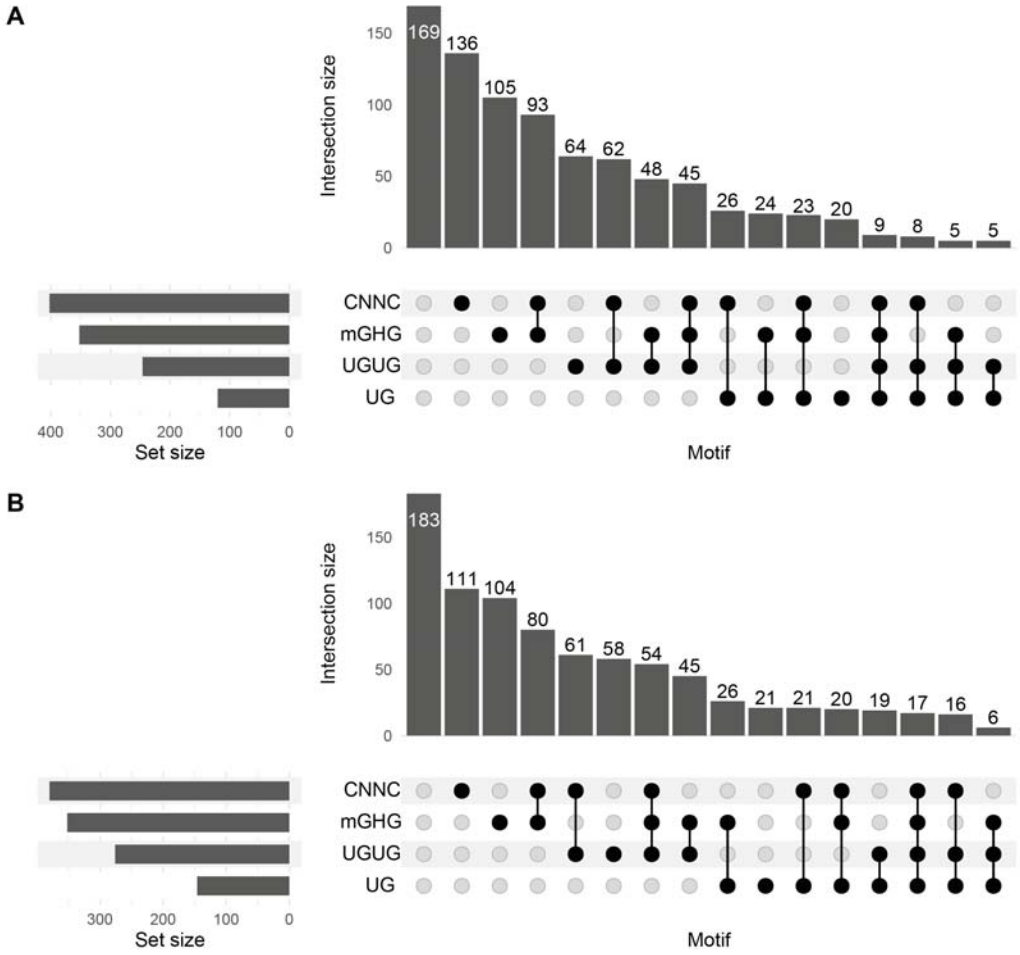


Figure 38. Co-occurrence of DROSHA processing motifs in human pri-miRNAs*. The analysis includes the 842 pri-miRNAs for which stems could be predicted successfully. Predictions of motifs were made by two methods:* ***A****. Based on distance to DROSHA cleavage sites ("positional");* ***B****. Based on distance to structural features ("structural").*

3.3.8. Annotation of miRNA variants

Using our novel framework, we annotated all variants that affected miRNA genes or their 200 bp flanking regions in all LUAD datasets. Here, we also included WES variants of 582 patients from TCGA-LUAD. Although WES was not useful for our lncRNA analysis because its exome capture design (Agilent SureSelect Human All Exon v4) did not cover most lncRNAs, it did cover 1676/1881 miRNA genes from miRBase 21.

3.3.8.1. Usefulness of WES, WGS, and targeted sequencing: differences in coverage of miRNA genes

Although the WES dataset of TCGA-LUAD included more samples than its WGS dataset and our targeted sequencing, it covered most miRNA genes at very low depths. Indeed, median sequencing depth in miRNA genes ±200 flanking bp was only ~2X in TCGA-LUAD WES data, in contrast to ~30X in our targeted sequencing and ~42X in TCGA-LUAD WGS. The reason for the markedly low coverage of WES was not a lack of baits, as WES baits covered ~90% of miRNA genes. Still, up to ~45% of the nucleotides of miRNA genes had zero coverage in WES samples (***Figure 39A***). This percentage was as low as ~4% in some samples, but even then the coverage of most nucleotides in miRNA genes did not reach 10X. The lower depth of TCGA-LUAD WES compared to the other datasets was reflected in a lower number of detected variants per sample (***Figure 39B***). On the other hand, the number of variants in miRNA genes and flanking regions was the highest in TCGA-LUAD WES data because the cohort was the largest (***Figure 39C***). In conclusion, WES did not cover miRNA genes at a high enough depth for a comprehensive mutational analysis per sample, but the large size of the TCGA-LUAD cohort compensated for its shallow sequencing depth.

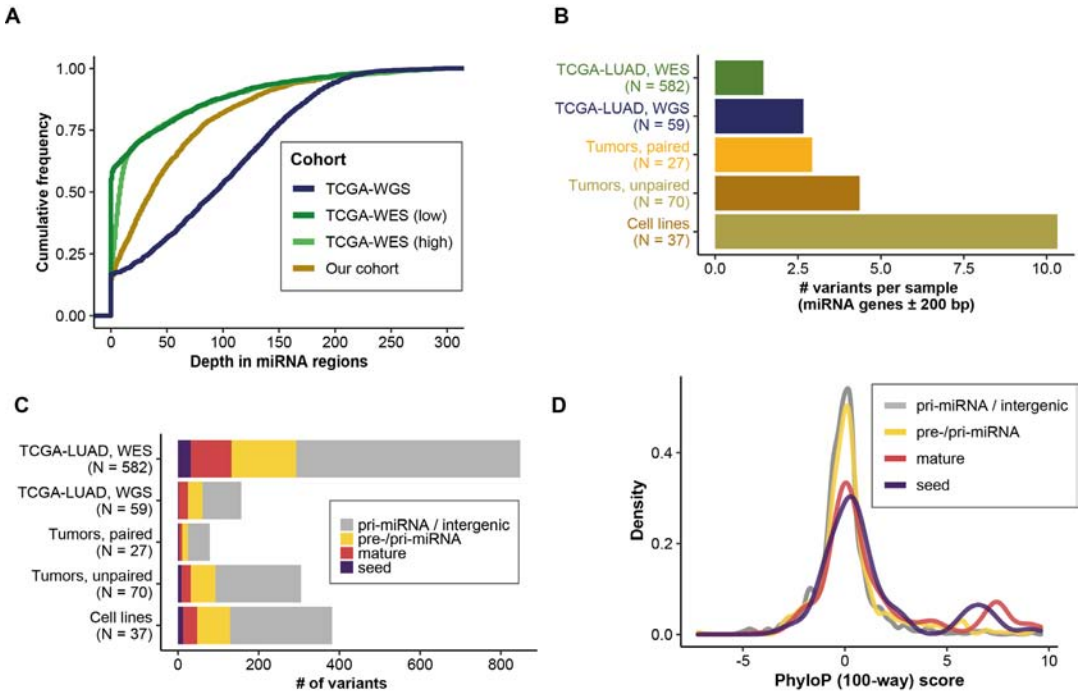


Figure 39. General statistics on miRNA variants. ***A****. Cumulative frequency distribution of the coverage of miRNA genes ±200 bp for four selected samples from three analyzed cohorts: TCGA-WGS (whole genome sequencing data from TCGA-LUAD); TCGA-WES (whole exome sequencing data from TCGA-LUAD); and our targeted sequencing cohort of primary tumors. For TCGA-WES, two samples were selected: one in which ~45% of the miRNAs had zero coverage ("low"), and one in which ~4% of the miRNAs had zero coverage ("high").* ***B****. Mean number of variants per sample in miRNA genes ±200 bp in the analyzed cohorts.* ***C****. Classification of the variants in miRNA regions based on the type of affected sequence.* ***D****. Distribution of phyloP 100-way scores for the variants that affected each type of sequence in miRNA regions. Here, all cohorts were grouped together.*

3.3.8.2. Variants in mature miRNAs are rare

Variants in miRNA genes were rare, as expected from their short length and high conservation (***Figure 39B-C***). Excluding flanking regions, we only found ~1 variant in miRNA genes per sample in our cohort of primary tumors and ~3.5 variants per sample in our cell lines. In mature miRNAs, we only found one variant for every ~2.5 primary tumors; in seeds, one variant for every ~9 primary tumors. Only 11 variants were recurrent, and none of them affected

mature miRNAs. Overall, our results support a low mutation rate of miRNA genes in LUAD.

To assess the functional relevance of the detected variants in miRNAs, we first evaluated the conservation (phyloP 100-way scores) of the nucleotides affected by the variants (***Figure 39D***). Most variants had phyloP scores centered around zero, even those in mature miRNAs and in seeds, suggesting that the affected sequences were under neutral evolution. However, there was an increase in the density of phyloP scores approximately at score ≥4, especially in mature miRNAs and seeds. The variants with phyloP score ≥4 included ~2% of the variants in flanking regions, ~6.4% of the variants in pre-miRNAs, ~17% of the variants in mature miRNAs, and ~15% of the variants in seeds. Overall, the conservation of mutated nucleotides in miRNAs was a mixture of two distributions: a major distribution of nucleotides under neutral evolution, and a minor distribution of highly conserved nucleotides.

Next, we aimed to select variants with high putative functional impact for further studies. In a first approach, we analyzed the hits from OncoDriveFML and OncoDriveCLUSTL (***Section 3.3.6.4***). Almost all of their variants affected poorly conserved positions outside of the mature miRNAs, and therefore their significance was unclear. The only exception was miR-4712-5p: one variant affected its seed in the H1373 cell line and another one affected the mature miRNA in the SKLU-1 cell line. However, both positions were poorly conserved and mir-4712 is not considered a high-confidence miRNA by neither miRBase nor MirGeneDB, casting doubts on its functional relevance (Fromm et al., 2015; Kozomara and Griffiths-Jones, 2014).

Due to our limited success among the hits from the driver analysis, we extended our study to all variants in miRNA seeds. We focused on seed variants from our cohorts that either affected a highly conserved nucleotide (phyloP ≥ 4) or that were detected in TCGA-LUAD or in COSMIC. Only five variants met these criteria: three in primary tumors and two in cell lines (***Supplementary Table 2***). Out of these, a somatic variant in the seed of miR-133b affected the most conserved nucleotide (phyloP score = 7.2) and it was predicted to be highly deleterious by FATHMM-MKL (score = 0.99) and

by CADD (score = 19.42). Therefore, we selected the miR-133b variant for further studies.

3.3.8.3. Seed variant in miR-133b: a case study

We selected a variant in the seed of miR-133b (chr6:52148992_G>T), detected in one of our primary tumor samples from the paired analysis, for further in-depth characterization. The VAF of the variant was 53%, which implied that it was present in a high percentage of tumor DNA molecules in the sample.

The *MIR133B* locus is located in chromosome 6, overlapping the lncRNA gene *LINCMD1* in its opposite strand (***Figure 40***). Mature miR-133b originates from the 3p arm of the mir-133b pre-miRNA. On the other hand, the 5p arm does not generate a detectable mature miRNA according to high-throughput sequencing data (Fromm et al., 2015; Kozomara and Griffiths-Jones, 2014). Therefore, miR-133b is a synonym for hsa-miR-133b-3p.

Recurrence in external datasets of somatic variants

Next, we evaluated whether *MIR133B* is mutated in external cancer cohorts and in non-cancer human populations. For cancer, we searched for pan-cancer somatic variants from COSMIC and from the International Cancer Genome Consortium (ICGC). ICGC reported variants in the seed of miR-133b in a chronic lymphocytic leukemia sample (ICGC sample ID: DO7172) and in a colon adenocarcinoma sample (DO8730) (***Figure 40***). Furthermore, ICGC found a variant in mature miR-133b in a glioblastoma sample (DO13270) and 15 more variants in the rest of the *MIR133B* gene, one of them in a LUAD sample (DO24692). COSMIC did not report any new somatic variants not described by ICGC. Regarding general populations, remarkably, we found no germline polymorphisms in miR-133b in non-cancer samples from gnomAD v3.0 (N = 71 702 WGS samples), dbSNP, or the 1000 Genomes Project (***Figure 40***). In conclusion, miR-133b is somatically mutated across multiple cancers, and germline polymorphisms in its seed are rare in healthy individuals.

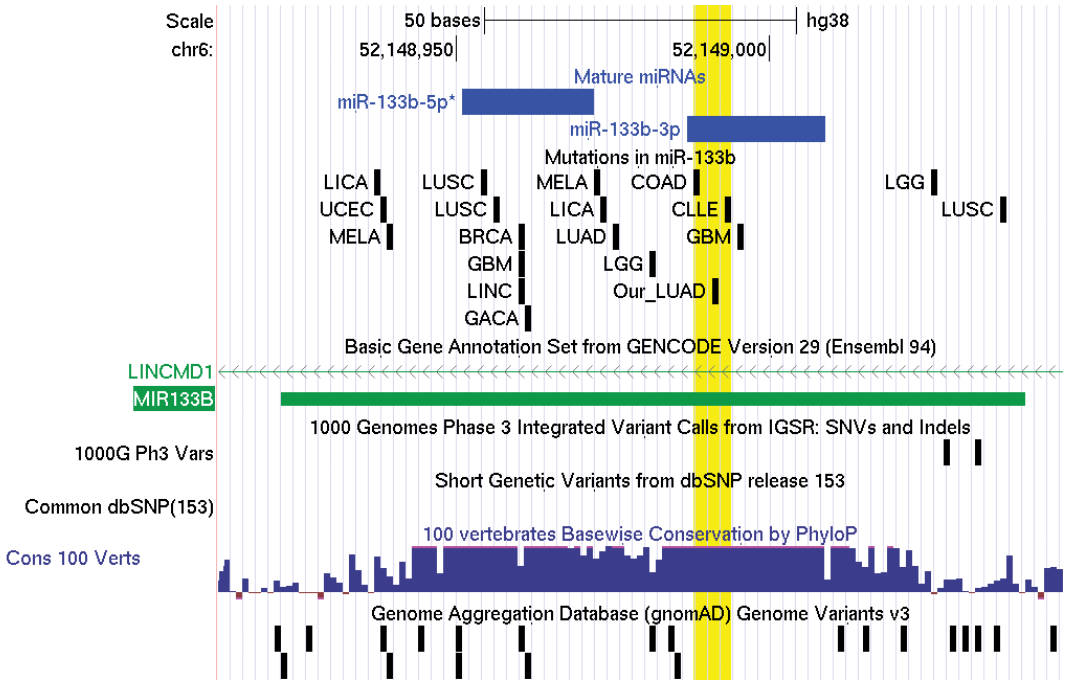


Figure 40. The MIR133B locus and its pan-cancer variants. *The 5p mature miRNA is not abundant according to high-throughput experiments, and hence it is marked with an asterisk. The highlighted yellow region marks the seed of mature miR-133b-3p. The variant found in our lung adenocarcinoma (LUAD) primary tumor is labeled as "Our_LUAD". The rest of the variants were reported by the International Cancer Genome Consortium (IGCG) across multiple cancers: LICA/LINC: liver cancer; UCEC: uterine corpus endometrial carcinoma; MELA: melanoma; LUSC: lung squamous cell carcinoma; BRCA: breast cancer; GBM: glioblastoma; GACA: gastric cancer; LGG: low grade glioma; COAD: colon adenocarcinoma; CLLE: chronic lymphocytic leukemia. Screenshot from the UCSC Genome Browser.*

Expression of miR-133b in LUAD and in normal lung

Even if the seed of a miRNA is mutated, the variant may not be functionally relevant if the miRNA is not expressed in the affected sample at biologically meaningful levels. Therefore, we determined whether miR-133b is expressed in normal lung and in LUAD samples. We used two independent miRNA-Seq datasets: one from TCGA-LUAD (N = 513 tumors and 46 normal samples) and one from Gillette et al (N = 100 normal samples and 107 LUAD tumors) (Gillette et al., 2020). All normal samples were from solid lung tissue adjacent to the tumors.

In both datasets, miR-133b was expressed in normal lung (***Figure 41A-B***). In particular, miR-133b expression was in the 88th percentile in normal samples from TCGA-LUAD and in the 84th percentile in normal samples from Gillette et al. In addition, miR-133b was downregulated in LUAD compared to normal lung both in TCGA-LUAD (fold change (FC) = -2.75, two-sample t test $p = 9.5 \cdot 10^{-19}$) and in Gillette et al (FC = -6.66, two-sample t test $p = 5.6 \cdot 10^{-31}$) (***Figure 41A-B***). In TCGA-LUAD, 24% of the tumor samples but none of the normal samples had undetectable expression of miR-133b. Furthermore, miR-133b was moderately expressed in LUAD cell lines from the CCLE (***Figure 41C***). The downregulation of miR-133b in LUAD suggested that it may be a tumor suppressor miRNA. Nevertheless, absolute expression of miR-133b in most samples across all cohorts was within the range of ~1-15 RPM, which may not be sufficient for a meaningful biological activity (Kilikevicius et al., 2022; Mullokandov et al., 2012).

Finally, we tested if miR-133b expression was associated with overall survival of patients from the TCGA-LUAD cohort. Although there was a trend for lower overall survival of "miR-133b low" patients compared to "miR-133b high" patients, it was not statistically significant (logrank p = 0.12) (***Figure 41D***). Furthermore, the trend was completely lost in a multivariate analysis that incorporated patient age and tumor stage as clinical covariates (Cox proportional hazards model: hazard ratio = 0.95 [95% CI: 0.78-1.18]; p = 0.65). Taken together, these results suggest that miR-133b is expressed in normal lung and that it is downregulated in LUAD, but there is no decisive evidence that its downregulation is associated with patient survival.

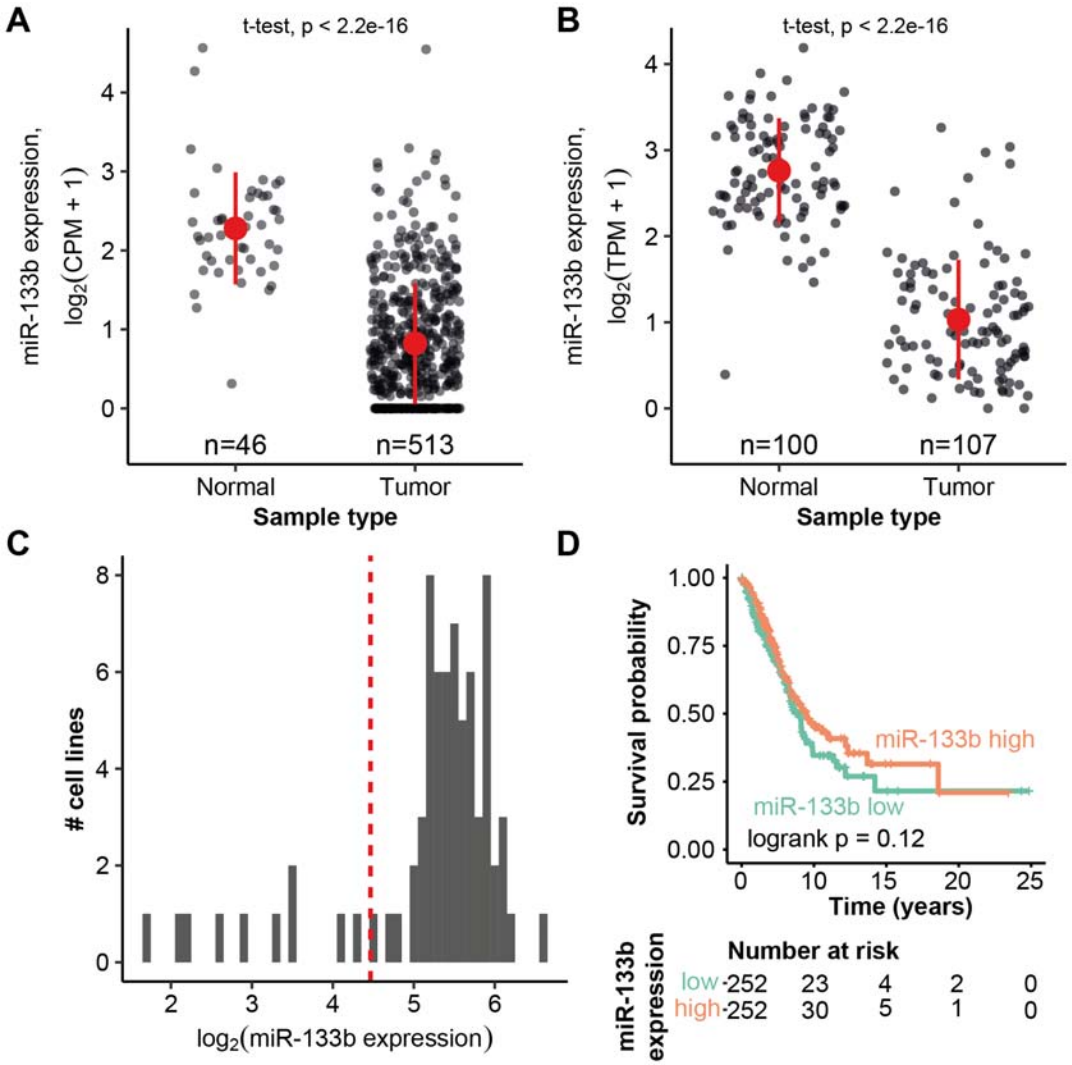


Figure 41. Expression analyses of miR-133b. A. *Differential expression of miR-133b between tumor and normal samples in TCGA-LUAD;* ***B.*** *Differential expression of miR-133b between tumor and normal samples in data from Gillette et al (2020).* ***C.*** *Histogram of miR-133b expression in LUAD cell lines from the Cancer Cell Line Encyclopedia. The red line displays the median expression of all miRNAs in all cell lines.* ***D.*** *Univariate survival analysis in TCGA-LUAD. Patients were split in two groups based on whether miR-133b expression was above or below the median. CPM: counts per million. TPM: transcripts per million.*

Predicted targets of miR-133b

Next, we predicted how the variant in the seed of miR-133b may affect the targets of the miRNA. The variant changed the seed from UUG**G**UC to UUG**U**UC. Because seeds are the main determinants of miRNA targets, the variant was expected to radically change the targets of miR-133b. Indeed, two independent target prediction tools (TargetScan and miRDB) estimated that the overlap between the targets of wild type and mutant miR-133b was <5% (***Figure 42A-B***). However, the agreement between both tools was modest: 30% for wild type and 14% for mutant miR-133b (***Figure 42C-D***). To prioritize the predicted targets, we focused on CGC genes in the intersection between the two prediction methods. We hypothesized that, if miR-133b has a tumor suppressor role, wild type miRNA should target oncogenes or, alternatively, mutant miR-133b should target tumor suppressor genes. Indeed, oncogenes such as *EGFR* were predicted targets of wild type, but not mutant, miR-133b (***Table 15***). Furthermore, tumor suppressor genes such as *DICER1* were predicted targets of mutant, but not wild type, miR-133b.

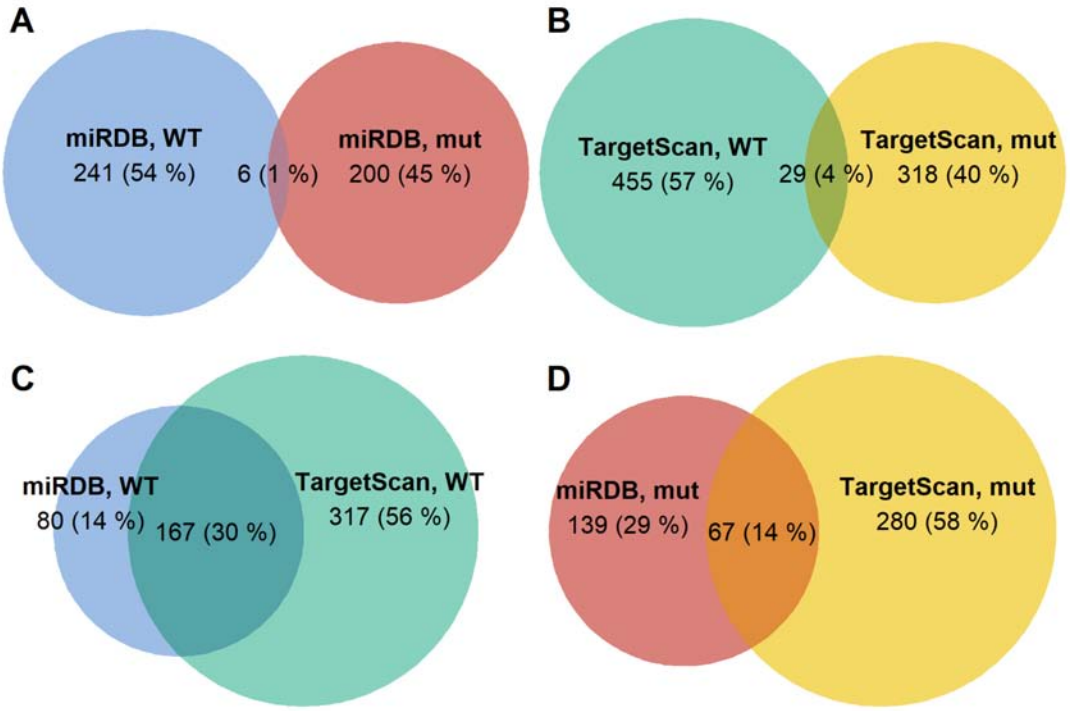


Figure 42. Agreement between target predictions for wild type (WT) and mutant (mut) miR-133b. ***A****. miRDB predictions, WT vs. mutant.* ***B****. TargetScan predictions, WT vs. mut.* ***C****. miRDB vs. TargetScan predictions for wild type miR-133b.* ***D****. miRDB vs. TargetScan predictions for mutant miR-133b.*

Table 15. Target predictions for mutant and wild type miR-133b.

Targeted by...	**TSGs**		**OGs**	**Others**	
WT but not mut	*MYH9*	*PTPRD*	*EGFR*	*EPHA7*	*LHFPL6*
	PML	*PTPRK*	*FGFR1*	*FOXL2*	*MLLT3*
	PRDM1	*SMARCD1*	*IDH1*	*JAZF1*	*MSN*
			SGK1	*LASP1*	*THRAP3*
			XPO1		
Mut but not WT	*ARHGEF12*		*JAK2*	-	
	DICER1				
	TSC1				

Only genes from the Cancer Gene Census predicted by both TargetScan and miRDB are included. Shaded cells highlight the genes that are consistent with a tumor suppressor role of wild type (WT) miR-133b or an oncogenic role of mutant (mut) miR-133b. TSG: tumor suppressor gene. OG: oncogene.

To obtain further support for the target predictions, we tested the correlation between expression of wild type miR-133b and its putative targets at mRNA and protein levels. We used the two independent cohorts of TCGA-LUAD (mRNA data only) and Gillette et al (mRNA and protein data). We focused on CGC targets predicted by both TargetScan and miRDB. Only *SMARCD1* mRNA expression and *EGFR* protein expression were significantly and negatively correlated with miR-133b expression (Kendall $\tau < -0.2$, $q < 0.05$), and only in Gillette et al's dataset (***Figure 43***). Although *PTPRK* was the only experimentally confirmed predicted target according to the TarBase database, it was not supported by the correlation analysis. Unexpectedly, *DICER1* protein expression was negatively correlated with miR-133b expression in Gillette et al's dataset (data not shown), even though it was predicted as a target of the mutant, but not of the wild type, miRNA. Among non-CGC targets, *EIF4A1* was the only one that was both experimentally validated according to TarBase and supported by the correlation analysis, in particular at the protein level in Gillette et al's dataset (***Figure 43***).

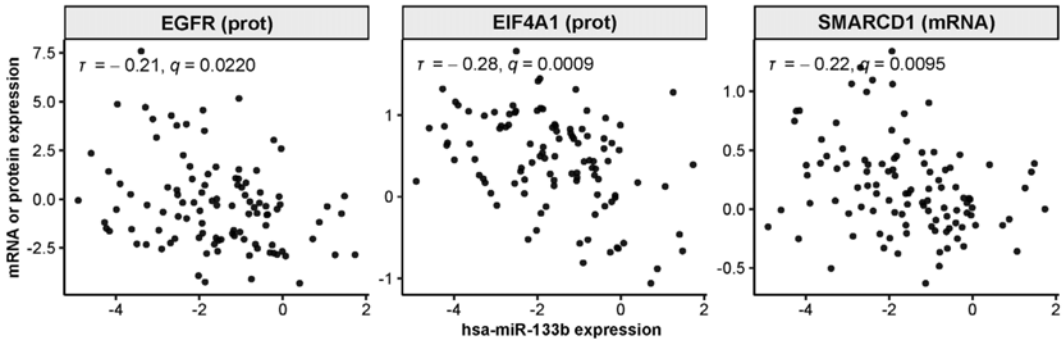


Figure 43. Correlation between miR-133b expression and mRNA or protein expression of three predicted targets*. Data were obtained from LUAD samples from Gillette et al (Gillette et al., 2020). Kendall correlation coefficients (τ) and q values (p values adjusted for false discovery rate) are shown.*

In conclusion, a somatic SNV in a highly conserved nucleotide of the seed of miR-133b was predicted to change its targets in a major way. Based on multiple independent methods, the highest confidence CGC targets of wild type miR-133b were *SMARCD1* and *EGFR*, and the highest confidence non-CGC target was *EIF4A1*. Importantly, mutant miR-133b was predicted to lose the ability to target the oncogene *EGFR*, which is consistent with a cancer-promoting role of the mutation. In future work, it should be experimentally determined whether miR-133b is expressed in our cohort at enough copies for it to have biological activity.

Do miR-133a-3p and miR-133b have redundant functions?

In the genome, miR-133b is located in a cluster (miR-206/133b) that shares its seed sequences with those of the miR-1/133a cluster, which has two copies in different chromosomes (Kozomara and Griffiths-Jones, 2014). In particular, miR-133b shares its seed with miR-133a-3p, and both mature miRNAs only differ in their last nucleotide, which has little impact on target specificity (Bartel, 2009). This raises the possibility that both miRNAs have redundant functions, and that miR-133a-3p may compensate a loss of of miR-133b.

In skeletal muscle, both miR-1/133a and miR-206/133b are highly expressed and they control cell growth and differentiation (Cesana et al., 2011; Chen et

al., 2006). However, miR-206/133b is not essential for this role in mice because miR-1/133a can compensate the loss of miR-206/133b (Boettger et al., 2014).

To determine whether miR-133a-3p and miR-133b functions are redundant in LUAD as well, we compared the expression of both miRNAs as well as their predicted targets. In both TCGA-LUAD and the dataset of Gillette et al, miR-133a-3p was more expressed than miR-133b in tumors and in normal samples (***Figure 44***). In particular, in LUAD samples, precursor mir-133a was expressed 4.18 times more than mir-133b in TCGA-LUAD (Student's t test, $p = 3.9 \cdot 10^{-18}$), whereas mature miR-133a-3p was expressed 10.8 times more than miR-133b in the dataset of Gillette et al (Student's t test, $p = 2.5 \cdot 10^{-22}$). Furthermore, expression of the top predicted targets of miR-133b, except for *EGFR*, was negatively correlated with miR-133a-3p expression as well (***Figure 45A-C***). More generally, the correlations between miR-133a-3p and proteome expression were highly similar to the correlations between miR-133b and proteome expression, suggesting that both miRNAs have highly similar activity in LUAD (Kendall $\tau = 0.76$, $p < 2 \cdot 10^{-16}$) (***Figure 45D***).

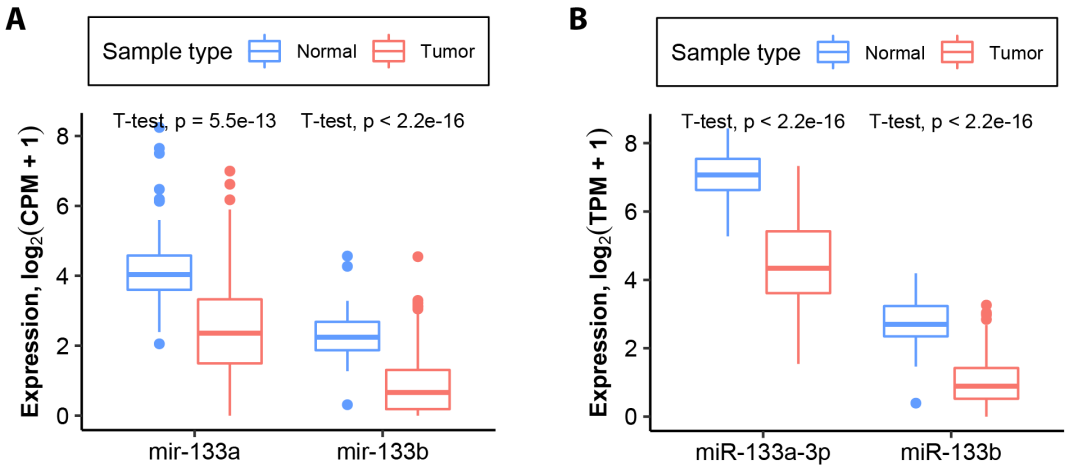


Figure 44. Expression of miR-133a-3p, miR-133b, and their precursors in LUAD primary tumors and normal samples. ***A****. In TCGA-LUAD. For each sample, expression values of both mir-133a precursors were added in the natural scale before taking the logarithm. CPM: counts per million.* ***B****. In Gillette et al (2020). TPM: transcripts per million.*

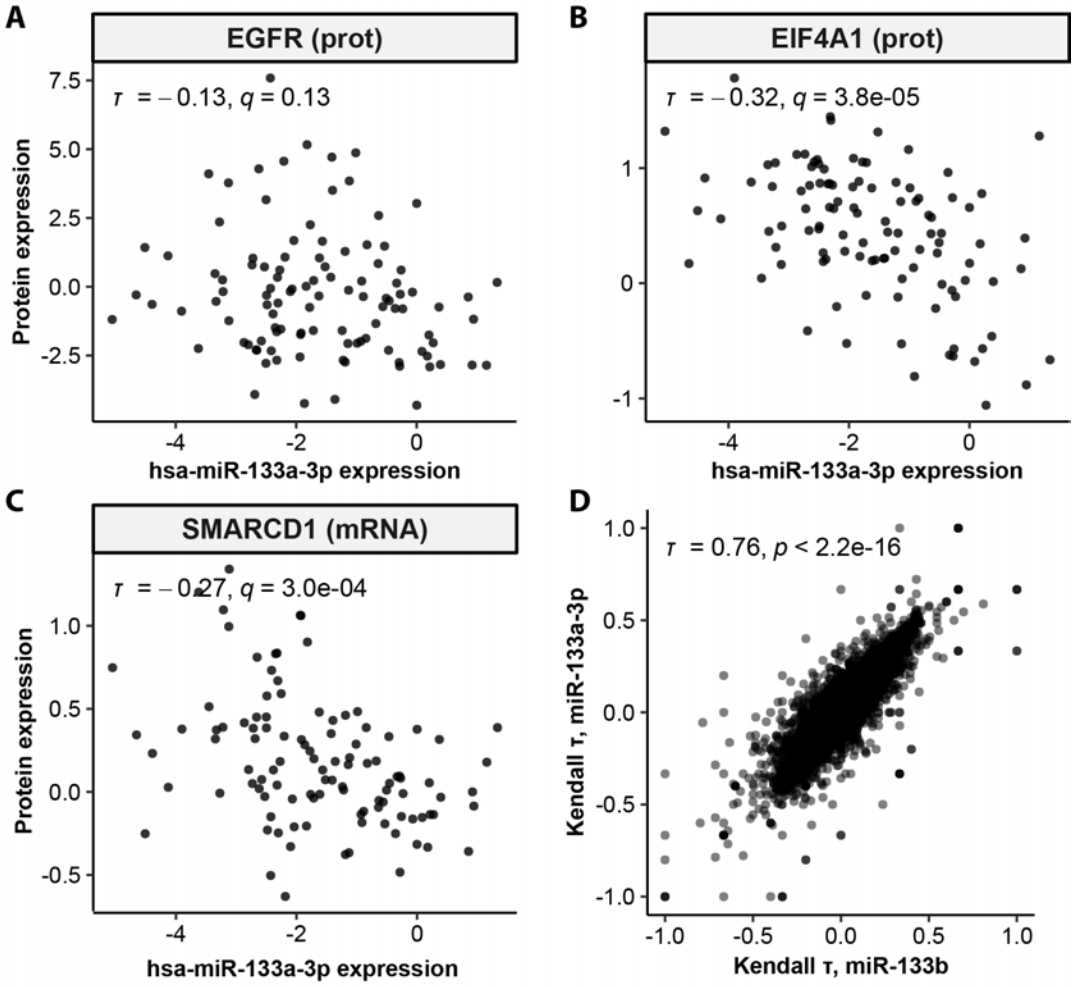


Figure 45. Correlation analyses between miR-133a-3p expression and protein or mRNA expression in the dataset of Gillette et al (2020). A-C. *Correlation analyses for the top miR-133b targets as in* ***Figure 43****. Median-centered miR-133a-3p expression was averaged across its two loci. Correlation coefficients (τ) and FDR-adjusted q values are shown for a Kendall correlation analysis.* ***D****. Correlation of the Kendall τ correlation coefficients: (i) between miR-133b expression and proteome expression, versus (ii) between miR-133a-3p expression and proteome expression.*

Taken together, our observations support that loss of function of miR-133b may be compensated by miR-133a-3p in LUAD. However, miR-133a-3p may not modulate *EGFR* as strongly as miR-133b. In addition, the miR-133b variant may generate a gain of function, for example by causing mutant miR-133b to target tumor suppressor mRNAs. In such a case, wild type miR-133a-3p would be unable to compensate for the new functions of mutant miR-133b.

3.3.8.4. Variants may cause major changes in DROSHA processing motifs

Besides variants within miRNA seeds, variants in DROSHA processing motifs may also have major functional consequences. Therefore, we determined which variants created or disrupted DROSHA processing motifs in our targeted sequencing samples (***Supplementary Table 3***) and in TCGA-LUAD WES and WGS data (***Supplementary Table 4***). In our samples, 9 variants affected DROSHA processing motifs in cell lines and 4 in primary tumors. The most recurrently affected motif was the mGHG motif (7 variants). Most importantly, in a primary tumor, a somatic variant in the mGHG motif of mir-139 was predicted to decrease DROSHA processing efficiency by ~12x (mGHG score from 86 to 6.9) (***Figure 46A***). In cell lines, mGHG variants in were found in mir-7-1 (~18x reduction of DROSHA processing efficiency) and in mir-7-2 (~28x reduction). Conservation of the positions affected by the three variants mentioned above was high (phyloP score ≥ 3.8). Variants also disrupted other types of DROSHA processing motifs, but their conservation was low, casting doubts on their functional relevance.

In TCGA-LUAD, 25 variants were found in DROSHA processing motifs (***Supplementary Table 4***). Again, they most frequently affected mGHG motifs (14 variants). The mGHG variant that had the highest predicted functional impact affected mir-551b. It was predicted to decrease DROSHA processing efficiency by 5.9x and the affected nucleotide was highly conserved (phyloP score = 5.9). Regarding other DROSHA processing motifs, most remarkably we found a variant disrupting a highly conserved CNNC motif in mir-301a (phyloP score = 6.93) (***Figure 46B***). In addition, two variants created basal UG motifs and two others created apical UGUG motifs. Overall, our method successfully annotated motif-disrupting and motif-creating variants, although the functional relevance of many of the motif variants was questionable.

Even if a variant affects a highly conserved DROSHA processing motif, it may not be functional if the affected miRNA gene is not expressed at biologically relevant levels in LUAD. In fact, only 12/38 (32%) of the miRNAs affected by motif mutations were expressed, on average, above 10 reads per million (RPM) in LUAD tumors from either the cohort of Gillette et al or TCGA-LUAD

(***Supplementary Table 3*** and ***Supplementary Table 4***). All but one (mir-551b) of our main candidates mentioned above were expressed at >10 RPM in at least one of the two external cohorts. Overall, expression and conservation data suggest that most of the variants that affect DROSHA processing motifs may not be functional, as only 6/38 (16%) variants affected highly conserved nucleotides (phyloP score > 3.8) in miRNAs expressed at moderate or high levels (>10 TPM) in LUAD tumors.

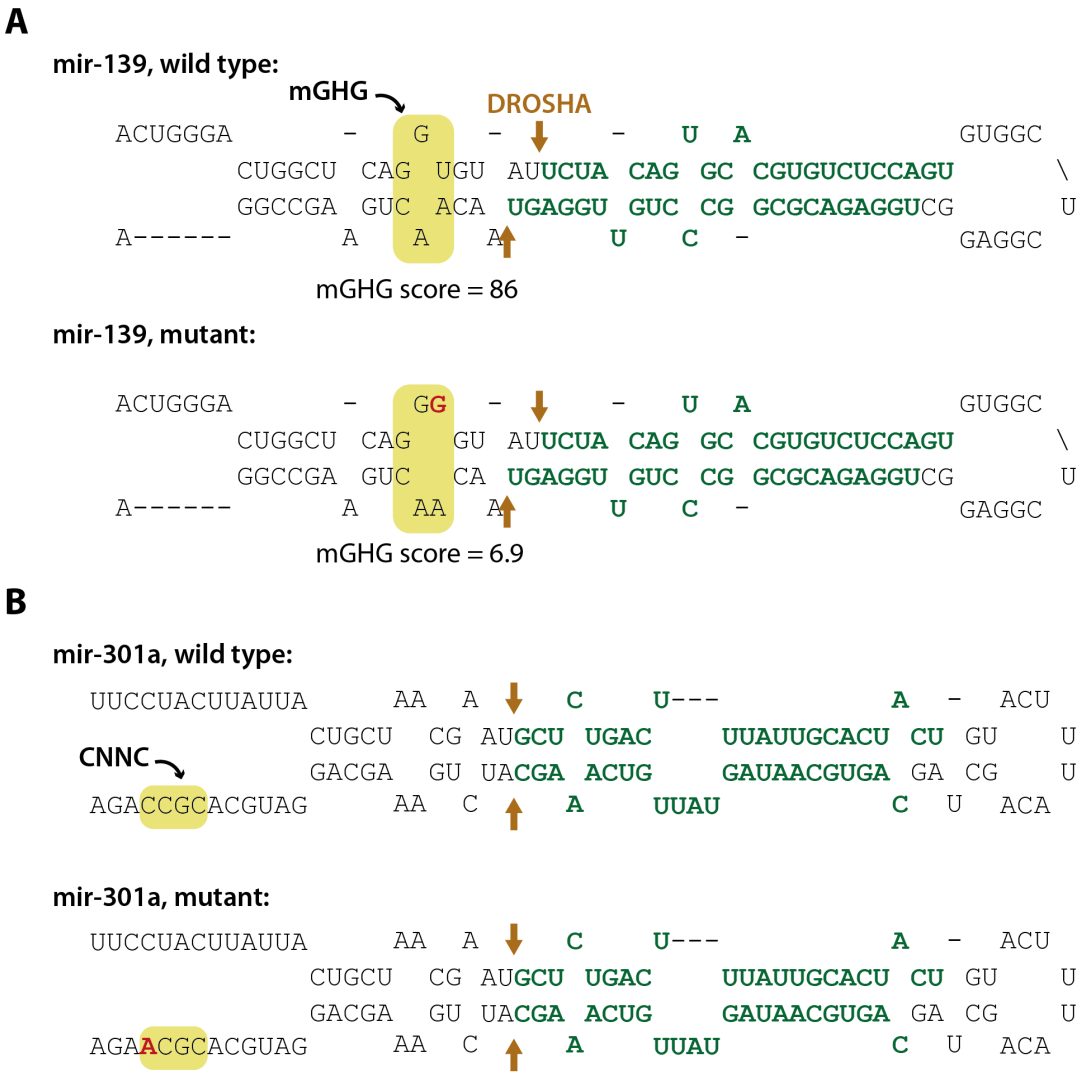


Figure 46. Representative examples of variants in DROSHA processing motifs in pri-miRNAs. ***A.*** *Somatic variant disrupting the mismatched GHG (mGHG) motif of mir-139 in one of our primary tumors.* ***B.*** *Somatic variant disrupting a CNNC motif in mir-301a in a primary tumor from TCGA-LUAD. Secondary structures were predicted using RNAfold. Mature miRNAs are highlighted in green. DROSHA cleavage sites are marked with brown arrows.*

3.4. Discussion

The non-coding genome may be viewed as a potential source of yet-to-be-discovered cancer-related alterations. Accordingly, thanks to the increasing viability of large-scale WGS projects, great efforts have been recently made to systematically search for cancer drivers within the non-coding genome (Fujimoto et al., 2016; Nik-Zainal et al., 2016; Rheinbay et al., 2020; Rheinbay et al., 2017). However, WGS projects have analyzed few primary LUAD samples so far, limiting their statistical power to detect non-coding drivers in the disease. In this work, we aimed to search for non-coding drivers in LUAD cohorts, expanding upon previous efforts, albeit using targeted sequencing with special focus on miRNAs, lncRNAs, and splice regions.

This ***Discussion*** is organized as follows. First, we comment on sequencing study and its limitations in the context of currently available datasets and algorithms. Second, we outline the past, present, and expected future of driver research in the non-coding genome. Third, we discuss our results in lncRNAs within the context of the challenges of studying lncRNA function. Fourth, we highlight the relevance of our findings at splice regions. Finally, we examine our contributions regarding the annotation of miRNA variants and the prediction of DROSHA processing motifs, and we contextualize our findings on miR-133b. Brief discussions on the rest of the individual candidate drivers were already included in the ***Results*** section.

3.4.1. Our targeted sequencing in LUAD: strengths and limitations

3.4.1.1. Cohort size, sequencing depth, and power for driver discovery

In this work, we have performed targeted DNA sequencing, with special focus on miRNAs, lncRNAs, and intronic splice regions, in 70 LUAD primary tumors and 37 LUAD cell lines. Our LUAD cohort was nearly twice as large as that of PCAWG (N = 70 vs. 38, respectively), and our sequencing depth in target regions was also higher (78% vs. 69% target nucleotides covered by at least 30 reads, respectively). However, PCAWG performed WGS and, therefore, it was overwhelmingly more comprehensive than our targeted sequencing. In addition, WES datasets from TCGA-LUAD (N = 582) included miRNA genes, but their coverage was mostly low or even zero. TCGA-LUAD WES data also included splice regions of PCGs, which have been analyzed elsewhere (Shiraishi et al., 2018). Overall, we have considerably increased the amount of sequencing data in non-coding RNAs in LUAD.

Despite our best efforts, our sample size was relatively low and, as a consequence, our statistical power for driver discovery was limited. For example, we only had 80% power to detect driver mutations at frequencies of ≥14% for lncRNA exons and ≥15% for CDSs of PCGs. Although we were more powered to detect driver mutations at splice sites, miRNAs, and promoters, this did not translate into a larger number of hits. There are several reasons for this. First, they are all short genomic features, and therefore it is unlikely for them to reach mutation frequencies of even ≥5% (Andrades et al., 2022). Notable exceptions include the *TERT* promoter and the *MIR142* gene in lymphoma (Rheinbay et al., 2020). Second, every single miRNA modulates a wide range of target mRNAs, thus possibly limiting the selective advantage of permanent sequence changes in miRNAs (Bartel, 2004). Third, promoters often have low sequencing coverage, in part due to their high GC content. For example, researchers from PCAWG estimated that, in their WGS dataset of 2658 tumors, they had <10% power for detecting drivers in nearly 10% of the promoters of cancer genes (Rheinbay et al., 2020). Fourth, in our work, neither

promoters nor splice regions were explicitly included among our sequencing targets, and instead we relied on the fact that regions adjacent to the targets also get sequenced, albeit with lower coverage. These last two factors were not considered in our power analysis, and as a result our power estimates for non-exonic regions were probably overoptimistic. In summary, power analyses for driver discovery must be interpreted within the context of technical limitations and the biological properties of the analyzed genomic regions.

As an example of our low statistical power, *EGFR* was not predicted as a driver in any of our analyses, even though it is one of the most mutated driver genes in LUAD (Bailey et al., 2018). The CDS of *EGFR* had non-synonymous variants in 3/39 cell lines (8%), 0/27 paired primary tumors (0%), 3/70 unpaired primary tumors (4%), and 4/59 TCGA-LUAD WGS samples (7%). These percentages were lower than previously reported in LUAD (~14%) (Bailey et al., 2018; Collisson et al., 2014). Although our low frequency of *EGFR* mutations could be due to differences between our cohort and TCGA-LUAD, it can also be a consequence of sampling variability in our relatively small cohort. This highlights the need for large cohort sizes to obtain comprehensive driver catalogs, as even the top driver genes may not be mutated with enough recurrence in small cohorts.

3.4.1.2. The problems with indels

A major limitation of driver discovery efforts from us and others has been the difficulty of working with indels (Campbell et al., 2020). Because indels insert or delete nucleotides, they usually have stronger functional impact than SNVs. However, indels pose several technical challenges that are currently unresolved and that, taken together, prompted us to discard indels longer than 1 bp for downstream analyses:

- **Indel calling is not robust**. Current variant calling tools strongly disagree in their indel calls, which limits the effectiveness of consensus-based variant calling in indels. For example, in their consensus-based approach, PCAWG only achieved ~60% sensitivity for indel calling, in contrast to their ~95% sensitivity for SNV calling (Campbell et al., 2020). We observed a similar issue in our datasets.

- **Variable number tandem repeats**. Human genomes are highly repetitive (Nurk et al., 2022). Many repetitive sequences, such as microsatellites and minisatellites, occur in tandem, and the number of times each sequence is repeated can vary across individuals. In our experience, a large proportion of the >1 bp indels detected by our consensus-based approach were short sequences repeated in tandem a variable number of times. We found these events to be prone to alignment artifacts that prompted variant calling algorithms to wrongly call them as somatic. In other cases, the difference between a tumor and its matched normal sample was the number of times the sequence was repeated. Finally, even if matched normal samples had not been sequenced, the affected region was often littered with indel polymorphisms in normal population datasets, but the exact indel present in the tumor was not present in the germline dataset and therefore it was not filtered out (Karczewski et al., 2020). In conclusion, indels longer than 1 bp were mostly artifacts that would have biased driver discovery analyses.
- **Driver discovery methods in indels are limited**. The discovery tools used in our work are mostly built for the analysis of SNVs (Arnedo-Pac et al., 2019; Mularoni et al., 2016). Although short indels could be re-coded as SNVs so that they are accepted by driver discovery tools, indel-generating background mutational processes are still poorly understood, and they cause artifacts in driver discovery (Imielinski et al., 2017; Rheinbay et al., 2020).

3.4.1.3. Benchmarking variant calling pipelines

Ideally, to benchmark a variant calling method, a high-confidence "ground truth" set of variants is required. However, acquiring such a dataset is challenging. As an alternative, we compared our pipelines to other pipelines applied to the same datasets. Each pipeline was expected to have a certain proportion of false positives and false negatives and, therefore, there was no "ground truth" set of variants. However, by thoroughly examining the discrepancies between pipelines, the biases and errors of each pipeline could be better understood.

There are other alternatives for benchmarking variant calling pipelines. For example, PCAWG performed targeted deep sequencing of a subset of their samples (Campbell et al., 2020). However, the process is costly, and it could be argued that certain biases of variant calling pipelines may still happen at high sequencing depths, and therefore the set of variants detected by deep sequencing may not be 100% true. Alternatively, tools such as BAMSurgeon simulate tumor BAM files by introducing known artificial variants in a BAM file obtained from a real non-tumor sample (Ewing et al., 2015). However, to simulate mutations reliably, the underlying mechanisms that generate mutations in real tumors must be firmly understood, and current knowledge is incomplete. Still, simulation can be a useful approximation.

We tested BAMSurgeon by simulating 1000 random SNVs and indels within our targets of interest using a normal BAM file from our cohort as input. We simulated homozygous and heterozygous variants, assuming a diploid genome with 80% tumor purity and Gaussian noise affecting the distribution of VAFs. We simulated indel lengths using Zipf's distribution with exponent 2. In our simulation, our paired variant calling pipeline achieved 78% sensitivity and 99.9% precision for SNVs and 63% sensitivity and 99.1% precision for indels. In contrast, PCAWG achieved a sensitivity of 95% (90% confidence interval: 88-98%) and a precision of 95% (71-99%) for SNVs, and a sensitivity of 60% (34-72%) and a precision of 91% (73-96%) for indels (Campbell et al., 2020). Although PCAWG's methods for benchmarking were different from ours, the benchmarks reflect that, in contrast to PCAWG, we prioritized precision over sensitivity, especially for SNVs. Importantly, most of the false negative variants missed by our pipeline were subclonal. We reasoned that, for our purposes of driver discovery, we preferred working with high-confidence somatic variants at the expense of missing a small proportion of subclonal variants, which were unlikely to be biologically relevant.

In summary, no method for benchmarking variant calling is completely accurate because it is currently not possible to obtain a set of "ground truth" variants that are faithfully representative of a real tumor sample. However, within this limitation, we successfully validated our pipelines by independent approaches.

3.4.2. Driver discovery in non-coding DNA

3.4.2.1. Methods for driver discovery and their limitations

Identifying non-coding cancer drivers in targeted sequencing data of small cohorts is a challenging task. In this work, we applied two state-of-the-art driver discovery tools that, in theory, were applicable to targeted sequencing data: OncoDriveFML and OncoDriveCLUSTL (Arnedo-Pac et al., 2019; Mularoni et al., 2016). However, we encountered performance issues with both tools.

Regarding OncoDriveFML, whereas it performed acceptably for CDSs, it underperformed for non-coding features such as lncRNA exons. Because non-coding regions are usually more weakly conserved than coding regions, the differences in functional impact scores between driver and passenger variants may be smaller in non-coding regions than in coding regions, and this could limit the sensitivity of tools that rely on such scores. Furthermore, non-coding regions may be harder to sequence than coding regions. For example, promoters usually have a high GC content, and non-coding RNAs often contain repetitive sequences (Rheinbay et al., 2020). This may have limited our power to detect variants in non-coding regions, which in turn may have limited our power for driver discovery. Taken together, these facts could explain the lower performance of OncoDriveFML in non-coding regions compared to coding regions.

Regarding OncoDriveCLUSTL, it consistently reported inflated p values, even though we had optimized its parameters. In agreement with this, in its original publication, OncoDriveCLUSTL also obtained inflated distributions of p values for some WES datasets from TCGA (Arnedo-Pac et al., 2019). This suggested that p value inflation may be a general issue of OncoDriveCLUSTL, at least under certain conditions. Indeed, we observed that MNVs biased cluster detection. When two or more consecutive nucleotides are mutated in the same sample, they have likely originated from a single mutational process, and therefore they should not be considered as independent events. OncoDriveFML handles MNVs correctly (Mularoni et al., 2016). However, based on our observations, OncoDriveCLUSTL seemed to consider MNVs as

multiple independent events, detecting about twice as many hits in our datasets compared to an analysis that considered MNVs as single events. To our knowledge, this behavior has not been documented. The bias may not be evident in large cohorts, such as the ones analyzed in the original publication, because MNVs may become "diluted" among a large number of SNVs (Arnedo-Pac et al., 2019). However, in our small datasets, two variants close enough to each other were usually sufficient for a genomic feature to be considered a significant hit. As a consequence, the effect of MNVs was more evident in our datasets than in previous work.

Although MNVs explained more than half of the likely false positive hits of OncoDriveCLUSTL, our final analysis that corrected for the MNV bias still had a large number of false positives. One possible explanation is that our cohorts were too small, and therefore any fortuitous occurrence of two or more variants at nearby genomic positions could be identified as a significant cluster. In addition, most hits from OncoDriveCLUSTL were detected in unpaired samples, especially those from our Basque cohort, suggesting that some germline SNPs specific to the Basque population may not have been removed by our unpaired variant calling pipeline. Importantly, even a small proportion of germline SNPs in the input data can bias driver discovery. Finally, localized hypermutation processes are still poorly understood, and they might generate SNVs at nearby but non-adjacent positions, which may also bias cluster detection even if MNVs are handled correctly. This bias may be partially mitigated if the cluster scoring algorithm accounts for variant phasing, i.e., whether two nearby variants from the same patient occur in the same copy or in different copies of the chromosome (Tewhey et al., 2011). However, only some variant calling algorithms, such as MuTect2, report variant phasing in short-read sequencing data.

In conclusion, although recent efforts such as PCAWG have greatly boosted the development of driver discovery methods for non-coding regions, there is still room for improvement (Campbell et al., 2020). Most importantly, future improvements should better account for localized hypermutation processes, which are still poorly understood for non-coding regions.

3.4.2.2. Rare non-coding drivers and the future of non-coding driver research

Our results are consistent with previous reports in that non-coding driver mutations are extremely rare, and that current driver discovery methods report many false positives in non-coding regions (Rheinbay et al., 2020). Beyond splice sites, the *TERT* promoter, and cancer-specific examples such as *MIR142*, few driver point mutations have been found in non-coding DNA (Elliott and Larsson, 2021). Importantly, statistical power in recent WGS efforts should have been enough to detect moderately or highly recurrent non-coding drivers in many cancer types (Fujimoto et al., 2016; Nik-Zainal et al., 2016; Rheinbay et al., 2020; Rheinbay et al., 2017). Although cohort sizes in LUAD (including our own cohorts) have been low, results in other cancer types are discouraging. On the other hand, technical and methodological limitations have prevented the discovery of some recurrent non-coding drivers. For example, promoter regions such as *TERT* are difficult to sequence at high depth due to their high GC content (Rheinbay et al., 2020). In another example, recurrent mutations in the *U1* small non-coding RNA were missed by PCAWG because *U1* has multiple copies and hundreds of pseudogenes in the human genome, complicating mutational analyses (Shuai et al., 2019; Suzuki et al., 2019). Despite these counterexamples, non-coding driver mutations are probably rare for mainly biological reasons, as most point mutations in non-coding DNA are expected to have little or no functional impact. Therefore, many researchers propose that the landscape of non-coding driver mutations is unlikely to change drastically from what PCAWG and previous studies have already revealed (Elliott and Larsson, 2021; Rheinbay et al., 2020).

Major recent WGS studies have not searched for non-coding drivers beyond the *TERT* promoter, reflecting that methods for detecting non-coding drivers may still be immature, or that there may be little hope of finding novel non-coding drivers using current methodologies. Most remarkably, the Hartwig Medical Foundation performed WGS on 2520 metastatic tumors and matched normal tissue, but they restricted driver analyses to CDSs, splice sites of PCGs, and the *TERT* promoter (Priestley et al., 2019). Similarly, WGS studies in

pediatric cancers have either limited their driver analyses to coding regions (Ma et al., 2018) (N = 651) or only found the *TERT* promoter as a non-coding driver (Gröbner et al., 2018) (N = 547). For now, and until better driver discovery methods are developed or more encouraging non-coding drivers are discovered, WGS information on non-coding regions is being mostly used for purposes other than driver discovery, such as studying genome-wide mutational patterns (Priestley et al., 2019).

In the future, as more whole tumor genomes are sequenced and cohort sizes increase to thousands or tens of thousands, very few (if any) novel drivers are expected to be found at moderate or high mutation frequencies (>5-10%). However, even the most comprehensive pan-cancer studies have failed to identify any driver events in at least 5% of tumors, and the driver catalog may be incomplete in at least part of the remaining 95% of tumors (Rheinbay et al., 2020). This raises the question of where these "missing" drivers are. Some researchers propose that the current definition of driver is too restrictive, and that it should be extended to incorporate epigenetic alterations and alterations that are not cell-autonomous (Alizadeh et al., 2015). Alternatively, the answer may reside within the long tail of variants at low or very low frequencies (<5%, or even <1%). In fact, the effect of low-frequency non-coding variants in different genomic features can converge on a single gene or pathway, a phenomenon that has been described for certain mutations in enhancers (Corona et al., 2020; Kim et al., 2016; Zhou et al., 2020). In addition, even if the functional impact of individual non-coding variants is low, cancer genomes may accumulate multiple low-impact variants so that their additive effects converge in an observable phenotype (Castro-Giner et al., 2015). Thus, there may be a large number of undiscovered, low-frequency, and mostly weak driver mutations in non-coding DNA, and even studies such as PCAWG would still be underpowered to detect many of them (Kumar et al., 2020). In this context, while our detected mutations in non-coding RNA genes such as *TUSC7* and *MIR133B* are certainly not highly recurrent, and they may even be tumor-specific, they may still be oncogenic in the tumor contexts where they have been discovered.

3.4.3. The challenges of studying lncRNA function in cancer

In our work, we aimed to identify putative driver lncRNAs in LUAD cohorts. However, evidence supporting even the top candidates was mixed. The issues encountered by us are recurrent in lncRNA research, and they mostly stem from the low expression of lncRNAs and the uncertainty on whether the function of a given lncRNA, if any, is dependent on the lncRNA sequence, on the act of transcription or splicing, on regulatory DNA elements, or on more than one of these possibilities (Kopp and Mendell, 2018).

Our top candidates had very low expression in external RNA-Seq datasets from LUAD and from normal lung samples. Most critically, *TUSC7* had zero expression in most external samples. However, we and others have detected *TUSC7* expression in LUAD and in normal lung by qPCR, thus contradicting RNA-Seq data (Liu et al., 2013; Zhou et al., 2019). Strong discrepancies between qPCR, RNA-Seq, and microarrays have been reported for other very lowly expressed lncRNAs (Seiler et al., 2017). More generally, while RNA-Seq and qPCR tend to agree for most transcripts, their discrepancies are enriched in lowly expressed RNAs (Everaert et al., 2017).

Then, when RNA-Seq and qPCR disagree, which one should be trusted? A first consideration is the amount of input RNA and the limit of detection of both techniques. Moreover, regarding qPCR, it should be confirmed that the lncRNA of interest was specifically amplified and that the signal did not arise from off-target amplification or from contamination with genomic or plasmid DNA (Seiler et al., 2017). On the other hand, RNA-Seq alignment and quantification tools may introduce biases, and the nature and the extent of such biases is not fully understood (Lahens et al., 2014; Robert and Watson, 2015). Therefore, if experimental design and controls are appropriate, low-throughput experiments may be more reliable than high-throughput ones for lowly expressed RNAs.

Generally, if a lncRNA has very low expression, it is unlikely to have a meaningful biological effect. However, a counterexample is the lncRNA *VELUCT*, whose expression in lung cancer cell lines was below the limit of

detection of both RNA-Seq and qPCR, but whose knockdown reproducibly decreased cell viability (Seiler et al., 2017). In such a case, the authors argued that *VELUCT* must have a biological activity in *cis*. Thus, even if *VELUCT* were quickly degraded after its transcription and never left the chromatin-associated RNA fraction, it could still modulate nearby genes, for example by directly binding to DNA. Intriguingly, whereas treatment of lung cancer cell lines with anti-*VELUCT* small interfering RNAs (siRNAs) and siRNA pools caused strong cellular phenotypes, *VELUCT* levels were unchanged upon siRNA treatment. In fact, the cellular machinery required for siRNA action is mostly active in the cytoplasm (Lennox and Behlke, 2015). Therefore, an open question is how, from a mechanistic point of view, cellular phenotypes were strongly and reproducibly affected by anti-*VELUCT* siRNAs. The answer to this question may have further implications in lncRNA research, as knockdown is a popular way of inferring whether a lncRNA of interest has biological activity, and some researchers use siRNAs for knocking down strictly nuclear and lowly expressed lncRNAs.

Low expression of lncRNAs can also distort differential expression analyses and survival analyses. In particular, small differences in the expression of very low-abundance lncRNAs can result in large fold changes, which may lead to the wrong conclusion that a lncRNA is strongly overexpressed or silenced. Standard tools for differential expression analyses of transcriptomic data handle lowly expressed genes by shrinking their fold change estimates (Love et al., 2014). However, in single-gene analyses, such techniques cannot be applied. In addition, low expression values are highly sensitive to statistical noise, which undermines the reproducibility of differential expression analyses and of survival analyses that rely on the stratification of patients by their "high" (above median) or "low" (below median) expression of a lncRNA. These issues are widespread in the lncRNA literature, as exemplified by *VELUCT*, *TUSC7*, *SOX2-OT*, *ZEB2-AS1*, and others (Guo et al., 2018; Hou et al., 2014; Seiler et al., 2017; Zhou et al., 2019). As a consequence, the absolute expression of a lncRNA should always be considered in differential expression analyses and in survival analyses.

Analyses on the potential activity of lncRNAs can be further complicated when the lncRNA overlaps a PCG, as was our case for *SOX2-OT* and for *ZEB2-AS1*. In these cases, it is difficult to untangle the activity of the lncRNA from that of its overlapping PCG. For example, if expression of a lncRNA is correlated with that of its overlapping PCG, it does not necessarily mean that one modulates the other. Instead, it could mean that common mechanisms modulate both genes simultaneously (Cuadros et al., 2019).

Because computational evidence on the driver role of our candidate lncRNAs is mixed, experiments will be needed to confirm whether the variants confer a competitive advantage to LUAD cells. An informative experiment is the competitive cell growth assay, in which cells overexpressing a wild type or mutant lncRNA as well as different fluorophores are co-cultured, and relative fluorescence intensity is tracked over time (Baliñas-Gavira et al., 2020). However, these systems are artificial because they express lncRNAs well above their endogenous levels, and their phenotypes may not represent real biology. Indeed, any proposed mechanism of action of a lncRNA must always take into account its number of copies per cell, more so for mutant lncRNAs, which only constitute a fraction of the total copies of a lncRNA in the cell.

3.4.4. Variants in intronic splice regions may cause major RNA aberrations

Variants that affect RNA splicing are expected to cause major aberrations in gene function, as they often disrupt open reading frames, remove entire exons, introduce or remove codons, or cause premature termination of transcription or translation (Andrades et al., 2022). To date, most research on variants that affect splicing has focused on the first and last two intronic nucleotides, which have the most conserved sequences. Here, we have explored deeper intronic regions. Importantly, this analysis has allowed us to identify a variant in the third position of intron 14 of *MET* that caused skipping of exon 14, a clinically actionable event in LUAD (Frampton et al., 2015; Mathieu et al., 2022). Moreover, we found splice-altering variants at the third and fifth intronic positions in *RBM10*, a LUAD driver gene whose biological role is yet to be fully elucidated (Bailey et al., 2018).

To our knowledge, few studies have searched for splice-altering mutations in cancer beyond the first and last two intronic nucleotides. Remarkably, Shiraishi et al performed such an analysis systematically in pan-cancer WES and matched RNA-Seq data from TCGA (Shiraishi et al., 2018). They found that ~37% of the intronic variants that affected splicing were not located at the first or last two nucleotides, and most of them were at the fifth or third positions in splice donor regions. In agreement with this, all of the splice-altering mutations beyond the first and last two intronic positions in our study affected the fifth or the third intronic positions at splice donor regions.

Finally, variants that are not located at proximal intronic splice regions may also cause aberrant splicing. In particular, splice-altering variants at the last nucleotide of exons may be roughly as frequent as or more frequent than those within introns (Jung et al., 2015; Shiraishi et al., 2018). Furthermore, synonymous variants within exons may also cause aberrant splicing (Supek et al., 2014). Moreover, splice-altering intronic variants may also affect the polypyrimidine tract and branch-site adenosines (Calabrese et al., 2020). In addition, variants located deep within introns may create novel splice sites, generating new exons (Calabrese et al., 2020; Shiraishi et al., 2018). Finally, variants may also affect splicing enhancers and silencers (De Conti et al., 2013). However, an analysis of splice-altering exonic variants was beyond the scope of our work, and our targeted sequencing approach limited our ability to detect variants far away from exon-intron junctions.

3.4.5. MicroRNA variants in cancer

3.4.5.1. Few variants affect miRNAs, and even fewer are cancer drivers

In our cohorts, the mutation rate of miRNAs was low, in agreement with their short length and high sequence constraint (Bartel, 2004). Furthermore, the majority of the already few miRNA variants were likely to be passengers. Our observations agree with those from the PCAWG Consortium, who analyzed whole genomes of over 2600 pan-cancer tumors but only found evidence of positive selection in one miRNA locus, *MIR142*, in one cancer cohort, B-cell

non-Hodgkin lymphoma, confirming previous findings (Kwanhian et al., 2012; Rheinbay et al., 2020). Although the low proportion of miRNA variants with high functional impact, even within mature miRNAs, might seem to contradict the high conservation of miRNA sequences, several factors can explain why many miRNA variants are not functional:

- **Annotation of miRNA genes is inaccurate**. Currently, miRBase is the gold standard annotation resource for miRNAs, and it is used by gene annotation consortia such as GENCODE (Frankish et al., 2019). However, miRBase is plagued with false positives, which may constitute up to two thirds of its entries for humans (Fromm et al., 2015). Indeed, its goal is not to provide a curated set of miRNAs, but to catalog published miRNA sequences (Griffiths-Jones, 2004). Although miRBase introduced a system to identify high-confidence miRNAs, complemented by public voting of which miRNAs its users believe to be "real" (Kozomara and Griffiths-Jones, 2014), this system has been questioned and improved by efforts such as MirGeneDB (Fromm et al., 2015; Fromm et al., 2020).
- **A mutated miRNA may not be expressed at meaningful levels in the affected cells**. Each cell type typically expresses a limited set of miRNAs (Landgraf et al., 2007). In addition, even if a cell expresses a miRNA, its levels may not be high enough for it to have biological activity (Mullokandov et al., 2012). More precisely, the relative levels of a miRNA and its target dictate whether the miRNA can effectively modulate its target (Kilikevicius et al., 2022).
- **Not all nucleotides within a miRNA may be required for its function**. The targets of a miRNA are mostly dictated by the seed sequence and, often, by the nucleotides immediately adjacent to the seed (Bartel, 2009). As a result, sequences within mature miRNAs but outside seeds may tolerate nucleotide changes.

Although variants in miRNA seeds are rare, they can indeed cause disease, as exemplified by miR-142 in lymphoma (Kwanhian et al., 2012; Rheinbay et al., 2020). In addition, rare germline variants in miRNA seeds have been associated with inherited diseases such as progressive hearing loss (miR-96),

retinal dystrophy (miR-204), and ocular syndromes related with cataract (miR-184) (Conte et al., 2015; Hughes et al., 2011; Iliff et al., 2012; Mencía et al., 2009).

Other groups have performed miRNA-centric reanalyses of somatic variants in lung cancer and in pan-cancer WES cohorts from TCGA (Galka-Marciniak et al., 2019; Urbanek-Trzeciak et al., 2020). In TCGA-LUAD, 8 pri-miRNAs accumulated variants over the expected mutation rate under a simple background model (Galka-Marciniak et al., 2019), and only one when using OncoDriveFML in a similar manner to us (Urbanek-Trzeciak et al., 2020). However, it was unclear whether the affected miRNAs were actually functional in LUAD and which variants, if any, had a biological effect. Variants were evenly spread along the affected pri-miRNAs, with no differences in variant frequencies between the seeds, the rest of the mature miRNAs, and the flanking sequences. The authors argued that these variant patterns could be consistent with those of tumor suppressor genes, but an alternative and more likely explanation is that the vast majority of the variants were passengers and no positive selection was acting on them. In fact, fewer than a third of the cancer-specific overmutated miRNAs identified by Urbanek-Trzeciak et al. were expressed in the affected cancer type, and only 7 (12%) were expressed at medium or high levels (Urbanek-Trzeciak et al., 2020). In addition, not all nucleotides within functional pri-miRNAs are necessary for their function. These observations highlight the importance of annotating miRNA variants with complementary metrics, such as expression in the affected tissue and conservation of the mutated nucleotide. Furthermore, we have shown that TCGA WES data has low coverage of most miRNA genes, limiting the power to detect the full spectrum of variants in the miRNome in these datasets. In conclusion, although previous reports set the groundwork for analyses of somatic variants in miRNAs, further improvements were necessary.

3.4.5.2. A novel approach for annotating variants in DROSHA motifs

To our knowledge, we have developed the first pipeline that annotates variants that create or disrupt DROSHA processing motifs in pri-miRNAs using exact positional and structural information, as well as the first that maps mutations in mGHG motifs to changes in mGHG scores. We are aware of two similar efforts to this end: miRNAmotif (Urbanek-Trzeciak et al., 2018) and ADmiRE (Oak et al., 2019). However, both approaches have critical limitations.

Regarding miRNAmotif, it searches for sequence motifs along either whole pre-miRNAs or very broad regions of them, ignoring the strict positional requirements of DROSHA processing motifs (Auyeung et al., 2013; Kim et al., 2021). Furthermore, miRNAmotif is limited to stem-loop sequences deposited in miRBase, which are often incomplete. Finally, miRNAmotif does not annotate mGHG motifs. Reports that use miRNAmotif share the same limitations (Galka-Marciniak et al., 2019; Urbanek-Trzeciak et al., 2020).

On the other hand, ADmiRE annotates variants that disrupt basal UG, apical UGUG, and downstream CNNC motifs by non-specified methods (Oak et al., 2019). Before developing our pipeline, we had initially applied ADmiRE to our data. However, its motif annotations in our data and in the supplementary material of ADmiRE's publication were incorrect. In particular, the sequence motifs predicted by ADmiRE were not at their expected positions, and the nucleotide sequences at those positions did not match the sequences of the motifs. For example, some variants annotated to affect downstream CNNC motifs were found in non-CNNC sequences upstream of the 5p miRNA. The issue seemed to be caused by a source file of pre-annotated motifs, whose content is incorrect. However, because the authors did not specify how they generated the motif annotation file and we found no patterns in the annotation errors, we were unable to further explore the issue.

The frequencies of our identified DROSHA processing motifs roughly agree with previous reports (Auyeung et al., 2013; Fang and Bartel, 2015; Kim et al., 2021; Kwon et al., 2019; Roden et al., 2017). Despite having analyzed different sets of pri-miRNAs by different methods, most reports from us and others agree that the frequency of pri-miRNAs containing basal UG motifs is ~15%;

of apical UGUG motifs, ~30%; of downstream CNNC motifs, ~50%; and of miRNAs with no motif, ~20% (***Table 16***). Auyeung et al reported an unusually high frequency for the basal UG motif (24.3%), possibly because they analyzed a small set of high-confidence pri-miRNAs (Auyeung et al., 2013). Roden et al reported a very low frequency of apical UGUG motifs, possibly because they were the only ones (other than us) to rely on structural criteria and, in contrast to us, they required the UGU/GUG sequence to be strictly 1 nt after the apical junction, which may not be its optimal position (Roden et al., 2017). Finally, the frequency of mGHG motifs in recent reports is higher than initially described because the definition of mGHG was updated recently (Kwon et al., 2019). Overall, our frequencies of DROSHA motifs agrees with previous work.

Table 16. Frequencies of DROSHA processing motifs in our work and in previous reports.

Report	n	UG	UGUG	mGHG	CNNC	None
Auyeung et al (2013)	204	24.3%	28.9%	-	59.4%	21%
Fang and Bartel (2015)	186	-	-	25%	-	-
Roden et al (2017)	1881	~12%	<10%	-	~50%	-
Kim et al (2021)	1816	13.8%	25.1%	35.6%	-	28.9%
Us, structural	842	14%	29%	42%	48%	20%
Us, positional	842	17%	33%	42%	45%	22%

n: number of analyzed pri-miRNAs. The sets of pri-miRNAs were: for Auyeung et al, miRBase 17 miRNAs conserved in mouse; for Fang and Bartel, curated miRBase 17 miRNAs conserved in mouse; for Roden et al, miRBase 21; for Kim et al, experimentally assayable pri-miRNAs from miRBase 21. For CNNC motifs from Roden et al, we report their permissive definition of 3-11 nt downstream of the 3p basal junction. Because Kim et al did not study CNNC motifs, the actual % of pri-miRNAs with no motifs in their dataset may be lower than reported.

3.4.5.3. Experimental reports of variants that affect DROSHA processing

Before the discovery of DROSHA processing motifs, a somatic variant 5 nt upstream of miR-142-5p was described in a diffuse large B-cell lymphoma patient (Kwanhian et al., 2012). This variant led to incorrect processing of miR-142-5p, yielding a mature miRNA that was longer than expected, as confirmed experimentally by Northern blot (Kwanhian et al., 2012). The sequence of the aberrant miRNA and the impact on its targets were not determined. Now, we can map the identified variant to the mGHG motif of mir-142 (Fang and Bartel, 2015; Kwon et al., 2019). The variant decreased DROSHA processing efficiency by ~4.2x (mGHG score from 10.96 to 2.61).

The finding in mir-142, reinterpreted under the lens of current knowledge, challenges current theories in two major ways. First, the mGHG score of wild type mir-142 is so low that the miRNA would be classified as lacking an mGHG motif. The fact that the mGHG variant altered the processing of the pri-miRNA raises the question of whether current methods for annotating mGHG motifs are accurate. Indeed, mGHG scores are based on experimental data on only three pri-miRNAs, which are then extrapolated to all human pri-miRNAs, but this extrapolation may not always be correct. Second, current theories cannot fully explain how an mGHG mutation leads to a longer mature miRNA. mGHG motifs increase the efficiency and accuracy of DROSHA cleavage, and strong mGHG motifs can alone dictate where DROSHA cleaves the pri-miRNA (Fang and Bartel, 2015; Kwon et al., 2019). One possibility is that the mGHG of mir-142 is actually strong and, when it mutated to a weak mGHG, the DROSHA cleavage site was dictated by the basal junction, which is 1 nt upstream from its optimal position. This would shift the 5p DROSHA cleavage site in mutant mir-142 1 nt towards the basal direction. If the DICER cleavage sites did not change, the mature miRNA would be 1 nt longer than the wild type. However, it would then be necessary to explain why DICER, which is thought to act as an accurate "molecular ruler", generated a longer mature miRNA instead of shifting its cleavage site by 1 nt (Park et al., 2011). In conclusion, further studies may be required to fully understand the role of DROSHA processing motifs and to quantify the strength of mGHG motifs.

SNPs that affect miRNA processing have been described in cancer and in other disease contexts. For example, rs2291418, which transforms an apical CGUG sequence in mir-1229 to UGUG, was associated with enhanced production of miR-1229-3p and with increased risk of Alzheimer's disease (Ghanbari et al., 2016). Other SNPs that have been linked to disease and that may affect DROSHA processing efficiency include rs2910164 (Shen et al., 2008), rs11614913 (Hu et al., 2008), rs2910164 (Jazdzewski et al., 2008), six SNPs associated with risk of schizophrenia (Sun et al., 2009), and a G>A SNP in the pri-miRNA of the human herpesvirus miRNA miR-K5 (Gottwein et al., 2006). None of these SNPs overlapped with any known motifs, but they often changed the secondary structure of the stem by introducing or removing mismatches. Therefore, DROSHA processing may not only be affected by sequence motifs, but also by structural features. However, more research is required to determine how mismatches, bulges, and loops affect the efficiency with which DROSHA cleaves pri-miRNAs.

3.4.5.4. Limitations of our motif annotation pipeline

Our pipeline for annotating DROSHA processing motifs in pri-miRNAs was an improvement upon previous methods (Oak et al., 2019; Urbanek-Trzeciak et al., 2018). However, it still suffered from many limitations:

- Current methods for quantifying the strength of mGHG motifs are based on experimental data in only three pri-miRNAs and they may not be applicable to all human pri-miRNAs.
- DROSHA processing efficiency may also be affected by structural features, such as bulges, mismatches, and loops. Although the RNAsnp software can evaluate the effects of SNVs on RNA secondary structure, it is not suitable for pri-miRNAs because it requires input sequences to be at least 200 nt long (Sabarinathan et al., 2013). Therefore, specific computational methods must be developed to predict the impact of variants on the structure of short (< 200 nt) non-coding RNAs.
- Secondary structure predictions of pri-miRNAs are not always accurate, limiting our ability to reliably predict motifs based on structural features.

- The impact of motif mutations on DROSHA processing efficiency is not yet fully understood. For example, downstream DNNC and CNND sequences may partially retain SRSF3 binding capability, limiting the impact of mutations at CNNC motifs (Kim et al., 2021).
- ~46% of the pri-miRNAs from miRBase 21 could not be analyzed due to their lack of annotated mature miRNAs.
- The positions affected by more than 80% of our detected motif variants had either low expression or low conservation, suggesting a high rate of passenger mutations among our findings. However, expression data were averaged across cohorts, and it cannot be ruled out that a miRNA that has low average expression in a cohort may be highly expressed in a subset of tumors, or in a subset of mutant cells within a tumor.

In conclusion, more research is required to fully understand which mutations affect DROSHA processing efficiency, and our method is expected to report many false positive findings and to miss false negatives that might be identified with improved knowledge in the future.

3.4.5.5. miR-133b and its proposed tumor suppressor role in LUAD

We detected a somatic variant in a conserved nucleotide in the seed of miR-133b that may affect its targeting to LUAD driver mRNAs such as *EGFR*. However, we were unable to determine if the variant had a biological effect using available data. Importantly, expression of miR-133b relative to its targets may be too low for it to meaningfully modulate them (Kilikevicius et al., 2022; Mullokandov et al., 2012). In addition, miR-133a-3p, which was expressed ~10 times more than miR-133b in external LUAD cohorts, may compensate for loss of function of miR-133b (Boettger et al., 2014). Still, the miR-133b variant may cause a gain of function, such as targeting new tumor suppressor mRNAs, which miR-133a-3p may not be able to compensate. To explore the possibly oncogenic role of the miR-133b variant, first the relative levels of wild type and mutant miR-133b, miR-133a-3p, and any putative target mRNAs should be experimentally measured in the affected sample.

External high- and low-throughput studies have confirmed that miR-133b is a bona fide miRNA. In particular, in a meta-analysis of 28 866 human small RNA-Seq datasets, mir-133b precursors consistently showed read patterns characteristic of true miRNAs: 5' end homogeneity, patterns of DICER processing, and detection of a clear mature miRNA within the precursor (Alles et al., 2019). In addition, exogenous mir-133b was efficiently processed by HEK 293T cells to generate mature miR-133b (Alles et al., 2019).

Expression of miR-133b was downregulated in LUAD compared to normal lung in two independent cohorts, in agreement with previous reports (Chen and Ruan, 2019; Zhang et al., 2021). However, miR-133b downregulation was not reported in meta-analyses of microarray-based miRNA expression datasets in lung cancer (Guan et al., 2012; Võsa et al., 2013). This discrepancy may be explained by differences between cohorts, technical differences between microarrays and miRNA-Seq, and because expression of miR-133b in LUAD is low. In particular, both in LUAD and in most normal lung samples, miR-133b expression was <10 TPM. Although these expression levels are generally considered to be too low for a miRNA to have biological activity, it cannot be ruled out that miR-133b expression may be higher in a subset of LUAD cells (Kilikevicius et al., 2022; Mullokandov et al., 2012).

Two previous reports claimed that low miR-133b expression is associated with poor survival in TCGA-LUAD (Chen and Ruan, 2019; Zhang et al., 2021). However, we failed to reproduce these findings. Both studies reported a smaller cohort size than that of TCGA-LUAD after excluding patients who lacked information on miRNA expression or survival (Chen and Ruan: N = 396; Zhang et al: N = 470; actual N = 513). None of the studies stated their sample inclusion criteria. In Chen and Ruan's report, the authors had most likely excluded all samples whose miR-133b expression was undetectable. Indeed, when we removed those samples, we reproduced their results (data not shown). However, such an analysis is flawed, as values of zero are valid data points. Regarding Zhang et al's report, we could not find or deduce how they performed the survival analysis. They may have used an older version of TCGA-LUAD that lacked information from some patients, which could explain their lower cohort size. In summary, there is no conclusive evidence that miR-133b expression is associated with LUAD patient survival.

Previous reports have explored the functional role of miR-133b in various cancers, including LUAD. A myriad of phenotypical effects and mRNA targets have been proposed for miR-133b, but little to none of them have been robustly validated. In LUAD, miR-133b may target *EGFR*, suppressing tumor phenotypes by inducing apoptosis, enhancing drug response, and inhibiting cell invasion (Liu et al., 2012). In agreement with this, we found *EGFR* to be a predicted target of miR-133b by multiple methods, and EGFR protein expression was negatively correlated with miR-133b expression. Other studies that link miR-133b to different cancer phenotypes and targets have been reviewed elsewhere (Li et al., 2017).

To our knowledge, the report by Liu et al is the most rigorous attempt at characterizing miR-133b in LUAD. Nevertheless, it has major flaws that are highly recurrent in the non-coding RNA bibliography (Kilikevicius et al., 2022; Liu et al., 2012). First, Liu et al measured miR-133b and *EGFR* mRNA expression in 27 LUAD patients, finding a negative correlation (Liu et al., 2012). However, they did not justify why they exclusively measured this miRNA-mRNA pair. Next, they showed that miR-133b targets the 3'-UTR of *EGFR* by luciferase assays in which miR-133b was overexpressed *in vitro*. However, they did not report the relative endogenous levels of miR-133b and *EGFR* mRNA. Therefore, it was unclear whether miR-133b could modulate *EGFR in vivo*, as miR-133b expression may be too low relative to *EGFR*. Furthermore, they knocked down endogenous miR-133b using an antisense oligonucleotide, observing phenotypical effects. However, here we have shown that: (i) miR-133a-3p has a nearly identical sequence to miR-133b; (ii) both miRNAs may be functionally redundant; and (iii) miR-133a-3p expression is an order of magnitude higher than that of miR-133b in LUAD. Therefore, miR-133a-3p may be an off-target of anti-miR-133b, and knockdown of miR-133a-3p may have played a major role in the observed phenotype. We were unable to further explore this possibility because the sequence of anti-miR-133b was not available. In summary, although numerous reports have associated miR-133b with tumor suppressor activity in LUAD and in other cancers, it is still unclear whether endogenous miR-133b has an actual function in LUAD.

Chapter 4. Non-coding mutations in diffuse large B cell lymphoma

This Chapter addresses ***Objective 5***, pertaining to the identification of cancer-promoting splice site mutations in diffuse large B-cell lymphoma. The content of this Chapter has been published in "Andrades et al (2022). Recurrent splice site mutations affect key diffuse large B-cell lymphoma genes. Blood 139, 2406-2410". Here, we include the contents of the article, with editing and additions for further context.

4.1. Background: diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most frequent lymphoid malignancy in adults. The most popular molecular classification of DLBCL is based on gene expression signatures, resulting in three major subtypes: germinal center B cell-like (GCB), activated B cell-like (ABC), and unclassified DLBCL (Reddy et al., 2017). Each DLBCL subtype has unique molecular and histological features:

- **GCB DLBCLs** are most frequently mutated in genes involved in chromatin remodeling, such as *KMT2D* and *EZH2* (Reddy et al., 2017; Schmitz et al., 2018; Young et al., 2019). Phenotypically, GCB DLBCLs resemble cells from the germinal center, a structure that is generated in lymph nodes when an organism is exposed to an antigen. In germinal centers, undifferentiated precursors of B cells experience somatic hypermutation to alter their repertoire of immunoglobulins (Igs) at their B cell receptors (BCRs) (Young et al., 2019). BCRs are composed of a membrane-bound immunoglobulin and a heterodimer of CD79A and CD79B, which mediates intracellular signaling and BCR degradation. The BCRs of most GCB DLBCLs contain IgG.

- **ABC DLBCLs** are most frequently mutated in genes involved in BCR signaling, such as *CD79B* and *MYD88*, because they rely on chronic active BCR signaling (Reddy et al., 2017; Schmitz et al., 2018; Young et al., 2019). Phenotypically, ABC DLBCLs resemble plasmablasts, which are the precursors of the antibody-secreting plasma cells (Young et al., 2019). The BCRs of most ABC DLBCLs contain IgM.
- **Unclassified DLBCLs** display intermediate gene expression patterns between GCB and ABC.

The expression-based molecular classification of DLBCL is gradually being replaced by a novel classification based on mutational patterns (Schmitz et al., 2018). Importantly, because each gene expression subtype has mutations in specific genes, there is a high correlation between both classifications. The new classification scheme proposes the following subtypes of DLBCL:

- **EZB**, which is defined by *EZH2* gain-of-function mutations and/or *BCL2* translocations. It is mostly composed of GCB DLBCLs.
- **MCD**, which is defined by gain-of-function mutations in *MYD88* and/or *CD79B*. It is mostly composed of ABC DLBCLs.
- **BN2**, which is defined by *BCL6* fusions and/or *NOTCH2* mutations. It is a rare subtype and it mostly consists of ABC DLBCLs.
- **N1**, which is defined by *NOTCH1* mutations. It includes GCB, ABC, and unclassified DLBCLs.

Recently, landmark multi-omic studies have provided comprehensive collections of molecular alterations in over 1700 DLBCLs ("Reddy et al": N = 1001; "Schmitz et al": N = 574; "Chapuy et al": N = 136) (Chapuy et al., 2018; Reddy et al., 2017; Schmitz et al., 2018). We previously reported that *BCL7A*, a tumor suppressor gene in DLBCL, is recurrently mutated at its first splice donor site in DLBCL (Baliñas-Gavira et al., 2020). Although these splice site mutations impaired the function of *BCL7A*, they had been overlooked by large-scale studies (Baliñas-Gavira et al., 2020). Importantly, Reddy et al provided the largest whole-exome sequencing dataset to date in DLBCL (N = 1001), but they did not analyze splice sites (Reddy et al., 2017). Based on our experience, we wondered if other genes undergo recurrent but overlooked splice site mutations in DLBCL.

4.2. Materials and methods

4.2.1. Variant calling

To identify previously missed splice site mutations in Reddy et al's dataset, we performed unpaired variant calling at splice sites followed by strict filtering. Due to the lack of data from matched normal samples, we performed unpaired variant calling. We used Mutect2 (GATK v4.1.4.0) providing the hg19 human genome, a germline resource (gnomAD v2.1.1) and a panel of normals from the GATK website (https://console.cloud.google.com/storage/browser/gatk-best-practices/somatic-b37). All other Mutect2 options were left at their default values. The variant calling was restricted to all exon boundaries ±10 bp according to GENCODE v28lift37.

We filtered Mutect2 calls using `FilterMutectCalls` after running `GetPileupSummaries`, both of them with default parameters.

4.2.2. Variant annotation and filtering

We annotated the Mutect2 filtered variants using vcf2maf (https://github.com/mskcc/vcf2maf) (Ensembl VEP 97). At this point, most gnomAD germline polymorphisms had already been removed by Mutect2. However, we noticed that ~0.5% of the mutations that passed Mutect2's filters had a gnomAD frequency above 1% and ~0.4% had a gnomAD frequency above 10%. These observations prompted us to apply an additional conservative gnomAD-based hard filter. Thus, we restricted the output to splice site variants whose allele frequencies were ≤ 0.01% in gnomAD. To decide the optimal allele frequency threshold, first we studied the effect of different thresholds on the number of variants that passed the filters, and we found that thresholds within the range of 0.005%-1% yielded roughly similar results, whereas thresholds below 0.005% led to overfiltering.

Still, the number of splice site mutations was almost 70,000, affecting over 15,000 genes, and we reasoned that further filters were required to identify the truly somatic and biologically relevant mutations. Therefore, we only considered mutations in nucleotide positions where Schmitz et al or Chapuy

et al reported at least one somatic mutation. The filters retained 426 genes that had at least one splice site mutation in the dataset of Reddy et al. All analyses were performed using hg19 coordinates.

Finally, we focused on the genes that had splice site mutations in at least 5 patients from Reddy et al, resulting in a final set of 29 genes.

4.2.3. Mutation frequencies per nucleotide in splice sites and in coding sequences

The ratio between the frequency of splice site mutations per nucleotide and the frequency of coding sequence (CDS) mutations per nucleotide was estimated for the 29 recurrent splice site mutant genes from the dataset of Reddy et al. The number of CDS mutations was retrieved from the original study, whereas the number of splice site mutations was estimated in our reanalysis. The lengths of the CDSs, in nucleotides, were estimated by summing the lengths of all non-overlapping CDSs of all isoforms of the genes according to GENCODE v28lift37. The lengths of the splice sites, in nucleotides, were estimated based on all protein-coding isoforms of the genes according to GENCODE v28lift37.

4.2.4. Analysis of RNA aberrations

To detect RNA aberrations induced by splice site mutations in our 29 selected genes, three complementary approaches were combined:

- **MAJIQ deltapsi**. In this approach, for each gene, we compared all samples mutated in splice sites against the rest of the samples. This allowed us to detect recurrent RNA aberrations induced by splice site mutations as long as there was a hotspot that accumulated most of the splice site mutations in the gene. However, it was not an appropriate approach when: (i) a gene had mutations in different splice sites and there was no clear hotspot; or (ii) mutations in the same splice site led to different RNA aberrations in different samples.

- **MAJIQ psi**. In this approach, we assessed RNA aberrations on a per-sample basis. This allowed us to detect RNA aberrations that were unique to one or few samples. To minimize false positives (e.g., RNA aberrations that are also present in wild type samples), we thoroughly studied each aberration in a random selection of wild type samples and we confirmed our observations on Integrative Genomics Viewer (IGV; see below).
- **Manual curation on IGV**. Each of the splice site mutant regions was observed on IGV. This allowed us to visually confirm the results from MAJIQ and to manually rescue events not detected by MAJIQ, which we labeled as "mutant splice site" (see "Classification of RNA aberrations" section below).

Because the RNA-Seq BAM files from Reddy et al and from Schmitz et al had been generated by different methods and had different characteristics, the configuration of MAJIQ was different for each dataset.

4.2.4.1. Reddy et al

For MAJIQ configuration, we set `strandedness=reverse`. We considered each sample as a separate condition and we used Ensembl 74 (GRCh37) as a transcript database, after removing non-coding transcripts, because it was the annotation database used by Reddy et al for their RNA-Seq alignment.

Due to the relatively low depth of the data, we decided to set parameters for majiq build, majiq psi, and majiq deltapsi that were less strict than the default ones.

For `majiq build`, we set `--minreads 3 --minpos 2 --min-denovo 3 --irnbins 0.1 --min-experiments 1`.

For `majiq psi` and `majiq deltapsi`, we set `--minreads 3 --minpos 2 --min-experiments 1`. In the deltapsi analysis, we considered an event as significant if abs(E(dPSI)) > 0.1 and confidence > 0.95.

4.2.4.2. Schmitz et al

Here, we set `strandedness=none` and the transcript database was GENCODE v22 (hg38), downloaded from the GDC Data Portal (https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files), after which we removed non-coding transcripts. The rest of the parameters and criteria were the same as for Reddy et al's dataset.

We retrieved most of the splice site mutant patients for each gene of interest from the MAF files downloaded from Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/DLBCL-2018). The only exception was *BCL7A*, for which we also incorporated the splice site mutations that we detected in our previous reanalysis of the dataset (Baliñas-Gavira et al., 2020).

4.2.4.3. Classification of RNA aberrations

RNA aberrations were classified in the following categories:

- **Intron retention**. The full splice site mutant intron is retained. To avoid false positives due to unspliced pre-mRNA detection, we required the intron to be retained at a higher psi than wild type samples and than adjacent introns from the same sample.
- **Cryptic splice site**. An alternative splice donor or acceptor site is used, leading to deletions (if the cryptic splice site is within an exon) or insertions (if the cryptic splice site is within an intron). There must be no canonical isoforms of the gene (in GENCODEv28lift37) that use these alternative splice sites.
- **Exon skipping**. An exon that is adjacent to the splice site mutation is skipped. There must be no canonical isoforms of the gene (in GENCODEv28lift37) that skip the exon.
- **Alternative isoform**. Increased usage of an isoform that should be unaffected by the splice site mutation. The isoform must be defined in GENCODEv28lift37.

- **Mutant splice site**. This term agglutinated two phenomena in which the intronic mutation was present at the RNA level, but read patterns did not suggest any of the other RNA aberrations described above:
    - Partial intron retention, as defined in ***Section 1.8.2***.
    - Cases of full intron retention in which the mutant splice site was detected in the RNA, but intron retention also occurred at similar rates in wild type samples.
    - In both cases, we required the intronic splice site to have at least 10% of the depth of the last adjacent exonic nucleotide.

4.2.5. Statistical analyses

Further statistical analyses were performed using R (version 4.0.2). Multivariate Cox survival analyses were performed using the `coxph()` function from the `survival` package. The cohort of Reddy et al had homogeneous treatment (rituximab-containing standard regimen) and all tumors were *de novo*, and therefore all patients who had available survival information were included. In the cohort of Schmitz et al, we limited the analysis to the patients that met the following criteria: (i) survival information was available; (ii) biopsy was acquired prior to any treatment; (iii) the patient was treated with standard chemoimmunotherapy. For a gene to be considered in the survival analyses, we required it to be mutated in more than 5 patients.

In R syntax, the survival model was as follows:

```
coxph(Surv(survival_time, event) ~ is_mutant +
IPI_group)
```

Where `is_mutant` was 0 if a particular gene of interest was wild type and 1 if it was mutant, and `IPI_group` was categorized as "Low" (IPI = 0-1), "Intermediate" (IPI = 2-3) or "High" (IPI = 4-5). We considered all exonic mutations plus the splice site mutations that caused RNA aberrations according to our analysis. We did not include genetic subtypes or gene expression subtypes as covariates to avoid multicollinearity with the mutational status. However, we performed survival analyses on the GCB or ABC subsets of the cohorts.

4.2.6. Cell lines

The ABC DLBCL cell lines U-2932 and Ri-1 were purchased from the Leibniz Institute DSMZ. They were grown in RPMI 1640 medium supplemented with 10% FBS, 1% L-glutamine 2 mM, and 100U/ml streptomycin and penicillin.

4.2.7. Plasmids

HIV packaging (psPAX2) and VSV-G (pMD2.G) plasmids were provided by Didier Trono (Addgene plasmids #12260 and #12259 respectively). The packaging plasmid psPAX2 encodes gag, pol, tat and rev genes. The pMD.G plasmid encodes the vesicular stomatitis virus (VSV) G protein. The lentiviral vectors encoding the three versions of CD79B ($CD79B^{WT}$, $CD79B^{Y196H}$ and $CD79B^{IR}$) were designed with Spleen focus-forming virus promoter driving CD79B expression and Human cytomegalovirus immediate early enhancer/promoter driving EGFP expression.

4.2.8. Lentiviral production and titration

Transfection of packaging cells was carried out by lipofection. Briefly, HEK293T cells, grown in DMEM (Cat #L0103-500, Biowest) supplemented with 10% fetal bovine serum (FBS), 100U/ml streptomycin and penicillin, were plated over a 10-cm tissue culture grade Petri dish (Cat #SIAL0167, Sigma-Aldrich) the day before transfection to ensure exponential growth and 80% confluence. pUltra-Chili-Luc vector plasmid, together with packaging (psPAX2) and envelope (pMD2.G) plasmids (18µg total DNA; plasmid proportions of 3:2:1, respectively) were resuspended in 0.5 ml free-serum DMEM and mixed at room temperature for 20 min with 45µL LipoD293 (Cat #SL100668, Signagen Laboratories, Rockville, MD, USA) previously diluted in 0.5 ml free-serum DMEM. The plasmid-lipoD293 mixture was added to cells. Transfection mixture was removed after 5h incubation and 7ml of total medium were carefully added. The viral supernatants were collected at 48 and 72h, and filtered through a 0.45 mm filter (Cat #FPE404030, JET Biofil, Guangzhou, China), aliquoted and immediately frozen at -80ºC.

The percentage of transduced cells was determined on the basis of fluorescence increase due to the expression of GFP. Viral titers (transduction units/ml) were calculated on the basis of the percentage of GFP+ cells detected in the linear range of a serial dilution of the supernatant. Viral titers were estimated in a highly permissive cell line such as K-562.

Transduction of U-2932 cells was carried out in three cycles of infection using 1 mL of the supernatant containing lentiviral particles and 8 μg/mL polybrene in 12x multiwells.

4.2.9. Immunoblot

Lysates (10^6 cells) from untransduced cells as well as from the three cell pools generated by transduction of the different versions of *CD79B* were boiled in Laemmli sample buffer, electrophoresed through a 10% Tris-glycine polyacrylamide gel, and transferred to a PVDF membrane. Immunoblotting antibodies included: mouse anti-CD79B (B29/123) (Santa Cruz Biotechnologies, Cat #sc-53210), goat anti-Lamin B (C-20) (Santa Cruz Biotechnologies, Cat #sc-6216), mouse anti-β-Actin (Sigma, Cat #A5441), rabbit anti-phospho-NF-κB p65 (Ser536) (93H1) (Cell Signaling Technology, Cat #3033), rabbit anti-phospho-Akt (Ser473) (D9E) (Cell Signaling Technology, Cat #4060) and mouse anti-GAPDH (Santa Cruz Biotechnologies, Cat #sc-47724).

4.2.10. Fluorescence-activated cell sorting

Surface expression of IgM in transduced cells was evaluated by flow-cytometry and compared with that of untransduced cells. The ABC DLBCL lines U-2932 and Ri-1 were transduced with $CD79B^{WT}$, $CD79B^{Y196H}$ mutant or $CD79B^{IR}$ and stained on ice with anti-human IgM-PE Monoclonal Antibody (SA-DA4) (eBioscience Catalog # 12-9998-42) and the viability marker 7AAD. Surface IgM expression was measured in transduced subpopulations gated by fluorescence-activated cell sorting to have equivalent GFP expression.

4.3. Results

We found 29 genes that had likely somatic splice site mutations in at least 5 patients from the cohort of Reddy et al (***Figure 47A***). Remarkably, the mutation frequency per nucleotide in these genes was a median of ~8x higher at splice sites than at coding sequences (***Figure 47B***). The accumulation of mutations at splice sites affected known DLBCL genes, such as *BCL7A* (19x), *SGK1* (12x), *CD79B* (9x), and *BCL6* (9x). The inclusion of splice site mutations increased the mutation frequency of *BCL7A* by ~44%, of *SGK1* by ~22%, of *CD79B* by ~32%, and of *BCL6* by ~18%. The mutation frequency of *BCL7A* was comparable to that from our previous report (Baliñas-Gavira et al., 2020). Our analysis also revealed novel genes that were not reported as mutated in the coding sequence by Reddy et al, including *ZFP36L1*, *POU2AF1*, *GRHPR*, *PABPC1*, *CD74*, *LAPTM5*, and *MYO1E.* Interestingly, *ZFP36L1* had a mutation frequency of ~10% in the other two analyzed datasets (~3-4% if only considering splice site mutations) and it has been proposed as a tumor suppressor gene in germinal center-derived B cell lymphomas (Caeser et al., 2019). The splice site mutation frequencies of our 29 selected genes in the dataset of Reddy et al correlated moderately with those from Schmitz et al (Kendall $\tau = 0.28$, $p = 0.04$) and Chapuy et al (Kendall $\tau = 0.44$, $p = 0.002$).

To evaluate the significance of our findings, we analyzed clinically relevant features in the 29 recurrent splice site mutant genes (***Figure 47A***). Of the 29 genes, 18 (62.1%) were known cancer genes according to the Cancer Gene Census (CGC) (Sondka et al., 2018). They included *SGK1*, *BCL7A*, *CD79B* and *PIM1*, among others. Three of the CGC genes had not been reported by Reddy et al: *POU2AF1*, *PABPC1*, and *CD74*. *POU2AF1* is involved in the formation of germinal centers in mice, and splice site mutations in this gene may be related to the transformation of follicular lymphoma to DLBCL (González-Rincón et al., 2019; Teitell, 2003). In addition, the 29 selected genes included 5 of the 12 (42%) genes whose mutations are specific to the GCB DLBCL subtype and 5 of the 8 (63%) genes whose mutations are specific ABC DLBCL subtype according to Reddy et al (Reddy et al., 2017). Furthermore, exonic mutations and RNA-altering splice site mutations (see below) in *CD79B* and in *PIM1* were associated with survival in both Reddy et al's and Schmitz et al's

cohorts according to our multivariate analysis (***Table 17***). Taken together, these results highlight the clinical relevance of splice site mutant genes.

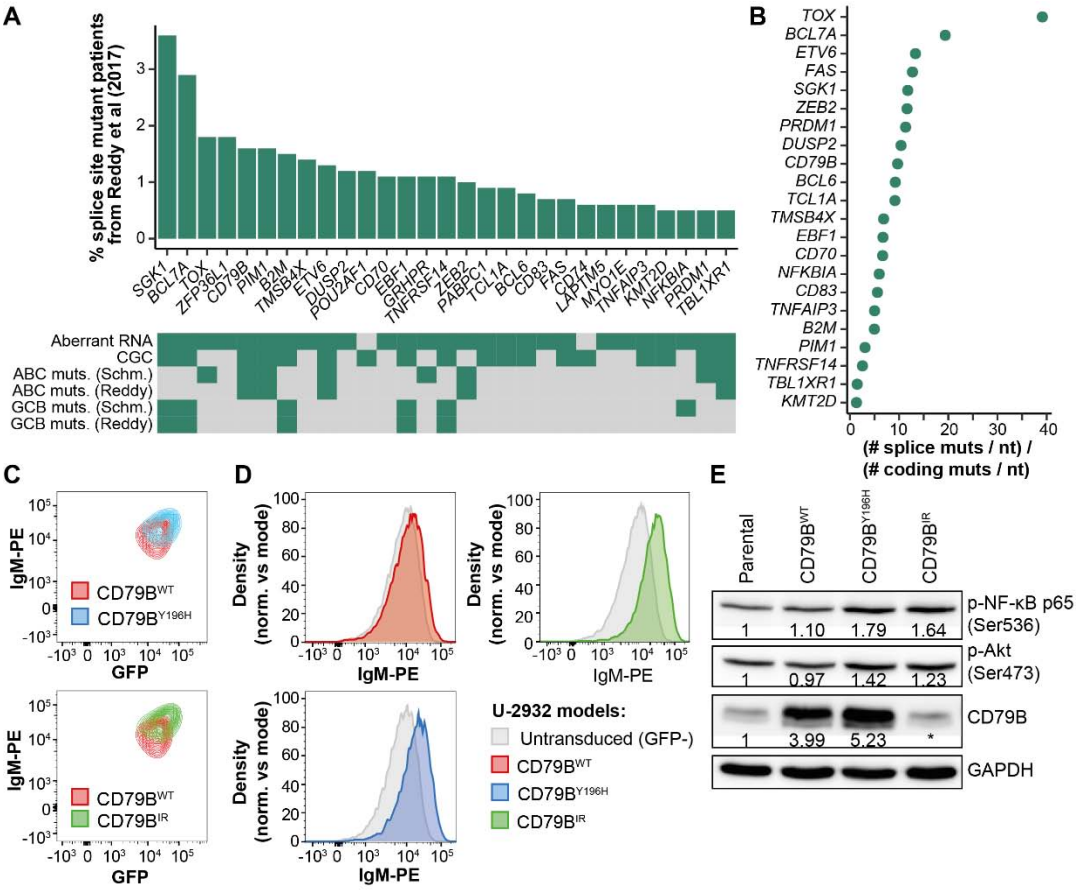


Figure 47. Splice site mutations in DLBCL. A. *Genes recurrently mutated at splice sites according to our reanalysis of Reddy et al's dataset (N = 1001). We only considered mutations that were confirmed as somatic by Schmitz et al (2018) or by Chapuy et al (2018). The genes shown here were mutated in at least 5 patients. For each gene, the frequency of splice site mutant patients is shown in the vertical axis. The heatmap below shows the presence (green) or absence (grey) of relevant features for each gene. "Aberrant RNA evidence": at least one splice site mutant patient had an RNA aberration that could be attributed to the mutation according to our analyses. "CGC": the gene is part of the Cancer Gene Census. "Muts. predict survival": genetic aberrations in the gene are associated with patient survival according to Reddy et al (2017). "GCB-/ABC muts.": the gene is preferentially mutated in the germinal center B cell-like (GCB) or in the activated B cell-like (ABC) subtype according to Reddy et al (2017) or to Schmitz et al (2018).* ***B.*** *Ratio between the frequency of splice site mutations per nucleotide and the frequency of coding mutations per nucleotide for the 22*

recurrent splice site mutant genes that had non-zero coding mutation frequency in Reddy et al (2017)'s dataset. ***C.*** *Mutant CD79B^{IR} increases surface IgM at a higher extent than CD79B^{Y196H}. The ABC DLBCL line U2932 was transduced with either CD79B^{WT}, mutant CD79B^{Y196H} or mutant CD79B^{IR}. Surface IgM is depicted gating on a co-transduced GFP marker to identify the subset of transduced cells with equivalent ectopic CD79B RNA expression.* ***D.*** *Surface IgM expression in U-2932 cells transduced with the indicated CD79B isoforms. U-2932 GFP-positive transduced cells are compared with untransduced cells. E. Western blot of CD79B, phosphorylated RELA/p65 (Ser536), and phosphorylated Akt (Ser473) in U-2932 cells after overexpression of mutant or wild type CD79B. The numbers indicate the fold change between each CD79B overexpression model and the parental cell line, previously normalized with GAPDH, according to a densitometry analysis using ImageJ. *Note that the epitope recognized by the anti-CD79B antibody is the region that is lost by CD79B^{IR}, which is why overexpression cannot be detected by this method. However, we confirmed CD79B^{IR} plasmid overexpression by flow cytometry (GFP+ signal).*

Table 17. Multivariate survival analyses on exonic or splice site mutations.

Cohort	Gene	HR	95% CI	p
Reddy	*PIM1*	1.36	0.99-18.6	0.056
Reddy	*SGK1*	0.43	0.22-0.85	0.015
Reddy	*CD79B*	1.61	1.02-2.55	0.043
Schmitz	*TOX*	3.35	1.59-7.07	0.002
Schmitz	*PIM1*	1.98	1.27-3.10	0.003
Schmitz	*TBL1XR1*	2.51	1.29-4.90	0.007
Schmitz	*CD79B*	2.08	1.16-3.74	0.015
Schmitz	*DUSP2*	1.90	1.06-3.39	0.030

HR: hazard ratio. CI: confidence interval.

Next, we evaluated whether splice site mutations altered the splicing of the affected RNAs using RNA-Seq data from the Reddy and Schmitz cohorts. We used the MAJIQ tool followed by manual curation (Vaquero-Garcia et al., 2016). In 27/29 (93%) genes, at least one splice site mutant patient had an RNA aberration (***Supplemental Files 5-6***, available online from (Andrades et al., 2022)). Furthermore, in 14/29 genes (48%), at least half of the splice site mutant samples in both datasets had RNA aberrations. The most frequent aberration according to a sample-by-sample analysis was intron retention (110 cases), followed by expression of the mutant splice site along with a few intronic nucleotides (94 cases), cryptic splice sites (92 cases), exon skipping (54 cases), and usage of alternative canonical isoforms (49 cases) (***Figure 48***). In *BCL7A*, mutations in the first splice donor site often led to the usage of a cryptic splice donor site in exon 1 that resulted in a loss of 27 aa, as we previously reported (Baliñas-Gavira et al., 2020). Overall, RNA-Seq data has allowed us to confirm the impact of splice site mutations in most genes.

The most frequent RNA aberration affected *CD79B* (***Figure 48***), which accumulated 23 mutations in its fourth splice donor site, out of which at least 18 caused retention of intron 4 (*CD79B*^{IR}). The retained intron introduced a premature stop codon just before the immunoreceptor tyrosine-based activation motif (ITAM) (***Figure 49***). *CD79B* and *CD79A* form heterodimers that, together with immunoglobulins at the B cell membrane, constitute B cell receptors (BCRs) (Davis et al., 2010). The ITAMs of *CD79A* and *CD79B* are involved in BCR signaling and internalization. In DLBCL, the ITAMs of *CD79A* and, most frequently, *CD79B* recurrently undergo point mutations and deletions that prevent BCR internalization, increasing surface BCR levels and causing overactive oncogenic BCR signaling (Davis et al., 2010; Wilson et al., 2015). Indeed, when we overexpressed the most frequent *CD79B*^{IR} variant in U-2932 and Ri-1 (Riva) cells, surface BCR levels increased compared to overexpression of wild type *CD79B* or of the most frequent exonic mutation, CD79B^{Y196H} (***Figure 47C-D***). Furthermore, CD79B^{IR} and CD79B^{Y196H}, but not CD79B^{WT}, increased phosphorylation of AKT and RELA/p65 in U-2932 cells, suggesting an increase in oncogenic signaling via AKT and NF-κB (***Figure 47E***). Taken together, our results highlight the functional relevance of splice site mutations in *CD79B* in DLBCL.

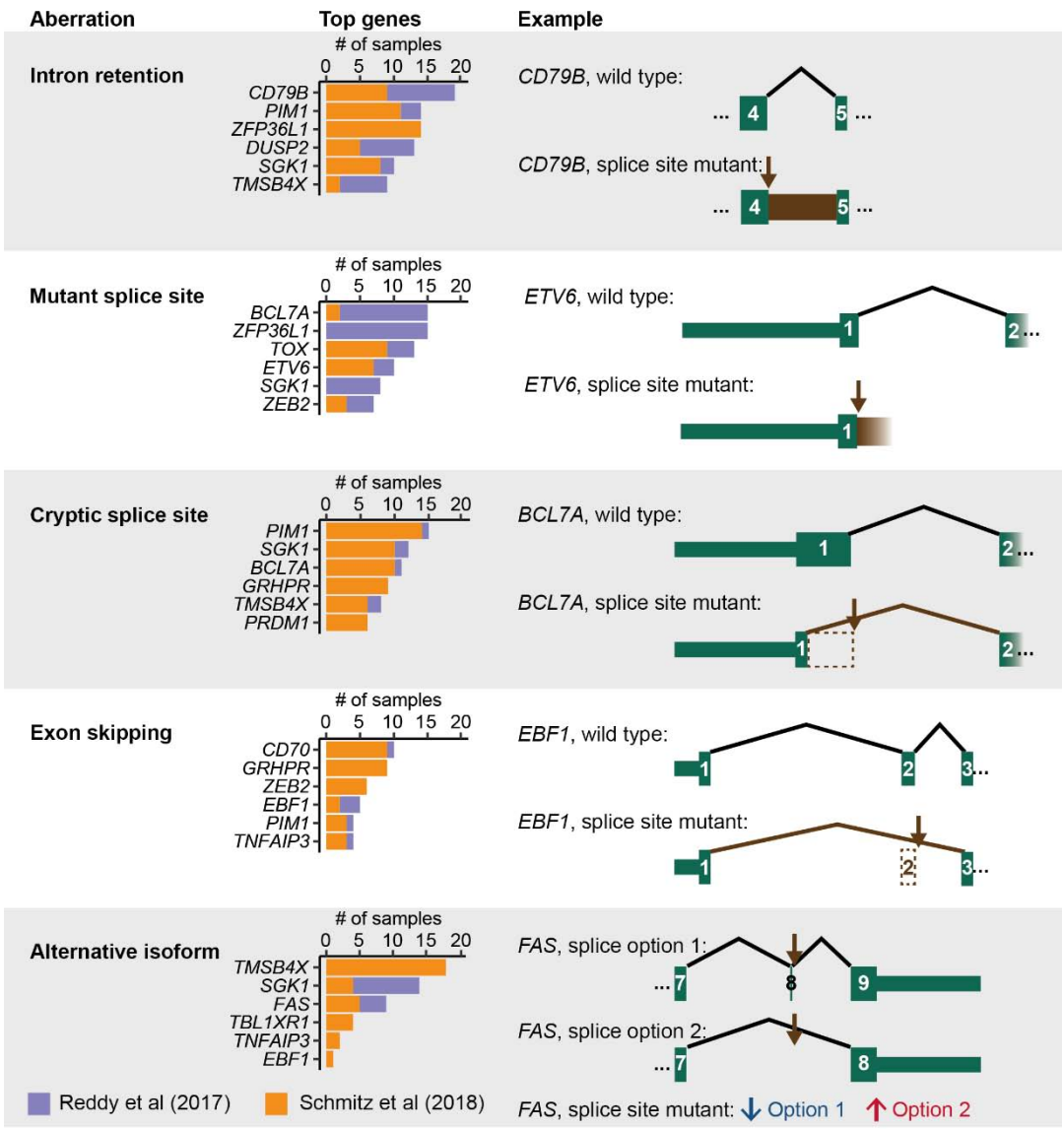


Figure 48. Summary of RNA aberrations identified in splice site mutant genes. *For each type of aberration, the top 6 most affected genes are shown. In addition, one illustrative example is represented schematically. Arrows mark the most recurrently mutated splice site in the gene. Exon numbering was based on the following transcripts from Ensembl v28: ENST00000006750.7 (CD79B), ENST00000396373.8 (ETV6), ENST00000261822.4 (BCL7A), ENST00000313708.10 (EBF1), ENST00000355740.6 (FAS, splice option 1), ENST00000355279.2 (FAS, splice option 2). In FAS, splice site mutant samples have a decreased usage of the "splice option 1" and an increased usage of the "splice option 2". Note that our analysis estimates the differential usage of each splice junction, not the expression of whole transcripts.*

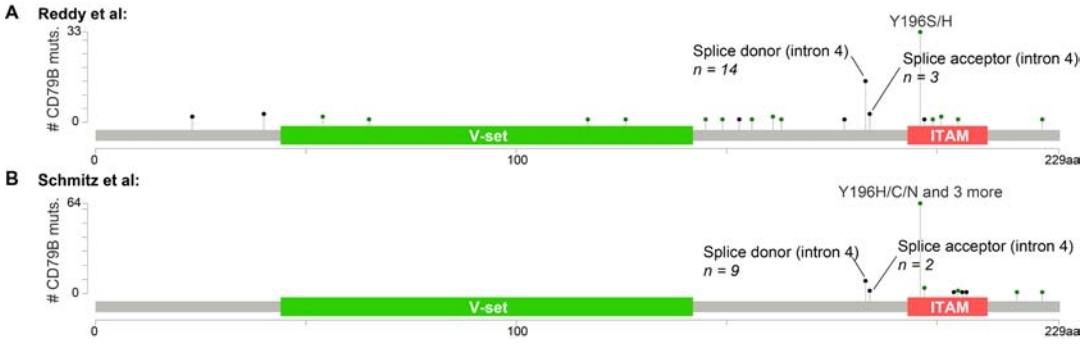


***Figure 49. Distribution of mutations in CD79B**: **A**. In Reddy et al's cohort, based on our reanalysis; **B**. In Schmitz et al's cohort, based on their original report, which already included splice sites. We highlight the splice site mutations in intron 4 as well as the Y196 position. Figure generated using cBioPortal (http://www.cbioportal.org/mutation_mapper).*

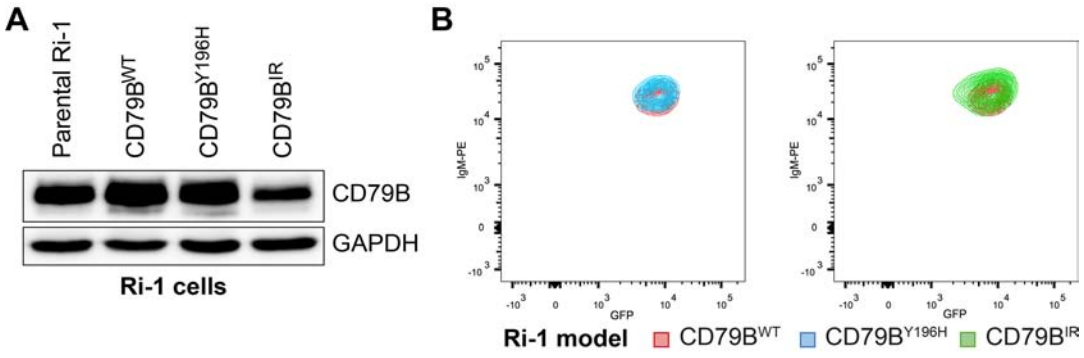


***Figure 50. Phenotypical assays on CD79B variants in Ri-1 cells. A**. Western blot of the overexpression of wild type or mutant CD79B in Ri-1 cells. Note that the epitope recognized by the antibody is the region that is lost by $CD79B^{IR}$, which is why overexpression cannot be detected by this method. However, we confirmed $CD79B^{IR}$ plasmid overexpression by flow cytometry (GFP+ signal). **B.** Changes in surface IgM expression in Ri-1 cells after overexpression of wild type or mutant CD79B. Mutant $CD79B^{IR}$ increases surface IgM at a higher extent than $CD79B^{Y196H}$. The ABC DLBCL line Ri-1 was transduced with either $CD79B^{WT}$, mutant $CD79B^{Y196H}$ or mutant $CD79B^{IR}$. Surface IgM is depicted gating on a co-transduced GFP marker to identify the subset of transduced cells with equivalent ectopic CD79B RNA expression*

4.4. Discussion

Splicing is critical for generating correct mature mRNAs containing full-length, in-frame coding sequences. Accordingly, genes contain highly conserved sequences that mark the boundaries between exons and introns (Sibley et al., 2016). Most of these sequences are intronic and, therefore, non-coding, and for this reason they have been overlooked by previous mutational studies (Reddy et al., 2017). In this work, we have provided a comprehensive collection of recurrent non-coding mutations affecting splice donor and acceptor sites in 1711 DLBCL patients. In addition, we have evaluated whether the splice site mutations altered splicing patterns at the RNA level.

Our analyses were limited by the availability of external data. Because the publicly available dataset of Reddy et al did not contain data from normal samples, we were forced to apply highly conservative filters when we reanalyzed it. As a result, we have most certainly excluded genuine somatic mutations from our report. We expect that an analysis that incorporates information from matched normal samples should reveal even more splice site mutant genes. In addition, more RNA-altering mutations may be discovered by analyzing intronic nucleotides beyond splice donor and acceptor sites, as well as the last exonic nucleotides and splice site-creating mutations (Shiraishi et al., 2018).

In our analysis, the gene most recurrently affected by RNA-altering splice site mutations was *CD79B*. The mutations concentrated in the fourth intron, especially at the splice donor site, creating a truncated CD79B protein. To our knowledge, only one previous report had described splice site mutations at *CD79B* (Zhang et al., 2020). However, the previously reported mutations affected the fourth splice acceptor site, which was rare in our analyzed datasets.

CD79B is one of the most clinically important genes in DLBCL because it is mutated in ~30% of ABC DLBCLs (Reddy et al., 2017; Schmitz et al., 2018). In addition, *CD79B* mutations, together with $MYD88^{L265P}$ mutations, define the MCD genetic subtype (Schmitz et al., 2018). MCD DLBCLs are characterized by a poor prognosis but good response to ibrutinib, a specific inhibitor of the Bruton's tyrosine kinase (BTK) (Schmitz et al., 2018; Wilson et al., 2015). The

degree of response to ibrutinib is largely dictated by whether only *CD79A/B*, only *MYD88*, both, or none are mutated (Wilson et al., 2015). However, clinical and phenotypical studies involving *CD79B* mutations have mostly focused on its point mutations at tyrosine 196 (Y196), which is its most recurrent mutational hotspot. Non-synonymous SNVs at Y196 inactivate an ITAM and, as a result, prevent BCRs from being internalized and degraded, causing overactive BCR signaling (Davis et al., 2010).

According to our findings, splice site mutations at the fourth intron of *CD79B* are recurrent in DLBCL and they truncate the protein, removing its whole ITAM-containing domain. Therefore, an intriguing question is whether the splice site mutations have functional and clinical implications similar to the Y196 mutations. At a functional level, in agreement with our results, a deletion that removes a splice acceptor site immediately before the ITAM of *CD79A* increases BCR levels in the surface of DLBCL cells (Wilson et al., 2015). Moreover, we have shown that overexpression of ITAM-truncated CD79B increases phosphorylation of downstream signaling proteins from the two main oncogenic pathways that are activated by BCRs: NF-κB and AKT (Young et al., 2019).

Then, how are BCRs containing ITAM-truncated CD79B able to transduce signals? Current evidence suggests that, within the CD79A/B heterodimer, CD79A is more important for signal transduction whereas CD79B has more of a regulatory role (Gazumyan et al., 2006). Therefore, BCRs containing ITAM-truncated CD79B may be able to transduce signals as long as CD79A remains functional. This would also explain why *CD79B* is overwhelmingly more mutated than *CD79A* in DLBCL (Reddy et al., 2017; Schmitz et al., 2018). Furthermore, because mutations in *CD79A/B* are heterozygous, mutant and wild type BCRs coexist in the same cell, creating complex interaction networks that act synergistically to promote oncogenesis (Phelan et al., 2018).

Finally, at a clinical level, primary central nervous system lymphomas harboring $CD79B^{Y196}$ and *CD79B* splice site mutations show similar responses to ibrutinib, which agrees with both mutations having similar functional consequences (Lionakis et al., 2017; Wilson et al., 2015).

In conclusion, splice site mutations recurrently affect key DLBCL genes, such as those related with disease subtype or with patient outcome. In particular, mutations in the fourth splice donor site of *CD79B* increase surface BCR levels similarly to the well-known oncogenic $CD79B^{Y196}$ mutations. Splice site mutations can have important clinical applications. In particular, splice site mutant genes may be targeted by anti-cancer drugs, such as the recently FDA-approved capmatinib and tepotinib, which target *MET* exon 14 skipping in metastatic non-small cell lung cancer (Frampton et al., 2015; Mathieu et al., 2022; Wolf et al., 2020). Furthermore, RNA aberrations caused by splice site mutations may generate neoepitopes for immunotherapy (Smart et al., 2018). Therefore, splice site mutations may be a major source of clinically relevant alterations in cancer.

Chapter 5. Conclusions

Conclusions

1. Our variant calling pipelines successfully detected high-confidence somatic variants in all analyzed datasets.
2. State-of-the-art driver discovery methods did not perform adequately in our targeted sequencing datasets of limited size: whereas a functional impact-based method lacked sensitivity in non-coding regions, a clustering-based method had a high false positive rate.
3. The lncRNAs *TUSC7*, *SOX2-OT*, and *ZEB2-AS1* were the top driver candidates in lung adenocarcinoma primary tumors. However, the analysis of the biological effect of their variants was inconclusive, and these lncRNAs may not have RNA sequence-dependent functions.
4. Somatic variants in mature miRNAs are rare in lung adenocarcinoma.
5. A novel somatic variant in the seed of miR-133b in a primary tumor of lung adenocarcinoma alters its predicted targets, which include *EGFR*, *SMARCD1*, and *EIF4A1*.
6. We have developed a computational pipeline that annotates cancer variants in a miRNA-centric manner, successfully predicting the variants that create or disrupt DROSHA processing motifs.
7. Variants located at or beyond the first and last two intronic nucleotides may cause aberrant RNA splicing in *cis* in cancer genes, such as *MET* in lung adenocarcinoma and *CD79B* in diffuse large B cell lymphoma.
8. In diffuse large B cell lymphoma, recurrent variants at the fourth splice donor site of *CD79B* generate a truncated protein that increases the number of surface B cell receptors, promoting oncogenic signaling.

Conclusiones

1. Nuestras metodologías para la detección de variantes somáticas funcionaron satisfactoriamente en todos los conjuntos de datos analizados.
2. Los métodos actuales para la detección de mutaciones conductoras no tuvieron un rendimiento adecuado en nuestros conjuntos de datos de secuenciación dirigida de tamaño limitado: mientras que un método basado en el impacto funcional tuvo baja sensibilidad en regiones no codificantes, un método basado en el agrupamiento de variantes tuvo una tasa elevada de falsos positivos.
3. Los ARNs largos no codificantes *TUSC7*, *SOX2-OT* y *ZEB2-AS1* fueron los principales candidatos a conductores en nuestra cohorte de tumores primarios de adenocarcinoma de pulmón. Sin embargo, el análisis del efecto biológico de sus variantes no fue concluyente, y estos ARNs largos no codificantes podrían tener funciones independientes de la secuencia de ARN.
4. Las variantes somáticas en microARNs maduros son infrecuentes en adenocarcinoma de pulmón.
5. Una mutación somática novedosa en la semilla de miR-133b en un tumor primario de adenocarcinoma de pulmón altera sus dianas predichas, que incluyen *EGFR*, *SMARCD1* y *EIF4A1*.
6. Hemos desarrollado un método computacional que anota variantes de una manera microARN-céntrica, prediciendo las variantes que crean o destruyen motivos de procesamiento por DROSHA.
7. Las variantes ubicadas en o más allá de los primeros y últimos dos nucleótidos intrónicos pueden causar corte y empalme aberrante de ARN en *cis* en genes relacionados con el cáncer, tales como *MET* en adenocarcinoma de pulmón y *CD79B* en linfoma difuso de células B grandes.
8. En linfoma difuso de células B grandes, variantes recurrentes en el cuarto sitio donador de corte y empalme de *CD79B* generan una proteína truncada que incrementa el número de receptores de células B en la superficie celular, promoviendo la señalización oncogénica.

References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*, e05005.

Alizadeh, A.A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., *et al.* (2015). Toward understanding and exploiting tumor heterogeneity. Nature Medicine *21*, 846-853.

Alles, J., Fehlmann, T., Fischer, U., Backes, C., Galata, V., Minet, M., Hart, M., Abu-Halima, M., Grässer, F.A., Lenhof, H.-P., *et al.* (2019). An estimate of the total number of true human miRNAs. Nucleic acids research *47*, 3353-3364.

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., and Mattick, J.S. (2010). lncRNAdb: a reference database for long noncoding RNAs. Nucleic acids research *39*, D146-D151.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Andersson, S., Wallin, K.L., Hellström, A.C., Morrison, L.E., Hjerpe, A., Auer, G., Ried, T., Larsson, C., and Heselmeyer-Haddad, K. (2006). Frequent gain of the human telomerase gene TERC at 3q26 in cervical adenocarcinomas. British journal of cancer *95*, 331-338.

Andrades, A., Álvarez-Pérez, J.C., Patiño-Mercau, J.R., Cuadros, M., Baliñas-Gavira, C., and Medina, P.P. (2022). Recurrent splice site mutations affect key diffuse large B-cell lymphoma genes. Blood *139*, 2406-2410.

Arenas, A.M., Andrades, A., Patiño-Mercau, J.R., Sanjuan-Hidalgo, J., Cuadros, M., García, D.J., Peinado, P., Rodríguez, M.I., Baliñas-Gavira, C., Álvarez-Pérez, J.C., *et al.* (2022). Opportunities of miRNAs in cancer therapeutics. In MicroRNA in Human Malignancies, M. Negrini, G. Calin, and C. Croce, eds. (London, United Kingdom: Elsevier).

Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics *35*, 4788-4790.

Athie, A., Marchese, F.P., González, J., Lozano, T., Raimondi, I., Juvvuna, P.K., Abad, A., Marin-Bejar, O., Serizay, J., Martínez, D., *et al.* (2020). Analysis of copy number alterations reveals the lncRNA ALAL-1 as a regulator of lung cancer immune evasion. Journal of Cell Biology *219*.

Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell *152*, 844-858.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., *et al.* (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385.e318.

Baliñas-Gavira, C., Rodríguez, M.I., and Andrades, A. (2020). Frequent mutations in the amino-terminal domain of BCL7A impair its tumor suppressor role in DLBCL. Leukemia *34*, 2722-2735.

Barta, J.A., Powell, C.A., and Wisnivesky, J.P. (2019). Global Epidemiology of Lung Cancer. Annals of Global Health *85*, 8.

Bartel, D.P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell *116*, 281-297.

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell *136*, 215-233.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome research *10*, 1001-1010.

Beltran, M., Puig, I., Peña, C., García, J.M., Álvarez, A.B., Peña, R., Bonilla, F., and de Herreros, A.G. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. Genes & Development *22*, 756-769.

Boettger, T., Wüst, S., Nolte, H., and Braun, T. (2014). The miR-206/133b cluster is dispensable for development, survival and regeneration of skeletal muscle. Skeletal Muscle *4*, 23.

Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J., and Olivier, M. (2016). TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Human mutation *37*, 865-876.

Bracken, C.P., Scott, H.S., and Goodall, G.J. (2016). A network-biology perspective of microRNA function and dysfunction in cancer. Nature Reviews Genetics *17*, 719-732.

Caeser, R., Di Re, M., Krupka, J.A., Gao, J., Lara-Chica, M., Dias, J.M.L., Cooke, S.L., Fenner, R., Usheva, Z., Runge, H.F.P., *et al.* (2019). Genetic modification of primary human B cells to model high-grade lymphoma. Nature Communications *10*, 4543.

Cai, H., Liu, X., Zheng, J., Xue, Y., Ma, J., Li, Z., Xi, Z., Li, Z., Bao, M., and Liu, Y. (2017). Long non-coding RNA taurine upregulated 1 enhances tumor-induced angiogenesis through inhibiting microRNA-299 in human glioblastoma. Oncogene *36*, 318-331.

Calabrese, C., Davidson, N.R., Demircioğlu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C.M., *et al.* (2020). Genomic basis for RNA alterations in cancer. Nature *578*, 129-136.

Campbell, B.B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., Davidson, S., Edwards, M., Elvin, J.A., Hodel, K.P., *et al.* (2017). Comprehensive Analysis of Hypermutation in Human Cancer. Cell *171*, 1042-1056.e1010.

Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., *et al.* (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93.

Cao, H., Wahlestedt, C., and Kapranov, P. (2018). Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. Trends in genetics : TIG *34*, 704-721.

Castro-Giner, F., Ratcliffe, P., and Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. Nature Reviews Cancer *15*, 680-685.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. Cell *147*, 358-369.

Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W., Chmielecki, J., Beer, D.G., Cope, L., Creighton, C.J., Danilova, L., *et al.* (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543-550.

Conte, I., Hadfield, K.D., Barbato, S., Carrella, S., Pizzo, M., Bhat, R.S., Carissimo, A., Karali, M., Porter, L.F., Urquhart, J., *et al.* (2015). MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proceedings of the National Academy of Sciences *112*, E3236-E3245.

Corona, R.I., Seo, J.-H., Lin, X., Hazelett, D.J., Reddy, J., Fonseca, M.A.S., Abassi, F., Lin, Y.G., Mhawech-Fauceglia, P.Y., Shah, S.P., *et al.* (2020). Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. Nature Communications *11*, 2020.

Costinean, S., Sandhu, S.K., Pedersen, I.M., Tili, E., Trotta, R., Perrotti, D., Ciarlariello, D., Neviani, P., Harb, J., Kauffman, L.R., *et al.* (2009). Src homology 2 domain–containing inositol-5-phosphatase and CCAAT enhancer-binding protein β are targeted by miR-155 in B cells of Eμ-MiR-155 transgenic mice. Blood *114*, 1374-1382.

Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138-163.

Cuadros, M., Andrades, Á., Coira, I.F., Baliñas, C., Rodríguez, M.I., Álvarez-Pérez, J.C., Peinado, P., Arenas, A.M., García, D.J., Jiménez, P., *et al.* (2019). Expression of the long non-coding RNA TCL6 is associated with clinical outcome in pediatric B-cell acute lymphoblastic leukemia. Blood cancer journal *9*, 93.

Chan, J.J., Zhang, B., Chew, X.H., Salhi, A., Kwok, Z.H., Lim, C.Y., Desi, N., Subramaniam, N., Siemens, A., Kinanti, T., *et al.* (2022). Pan-cancer pervasive upregulation of 3′ UTR splicing drives tumourigenesis. Nature Cell Biology.

Chapuy, B., Stewart, C., Dunford, A.J., Kim, J., Kamburov, A., Redd, R.A., Lawrence, M.S., Roemer, M.G.M., Li, A.J., Ziepert, M., *et al.* (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nature Medicine *24*, 679-690.

Chen, G.-Y., and Ruan, L. (2019). Downregulation Of microRNA-133b And Its Clinical Value In Non-Small Cell Lung Cancer. Onco Targets Ther *12*, 9421-9434.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research *41*, D983-986.

Chen, J.-F., Mandel, E.M., Thomson, J.M., Wu, Q., Callis, T.E., Hammond, S.M., Conlon, F.L., and Wang, D.-Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nature Genetics *38*, 228-233.

Chen, Y., and Wang, X. (2019). miRDB: an online database for prediction of functional microRNA targets. Nucleic acids research *48*, D127-D131.

Chen, Z., Chen, Z., Xu, S., and Zhang, Q. (2022). LncRNA SOX2-OT/miR-30d-5p/PDK1 Regulates PD-L1 Checkpoint Through the mTOR Signaling Pathway to Promote Non-small Cell Lung Cancer Progression and Immune Escape. Frontiers in Genetics *12*.

Davis, R.E., Ngo, V.N., Lenz, G., Tolar, P., Young, R.M., Romesser, P.B., Kohlhammer, H., Lamy, L., Zhao, H., Yang, Y., *et al.* (2010). Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature *463*, 88-92.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley interdisciplinary reviews RNA *4*, 49-60.

de Goede, O.M., Nachun, D.C., Ferraro, N.M., Gloudemans, M.J., Rao, A.S., Smail, C., Eulalio, T.Y., Aguet, F., Ng, B., Xu, J., *et al.* (2021). Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. Cell *184*, 2633-2648.e2619.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Domcke, S., Sinha, R., Levine, D.A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. Nature Communications *4*, 2126.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. Proceedings of the National Academy of Sciences *110*, 5294-5300.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Eddy, S.R. (2013). The ENCODE project: Missteps overshadowing a success. Current Biology *23*, R259-R261.

Elliott, K., and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nature Reviews Cancer *21*, 500-509.

Everaert, C., Luypaert, M., Maag, J.L.V., Cheng, Q.X., Dinger, M.E., Hellemans, J., and Mestdagh, P. (2017). Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Scientific reports *7*, 1559.

Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., *et al.* (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature Methods *12*, 623-630.

Fang, W., and Bartel, D.P. (2015). The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. Mol Cell *60*, 131-145.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., *et al.* (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*, bax028.

Frampton, G.M., Ali, S.M., Rosenzweig, M., Chmielecki, J., Lu, X., Bauer, T.M., Akimov, M., Bufill, J.A., Lee, C., Jentz, D., *et al.* (2015). Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. Cancer discovery *5*, 850-859.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., *et al.* (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic acids research *47*, D766-D773.

Fredriksson, N.J., Ny, L., Nilsson, J.A., and Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nature Genetics *46*, 1258-1263.

Fromm, B., Billipp, T., Peck, L.E., Johansen, M., Tarver, J.E., King, B.L., Newcomb, J.M., Sempere, L.F., Flatmark, K., Hovig, E., *et al.* (2015). A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. Annual Review of Genetics *49*, 213-242.

Fromm, B., Domanska, D., Høye, E., Ovchinnikov, V., Kang, W., Aparicio-Puerta, E., Johansen, M., Flatmark, K., Mathelier, A., Hovig, E., *et al.* (2020). MirGeneDB 2.0: the metazoan microRNA complement. Nucleic acids research *48*, D132-d141.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biology *15*, 480.

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., *et al.* (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nature Genetics *48*, 500-509.

Galka-Marciniak, P., Urbanek-Trzeciak, M.O., Nawrocka, P.M., Dutkiewicz, A., Giefing, M., Lewandowska, M.A., and Kozlowski, P. (2019). Somatic Mutations in miRNA Genes in Lung Cancer-Potential Functional Consequences of Non-Coding Sequence Variants. Cancers (Basel) *11*.

Gazumyan, A., Reichlin , A., and Nussenzweig , M.C. (2006). Igβ tyrosine residues contribute to the control of B cell receptor signaling by regulating receptor internalization. Journal of Experimental Medicine *203*, 1785-1794.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., *et al.* (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122-128.

Ghanbari, M., Ikram, M.A., de Looper, H.W.J., Hofman, A., Erkeland, S.J., Franco, O.H., and Dehghan, A. (2016). Genome-wide identification of microRNA-related variants associated with risk of Alzheimer's disease. Scientific reports *6*, 28387.

Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., *et al.* (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell *182*, 200-225.e235.

Gonzalez-Perez, A., Sabarinathan, R., and Lopez-Bigas, N. (2019). Local Determinants of the Mutational Landscape of the Human Genome. Cell *177*, 101-114.

González-Rincón, J., Méndez, M., Gómez, S., García, J.F., Martín, P., Bellas, C., Pedrosa, L., Rodríguez-Pinilla, S.M., Camacho, F.I., Quero, C., *et al.* (2019).

Unraveling transformation of follicular lymphoma to diffuse large B-cell lymphoma. PLOS ONE *14*, e0212813.

Goodall, G.J., and Wickramasinghe, V.O. (2021). RNA in cancer. Nature Reviews Cancer *21*, 22-36.

Gottwein, E., Cai, X., and Cullen Bryan, R. (2006). A Novel Assay for Viral MicroRNA Function Identifies a Single Nucleotide Polymorphism That Affects Drosha Processing. Journal of Virology *80*, 5321-5326.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A., and Elhaik, E. (2013). On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. Genome Biology and Evolution *5*, 578-590.

Gregory, T.R. (2005). Synergy between sequence and size in Large-scale genomics. Nature Reviews Genetics *6*, 699-708.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic acids research *32*, D109-D111.

Gröbner, S.N., Worst, B.C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V.A., Johann, P.D., Balasubramanian, G.P., Segura-Wang, M., Brabetz, S., *et al.* (2018). The landscape of genomic alterations across childhood cancers. Nature *555*, 321-327.

Gruber, A.J., and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. Nature Reviews Genetics *20*, 599-614.

Guan, P., Yin, Z., Li, X., Wu, W., and Zhou, B. (2012). Meta-analysis of human lung cancer microRNA expression profiling studies comparing cancer tissues with normal tissues. Journal of Experimental & Clinical Cancer Research *31*, 54.

Guo, Y., Hu, Y., Hu, M., He, J., and Li, B. (2018). Long non-coding RNA ZEB2-AS1 promotes proliferation and inhibits apoptosis in human lung cancer cells. Oncol Lett *15*, 5220-5226.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., *et al.* (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071-1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Haerty, W., and Ponting, C.P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome Biology *14*, R49.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer discovery *12*, 31-46.

Hessels, D., Klein Gunnewiek, J.M., van Oort, I., Karthaus, H.F., van Leenders, G.J., van Balken, B., Kiemeney, L.A., Witjes, J.A., and Schalken, J.A. (2003). DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. European urology *44*, 8-15; discussion 15-16.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., *et al.* (2013). TERT promoter mutations in familial and sporadic melanoma. Science (New York, NY) *339*, 959-961.

Hou, Z., Zhao, W., Zhou, J., Shen, L., Zhan, P., Xu, C., Chang, C., Bi, H., Zou, J., Yao, X., *et al.* (2014). A long noncoding RNA Sox2ot regulates lung cancer cell proliferation and is a prognostic indicator of poor survival. The International Journal of Biochemistry & Cell Biology *53*, 380-388.

Hu, Z., Chen, J., Tian, T., Zhou, X., Gu, H., Xu, L., Zeng, Y., Miao, R., Jin, G., Ma, H., *et al.* (2008). Genetic variants of miRNA sequences and non–small cell lung cancer survival. The Journal of Clinical Investigation *118*, 2600-2608.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science (New York, NY) *339*, 957-959.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat Med *21*, 1253-1261.

Hughes, A.E., Bradley, D.T., Campbell, M., Lechner, J., Dash, D.P., Simpson, D.A., and Willoughby, C.E. (2011). Mutation altering the miR-184 seed region causes familial keratoconus with cataract. American journal of human genetics *89*, 628-633.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., *et al.* (2011). Extensive and

coordinated transcription of noncoding RNAs within cell-cycle promoters. Nature Genetics *43*, 621-629.

Ibrahim, M.M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., Tursun, B., Zinzen, R.P., Lacadie, S.A., and Ohler, U. (2018). Determinants of promoter and enhancer transcription directionality in metazoans. Nature Communications *9*, 4472.

Iliff, B.W., Riazuddin, S.A., and Gottsch, J.D. (2012). A Single-Base Substitution in the Seed Region of miR-184 Causes EDICT Syndrome. Investigative Ophthalmology & Visual Science *53*, 348-353.

Imielinski, M., Guo, G., and Meyerson, M. (2017). Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. Cell *168*, 460-472.e414.

Jazdzewski, K., Murray, E.L., Franssila, K., Jarzab, B., Schoenberg, D.R., and Chapelle, A.d.l. (2008). Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. Proceedings of the National Academy of Sciences *105*, 7269-7274.

Jeggari, A., Marks, D.S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics *28*, 2062-2063.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., *et al.* (2003). 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.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nature Genetics *47*, 1242-1248.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443.

Karimzadeh, M., Ernst, C., Kundaje, A., and Hoffman, M.M. (2018). Umap and Bismap: quantifying genome and methylome mappability. Nucleic acids research *46*, e120-e120.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al.* (2013). Integrative

Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science (New York, NY) *342*, 1235587.

Kilikevicius, A., Meister, G., and Corey, D.R. (2022). Reexamining assumptions about miRNA-guided gene silencing. Nucleic acids research *50*, 617-634.

Kim, K., Baek, S.C., Lee, Y.-Y., Bastiaanssen, C., Kim, J., Kim, H., and Kim, V.N. (2021). A quantitative map of human primary microRNA processing sites. Molecular Cell *81*, 3422-3439.e3411.

Kim, K., Jang, K., Yang, W., Choi, E.-Y., Park, S.-M., Bae, M., Kim, Y.-J., and Choi, J.K. (2016). Chromatin structure–based prediction of recurrent noncoding mutations in cancer. Nature Genetics *48*, 1321-1326.

Kim, M., Rhee, J.K., Choi, H., Kwon, A., Kim, J., Lee, G.D., Jekarl, D.W., Lee, S., Kim, Y., and Kim, T.M. (2017). Passage-dependent accumulation of somatic mutations in mesenchymal stromal cells during in vitro culture revealed by whole genome sequencing. Scientific reports *7*, 14508.

Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., *et al.* (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods *15*, 591-594.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics *46*, 310-315.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research *22*, 568-576.

Kopp, F., and Mendell, J.T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell *172*, 393-407.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene *30*, 1956-1962.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-D73.

Kumar, S., Warrell, J., Li, S., McGillivray, P.D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C.W.Y., Nielsen, M.M., *et al.* (2020). Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell *180*, 915-927.e916.

Kwanhian, W., Lenze, D., Alles, J., Motsch, N., Barth, S., Döll, C., Imig, J., Hummel, M., Tinguely, M., Trivedi, P., *et al.* (2012). MicroRNA-142 is mutated in about 20% of diffuse large B-cell lymphoma. Cancer Medicine *1*, 141-155.

Kwon, S.C., Baek, S.C., Choi, Y.-G., Yang, J., Lee, Y.-s., Woo, J.-S., and Kim, V.N. (2019). Molecular Basis for the Single-Nucleotide Precision of Primary microRNA Processing. Molecular Cell *73*, 505-518.e505.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., *et al.* (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biology *15*, R86.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., *et al.* (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. Cell *129*, 1401-1414.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., *et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Lennox, K.A., and Behlke, M.A. (2015). Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. Nucleic acids research *44*, 863-877.

Li, D., Xia, L., Chen, M., Lin, C., Wu, H., Zhang, Y., Pan, S., and Li, X. (2017). miR-133b, a particular member of myomiRs, coming into playing its unique pathological role in human cancer. Oncotarget *8*.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.

Lionakis, M.S., Dunleavy, K., Roschewski, M., Widemann, B.C., Butman, J.A., Schmitz, R., Yang, Y., Cole, D.E., Melani, C., Higham, C.S., *et al.* (2017). Inhibition of B Cell Receptor Signaling by Ibrutinib in Primary CNS Lymphoma. Cancer Cell *31*, 833-843.e835.

Liu, L., Dilworth, D., Gao, L., Monzon, J., Summers, A., Lassam, N., and Hogg, D. (1999). Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. Nature Genetics *21*, 128-132.

Liu, L., Shao, X., Gao, W., Zhang, Z., Liu, P., Wang, R., Huang, P., Yin, Y., and Shu, Y. (2012). MicroRNA-133b inhibits the growth of non-small-cell lung cancer by targeting the epidermal growth factor receptor. The FEBS Journal *279*, 3800-3812.

Liu, Q., Huang, J., Zhou, N., Zhang, Z., Zhang, A., Lu, Z., Wu, F., and Mo, Y.Y. (2013). LncRNA loc285194 is a p53-regulated tumor suppressor. Nucleic acids research *41*, 4976-4987.

Liu, Y., Shi, M., He, X., Cao, Y., Liu, P., Li, F., Zou, S., Wen, C., Zhan, Q., Xu, Z., *et al.* (2022). LncRNA-PACERR induces pro-tumour macrophages via interacting with miR-671-3p and m6A-reader IGF2BP2 in pancreatic ductal adenocarcinoma. Journal of Hematology & Oncology *15*, 52.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26-26.

Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., *et al.* (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nature Biotechnology *39*, 1453-1465.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Ma, X., Liu, Y., Liu, Y., Alexandrov, L.B., Edmonson, M.N., Gawad, C., Zhou, X., Li, Y., Rusch, M.C., Easton, J., *et al.* (2018). Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature *555*, 371-376.

Marchese, F.P., Raimondi, I., and Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. Genome Biology *18*, 206.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 3.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041.e1021.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., *et al.* (2020). A compendium of mutational cancer driver genes. Nature Reviews Cancer *20*, 555-572.

Mathieu, L.N., Larkins, E., Akinboro, O., Roy, P., Amatya, A.K., Fiero, M.H., Mishra-Kalyani, P.S., Helms, W.S., Myers, C.E., Skinner, A.M., *et al.* (2022). FDA Approval Summary: Capmatinib and Tepotinib for the Treatment of Metastatic NSCLC Harboring MET Exon 14 Skipping Mutations or Alterations. Clinical Cancer Research *28*, 249-254.

Mattick, J.S., and Dinger, M.E. (2013). The extent of functionality in the human genome. The HUGO Journal *7*, 2.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (New York, NY) *337*, 1190-1195.

Mayr, C., and Bartel, D.P. (2009). Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell *138*, 673-684.

Mayr, C., Hemann, M.T., and Bartel, D.P. (2007). Disrupting the Pairing Between *let-7* and *Hmga2* Enhances Oncogenic Transformation. Science (New York, NY) *315*, 1576-1579.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B., Donnelly, P., and The, W.G.S.C. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Medicine *6*, 26.

Meader, S., Ponting, C.P., and Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. Genome research *20*, 1335-1343.

Medina, P.P., Nolde, M., and Slack, F.J. (2010). OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. Nature *467*, 86-90.

Melton, C., Reuter, J.A., Spacek, D.V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nature Genetics *47*, 710-716.

Mencía, Á., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L.A., del Castillo, I., Steel, K.P., Dalmay, T., *et al.* (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nature Genetics *41*, 609-613.

Messeguer, X., Escudero, R., Farré, D., Núñez, O., Martínez, J., and Albà, M.M. (2002). PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics *18*, 333-334.

Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A., and Lopez-Bigas, N. (2021). In silico saturation mutagenesis of cancer genes. Nature *596*, 428-432.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biology *17*, 128.

Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A.D., Tung, N., Israelow, B., Evans, M.J., Sachidanandam, R., and Brown, B.D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nature Methods *9*, 840-846.

Nicholson, A.G., Tsao, M.S., Beasley, M.B., Borczuk, A.C., Brambilla, E., Cooper, W.A., Dacic, S., Jain, D., Kerr, K.M., Lantuejoul, S., *et al.* (2022). The

2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. Journal of Thoracic Oncology *17*, 362-387.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., *et al.* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47-54.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., *et al.* (2022). The complete sequence of a human genome. Science (New York, NY) *376*, 44-53.

Oak, N., Ghosh, R., Huang, K.-l., Wheeler, D.A., Ding, L., and Plon, S.E. (2019). Framework for microRNA variant annotation and prioritization using human population and disease datasets. Human mutation *40*, 73-89.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292-294.

Park, J.-E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., and Kim, V.N. (2011). Dicer recognizes the 5′ end of RNA for efficient and accurate processing. Nature *475*, 201-205.

Partin, A.C., Zhang, K., Jeong, B.-C., Herrell, E., Li, S., Chiu, W., and Nam, Y. (2020). Cryo-EM Structures of Human Drosha and DGCR8 in Complex with Primary MicroRNA. Molecular Cell *78*, 411-422.e414.

Peinado, P., Andrades, A., Cuadros, M., Rodriguez, M.I., Coira, I.F., Garcia, D.J., Álvarez-Perez, J.C., Baliñas-Gavira, C., Arenas, A.M., Patiño-Mercau, J.R., *et al.* (2020). Comprehensive Analysis of SWI/SNF Inactivation in Lung Adenocarcinoma Cell Models. Cancers (Basel) *12*, 3712.

Peinado, P., Andrades, A., Cuadros, M., Rodriguez, M.I., Coira, I.F., Garcia, D.J., Benitez-Cantos, M.S., Cano, C., Zarzuela, E., Muñoz, J., *et al.* (2022). Multi-omic alterations of the SWI/SNF complex define a clinical subgroup in lung adenocarcinoma. Clinical Epigenetics *14*, 42.

Perera, D., Poulos, R.C., Shah, A., Beck, D., Pimanda, J.E., and Wong, J.W.H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature *532*, 259-263.

Phelan, J.D., Young, R.M., Webster, D.E., Roulland, S., Wright, G.W., Kasbekar, M., Shaffer, A.L., Ceribelli, M., Wang, J.Q., Schmitz, R., *et al.* (2018).

A multiprotein supercomplex controlling oncogenic signalling in lymphoma. Nature *560*, 387-391.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome research *20*, 110-121.

Ponting, C.P., and Hardison, R.C. (2011). What fraction of the human genome is functional? Genome research *21*, 1769-1776.

Prensner, J.R., Chen, W., Iyer, M.K., Cao, Q., Ma, T., Han, S., Sahu, A., Malik, R., Wilder-Romans, K., Navone, N., *et al.* (2014). PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. Cancer research *74*, 1651-1660.

Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., *et al.* (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210-216.

Puente, X.S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J.I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., *et al.* (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature *526*, 519-524.

Reddy, A., Zhang, J., Davis, N.S., Moffitt, A.B., Love, C.L., Waldrop, A., Leppa, S., Pasanen, A., Meriranta, L., Karjalainen-Lindsberg, M.L., *et al.* (2017). Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell *171*, 481-494.e415.

Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., *et al.* (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102-111.

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., *et al.* (2017). Recurrent and functional regulatory mutations in breast cancer. Nature *547*, 55-60.

Robert, C., and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biology *16*, 177.

Roden, C., Gaillard, J., Kanoria, S., Rennie, W., Barish, S., Cheng, J., Pan, W., Liu, J., Cotsapas, C., Ding, Y., *et al.* (2017). Novel determinants of mammalian

primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. Genome research *27*, 374-384.

Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., *et al.* (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nature Genetics *44*, 1111-1116.

Sabarinathan, R., Tafer, H., Seemann, S.E., Hofacker, I.L., Stadler, P.F., and Gorodkin, J. (2013). RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. Human mutation *34*, 546-556.

Schmitz, R., Wright, G.W., Huang, D.W., Johnson, C.A., Phelan, J.D., Wang, J.Q., Roulland, S., Kasbekar, M., Young, R.M., Shaffer, A.L., *et al.* (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. New England Journal of Medicine *378*, 1396-1407.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS genetics *7*, e1002071.

Schuster, S.L., and Hsieh, A.C. (2019). The Untranslated Regions of mRNAs in Cancer. Trends in Cancer *5*, 245-262.

Seiler, J., Breinig, M., Caudron-Herger, M., Polycarpou-Schwarz, M., Boutros, M., and Diederichs, S. (2017). The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance. Nucleic acids research *45*, 5458-5469.

Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., *et al.* (2018). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Reports *23*, 282-296.e284.

Shen, J., Ambrosone, C.B., DiCioccio, R.A., Odunsi, K., Lele, S.B., and Zhao, H. (2008). A functional polymorphism in the miR-146a gene and age of familial breast/ovarian cancer diagnosis. Carcinogenesis *29*, 1963-1966.

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536-1543.

Shiraishi, Y., Kataoka, K., Chiba, K., Okada, A., Kogure, Y., Tanaka, H., Ogawa, S., and Miyano, S. (2018). A comprehensive characterization of cis-acting splicing-associated variants in human cancer. Genome research.

Shuai, S., Suzuki, H., Diaz-Navarro, A., Nadeu, F., Kumar, S.A., Gutierrez-Fernandez, A., Delgado, J., Pinyol, M., López-Otín, C., Puente, X.S., *et al.* (2019). The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature *574*, 712-716.

Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nature Reviews Genetics *17*, 407-421.

Skoulidis, F., and Heymach, J.V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. Nature Reviews Cancer *19*, 495-509.

Slack, F.J., and Chinnaiyan, A.M. (2019). The Role of Non-coding RNAs in Oncology. Cell *179*, 1033-1055.

Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.-K., and Van Allen, E.M. (2018). Intron retention is a source of neoepitopes in cancer. Nature Biotechnology *36*, 1056-1058.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer *18*, 696-705.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nature Reviews Molecular Cell Biology *22*, 96-118.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nature structural & molecular biology *14*, 103-105.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D.A., Sommer, S.S., and Rossi, J.J. (2009). SNPs in human miRNA genes affect biogenesis and function. RNA *15*, 1640-1651.

Supek, F., and Lehner, B. (2019). Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair *81*, 102647.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324-1335.

Suzuki, H., Kumar, S.A., Shuai, S., Diaz-Navarro, A., Gutierrez-Fernandez, A., De Antonellis, P., Cavalli, F.M.G., Juraschka, K., Farooq, H., Shibahara, I., *et al.* (2019). Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. Nature *574*, 707-711.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238-2244.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., *et al.* (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research *47*, D941-D947.

Teitell, M.A. (2003). OCA-B regulation of B-cell development and function. Trends in Immunology *24*, 546-553.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nature Reviews Genetics *12*, 215-223.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., *et al.* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, NY) *310*, 644-648.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics *17*, 601-614.

Urbanek-Trzeciak, M.O., Galka-Marciniak, P., Nawrocka, P.M., Kowal, E., Szwec, S., Giefing, M., and Kozlowski, P. (2020). Pan-cancer analysis of somatic mutations in miRNA genes. eBioMedicine *61*.

Urbanek-Trzeciak, M.O., Jaworska, E., and Krzyzosiak, W.J. (2018). miRNAmotif—A Tool for the Prediction of Pre-miRNA–Protein Interactions. International Journal of Molecular Sciences *19*, 4075.

Van den Eynden, J., Basu, S., and Larsson, E. (2016). Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution. PLoS genetics *12*, e1006506.

Vancura, A., Lanzós, A., Bosch-Guiteras, N., Esteban, M.T., Gutierrez, A.H., Haefliger, S., and Johnson, R. (2021). Cancer LncRNA Census 2 (CLC2): an enhanced resource reveals clinical features of cancer lncRNAs. NAR Cancer *3*.

Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife *5*, e11752.

Võsa, U., Vooder, T., Kolde, R., Vilo, J., Metspalu, A., and Annilo, T. (2013). Meta-analysis of microRNA expression in lung cancer. International Journal of Cancer *132*, 2884-2893.

Wang, M., Herbst, R.S., and Boshoff, C. (2021). Toward personalized treatment approaches for non-small-cell lung cancer. Nature Medicine *27*, 1345-1356.

Ward, L.D., and Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science (New York, NY) *337*, 1675-1678.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nature Genetics *46*, 1160-1165.

Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G., *et al.* (2007). Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood *109*, 4599-4606.

Wild, C.P., Weiderpass, E., and Stewart, B.W. (2020). World Cancer Report: Cancer Research for Cancer Prevention (Lyon, France: International Agency for Research on Cancer).

Wilson, W.H., Young, R.M., Schmitz, R., Yang, Y., Pittaluga, S., Wright, G., Lih, C.-J., Williams, P.M., Shaffer, A.L., Gerecitano, J., *et al.* (2015). Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. Nature Medicine *21*, 922-926.

Willingham, A.T., and Gingeras, T.R. (2006). TUF Love for "Junk" DNA. Cell *125*, 1215-1220.

Wolf, J., Seto, T., Han, J.-Y., Reguart, N., Garon, E.B., Groen, H.J.M., Tan, D.S.W., Hida, T., de Jonge, M., Orlov, S.V., *et al.* (2020). Capmatinib in MET Exon 14–Mutated or MET-Amplified Non–Small-Cell Lung Cancer. New England Journal of Medicine *383*, 944-957.

World Health Organization (2022).

Yang, F., Zhang, H., Mei, Y., and Wu, M. (2014). Reciprocal Regulation of HIF-1α and LincRNA-p21 Modulates the Warburg Effect. Molecular Cell *53*, 88-100.

Young, R.M., Phelan, J.D., Wilson, W.H., and Staudt, L.M. (2019). Pathogenic B-cell receptor signaling in lymphoid malignancies: New insights to improve treatment. Immunological Reviews *291*, 190-213.

Zhang, J., Cao, Z., Yang, G., You, L., Zhang, T., and Zhao, Y. (2019). MicroRNA-27a (miR-27a) in Solid Tumors: A Review Based on Mechanisms and Clinical Observations. Frontiers in Oncology *9*.

Zhang, M., Lan, X., and Chen, Y. (2021). MiR-133b suppresses the proliferation, migration and invasion of lung adenocarcinoma cells by targeting SKA3. Cancer Biology & Therapy *22*, 571-578.

Zhang, W., Yang, L., Guan, Y.Q., Shen, K.F., Zhang, M.L., Cai, H.D., Wang, J.C., Wang, Y., Huang, L., Cao, Y., *et al.* (2020). Novel bioinformatic classification system for genetic signatures identification in diffuse large B-cell lymphoma. BMC Cancer *20*, 714.

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., Bu, D., Li, H., Sun, L., Pei, D., *et al.* (2021a). NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic acids research *49*, D165-d171.

Zhao, Z., Xu, Q., Wei, R., Huang, L., Wang, W., Wei, G., and Ni, T. (2021b). Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers. Nucleic acids research *49*, 10369-10381.

Zhou, H., Chen, A., Shen, J., Zhang, X., Hou, M., Li, J., Chen, J., Zou, H., Zhang, Y., Deng, Q., *et al.* (2019). Long non-coding RNA LOC285194 functions as a tumor suppressor by targeting p53 in non-small cell lung cancer. Oncology reports *41*, 15-26.

Zhou, S., Hawley, J.R., Soares, F., Grillo, G., Teng, M., Madani Tonekaboni, S.A., Hua, J.T., Kron, K.J., Mazrooei, P., Ahmed, M., *et al.* (2020). Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. Nature Communications *11*, 441.

Zhou, Y., Zhong, Y., Wang, Y., Zhang, X., Batista, D.L., Gejman, R., Ansell, P.J., Zhao, J., Weng, C., and Klibanski, A. (2007). Activation of p53 by MEG3 Non-coding RNA. Journal of Biological Chemistry *282*, 24731-24742.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research *31*, 3406-3415.

Publications during the PhD

The * symbol denotes equal contribution.

Original articles as a first author:

Peinado, P.*, Andrades, A.*, Cuadros, M.*, Rodriguez, M. I.*, Coira, I. F., Garcia, D. J., Álvarez-Perez, J. C., Baliñas-Gavira, C., Arenas, A. M., Patiño-Mercau, J. R., *et al.* Multi-omic alterations of the SWI/SNF complex define a clinical subgroup in lung adenocarcinoma. (2022). Clinical Epigenetics *14*, 42.

Andrades, A.*, Álvarez-Pérez, J. C.*, Patiño-Mercau, J. R., Cuadros, M., Baliñas-Gavira, C., and Medina, P. P. (2022). Recurrent splice site mutations affect key diffuse large B-cell lymphoma genes. Blood *139*, 2406-2410, doi: 10.1182/blood.2021011708.

Cuadros, M.*, García, D. J.*, Andrades, A.*, Arenas, A. M., Coira, I. F., Baliñas-Gavira, C., Peinado, P., Rodríguez, M. I., Álvarez-Pérez, J. C., Ruiz-Cabello, F., *et al.* (2020). LncRNA-mRNA Co-Expression Analysis Identifies AL133346.1/CCN2 as Biomarkers in Pediatric B-Cell Acute Lymphoblastic Leukemia. Cancers *12*, 3803.

Peinado, P.*, Andrades, A.*, Cuadros, M.*, Rodriguez, M. I.*, Coira, I. F., Garcia, D. J., Álvarez-Perez, J. C., Baliñas-Gavira, C., Arenas, A. M., Patiño-Mercau, J. R., *et al.* (2020). Comprehensive Analysis of SWI/SNF Inactivation in Lung Adenocarcinoma Cell Models. Cancers *12*, 3712.

Cuadros, M.*, Andrades, A.*, Coira, I. F., Baliñas, C., Rodríguez, M. I., Álvarez-Pérez, J. C., Peinado, P., Arenas, A. M., García, D. J., Jiménez, P., *et al.* (2019). Expression of the long non-coding RNA TCL6 is associated with clinical outcome in pediatric B-cell acute lymphoblastic leukemia. Blood Cancer Journal *9*, 93.

Book chapters:

Arenas, A. M.*, Andrades, A.*, Patiño-Mercau, J. R., Sanjuan-Hidalgo, J., Cuadros, M., García, D. J., Peinado, P., Rodríguez, M. I., Baliñas-Gavira, C., Álvarez-Pérez, J. C., and Medina, P. P. (2022). Opportunities of miRNAs in cancer therapeutics. In MicroRNA in Human Malignancies, M. Negrini, G. Calin, and C. Croce, eds. (London, UK: Academic Press, Elsevier)

Original articles as a non-first author:

Esposito, R., Polidori, T., Meise, D. F., [...], Andrades, A., [...], and Johnson, R. (2021). Multi-hallmark long noncoding RNA maps reveal non-small cell lung cancer vulnerabilities. bioRxiv, doi: 10.1101/2021.10.19.464956.

Boyero, L., Martin-Padron, J., Fárez-Vidal, M. E., Rodríguez, M. I., Andrades, A., Peinado, P., Arenas, A. M., Ritoré-Salazar, F., Alvarez-Perez, J. C., Cuadros, M., and Medina, P. P. (2022). PKP1 and MYC create a feedforward loop linking transcription and translation in squamous cell lung cancer. Cellular Oncology, *in press*, doi: 10.1007/s13402-022-00660-1.

Romero, O. A., Vilarrubi, A., Alburquerque-Bejar, J. J., Gomez, A., Andrades, A., Trastulli, D., Pros, E., Setien, F., Verdura, S., Farré, L., *et al.* (2021). SMARCA4 deficient tumours are vulnerable to KDM6A/UTX and KDM6B/JMJD3 blockade. Nature Communications *12*, 4319.

Peinado, P., Andrades, A., Martorell-Marugán, J., Haswell, J. R., Slack, F. J., Carmona-Sáez, P., and Medina, P. P. (2021). The SWI/SNF complex regulates the expression of miR-222, a tumor suppressor microRNA in lung adenocarcinoma. Human Molecular Genetics *30*, 2263-2271.

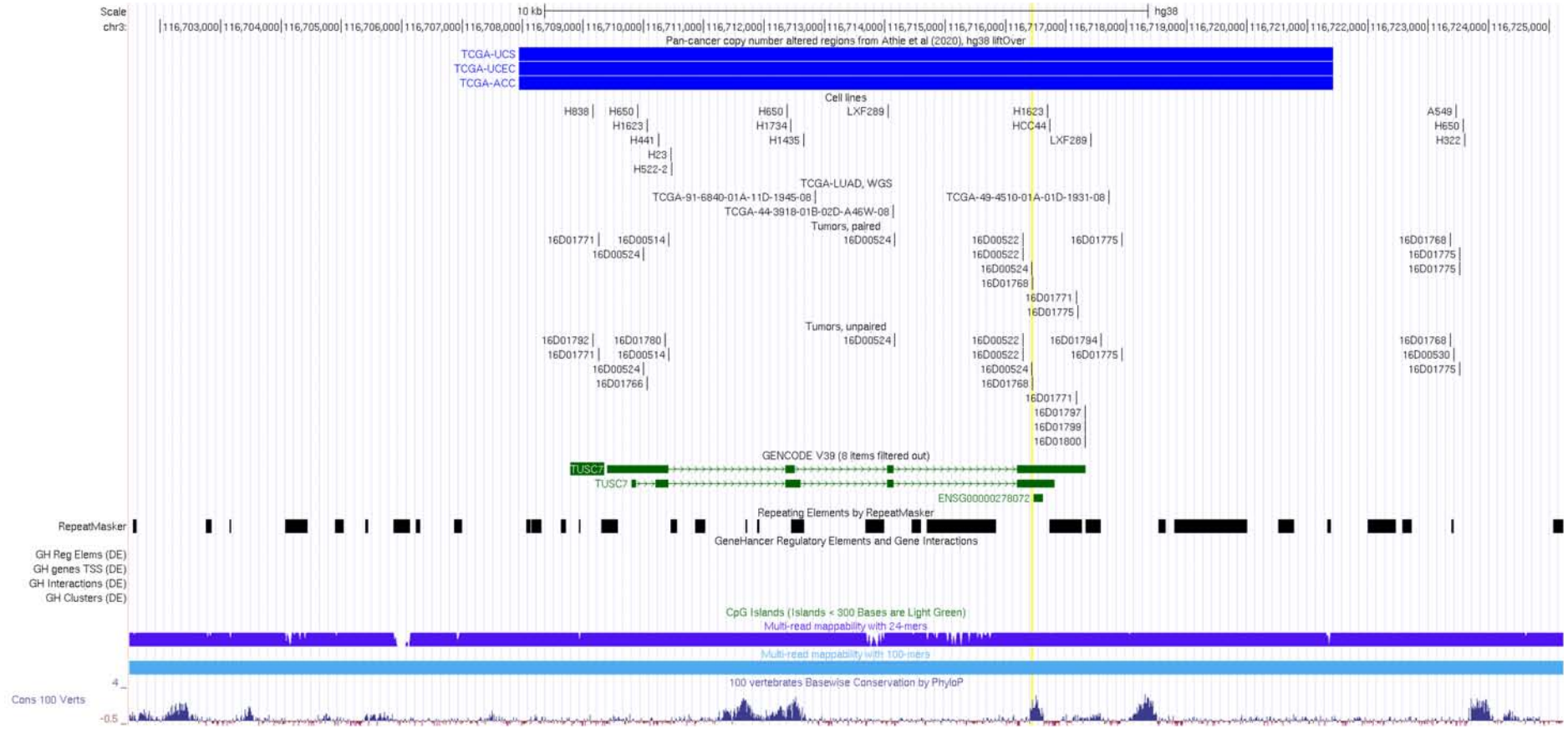
Arenas, A. M.*, Cuadros, M.*, Andrades, A., García, D. J., Coira, I. F., Rodríguez, M. I., Baliñas-Gavira, C., Peinado, P., Álvarez-Pérez, J. C., and Medina, P. P. (2020). LncRNA DLG2-AS1 as a Novel Biomarker in Lung Adenocarcinoma. Cancers *12*, 2080.

Baliñas-Gavira, C., Rodríguez, M. I., Andrades, A., Cuadros, M., Álvarez-Pérez, J. C., Álvarez-Prado, Á., de Yébenes, V. G., Sánchez-Hernández, S., Fernández-Vigo, E., Muñoz, J., *et al.* (2020). Frequent mutations in the amino-terminal domain of BCL7A impair its tumor suppressor role in DLBCL. Leukemia *34*, 2722-2735.

Martin-Padron, J., Boyero, L., Rodriguez, M. I., Andrades, A., Díaz-Cano, I., Peinado, P., Baliñas-Gavira, C., Alvarez-Perez, J. C., Coira, I. F., Fárez-Vidal, M. E., and Medina, P. P. (2020). Plakophilin 1 enhances MYC translation, promoting squamous cell lung cancer. Oncogene *39*, 5479-5493.

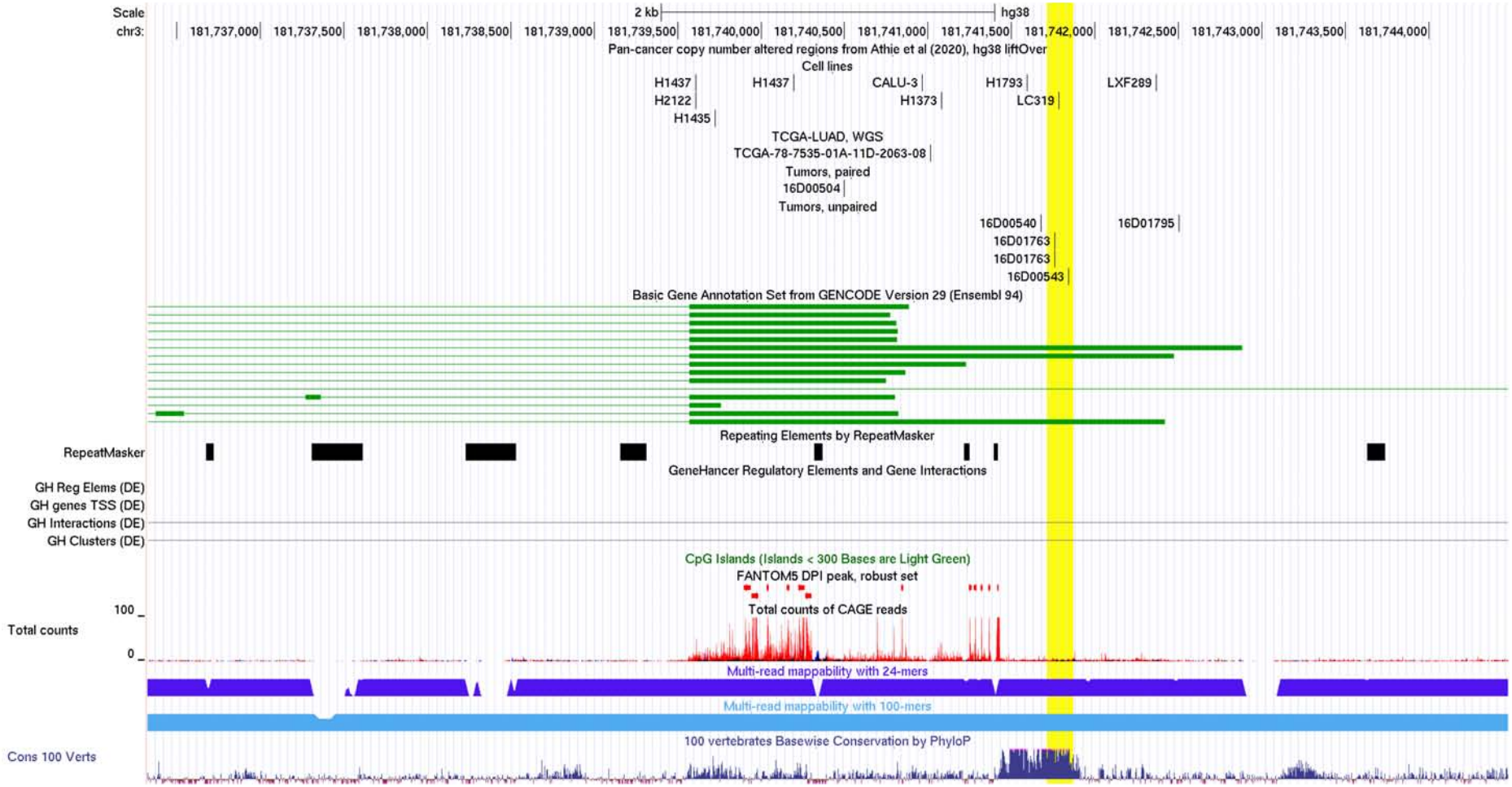
Galindo, I., Gómez-Morales, M., Díaz-Cano, I., Andrades, A., Caba-Molina, M., Miranda-León, M. T., Medina, P. P., Martín-Padron, J., and Fárez-Vidal, M. E. (2020). The value of desmosomal plaque-related markers to distinguish squamous cell carcinoma and adenocarcinoma of the lung. Upsala Journal of Medical Sciences *125*, 19-29.

Supplementary Figures

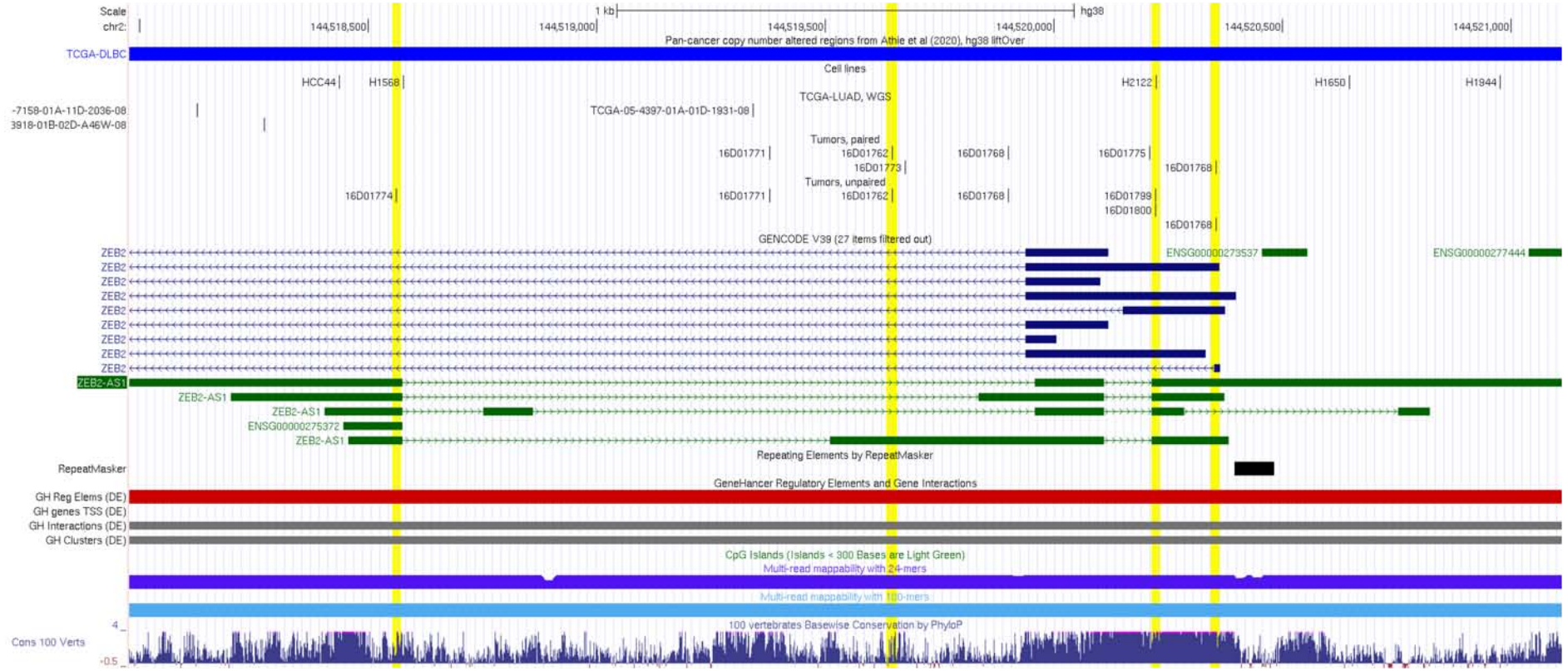


Supplementary Figure 1. Genomic features of the TUSC7 locus. *The position of the variant predicted to have high functional impact is highlighted in yellow.*

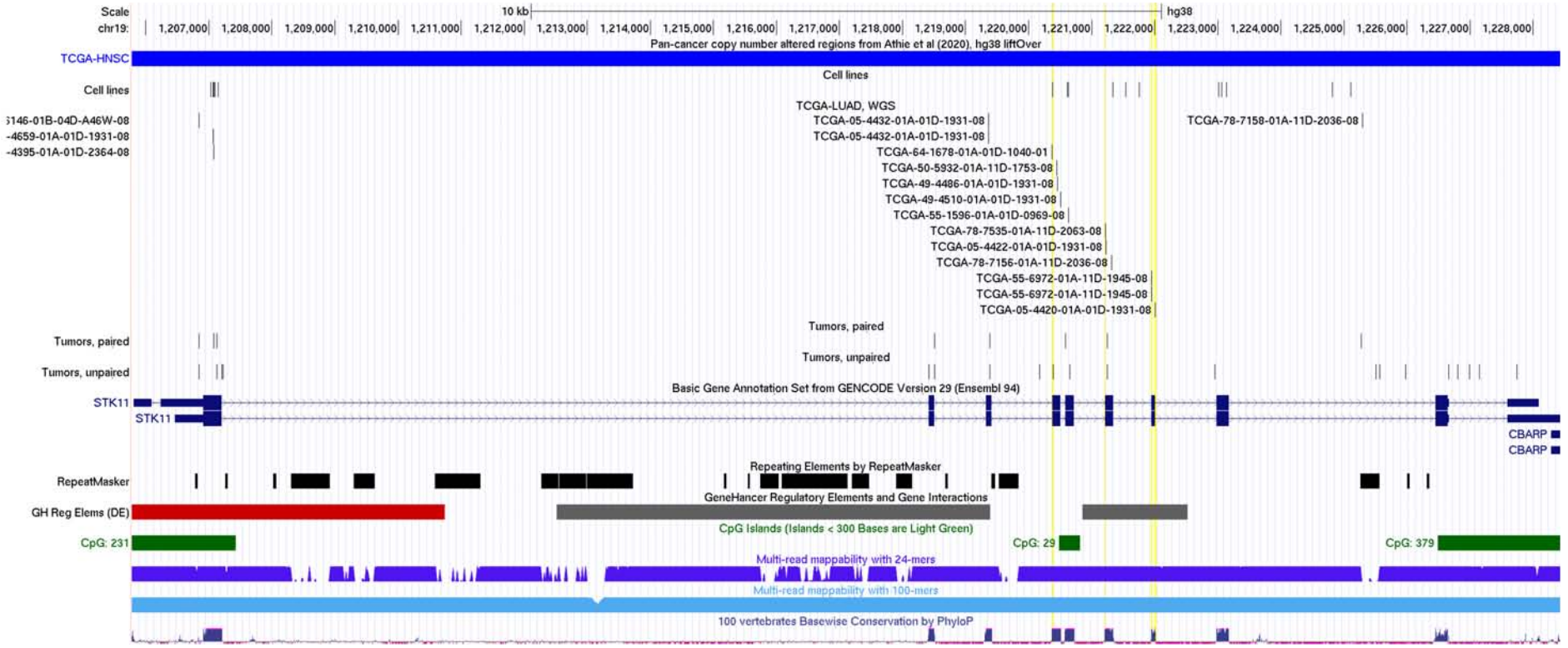
Supplementary Figure 2. Genomic features of the SOX2-OT locus. *The region of the variants predicted to have high functional impact is highlighted in yellow.*



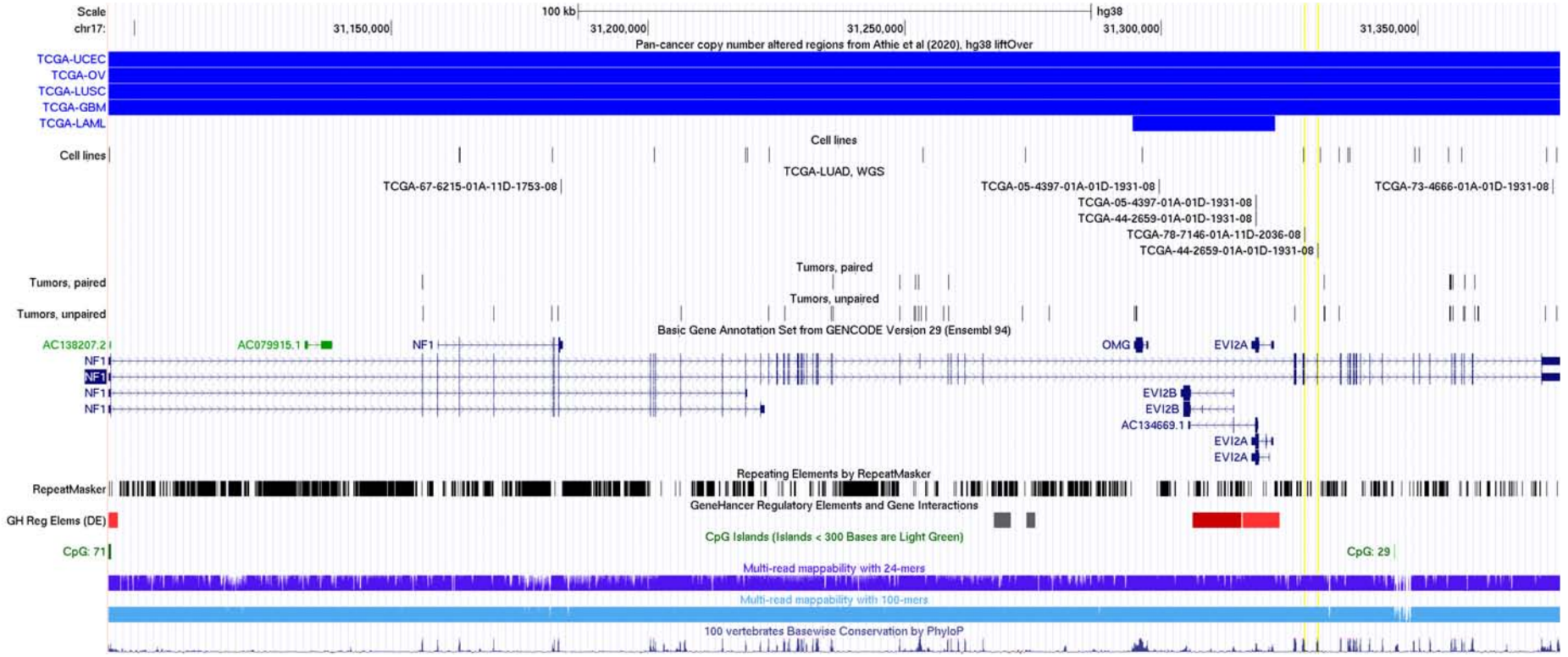
Supplementary Figure 3. Genomic features of the SOX2-OT locus (zoomed in). *The region of the variants predicted to have high functional impact is highlighted in yellow.*



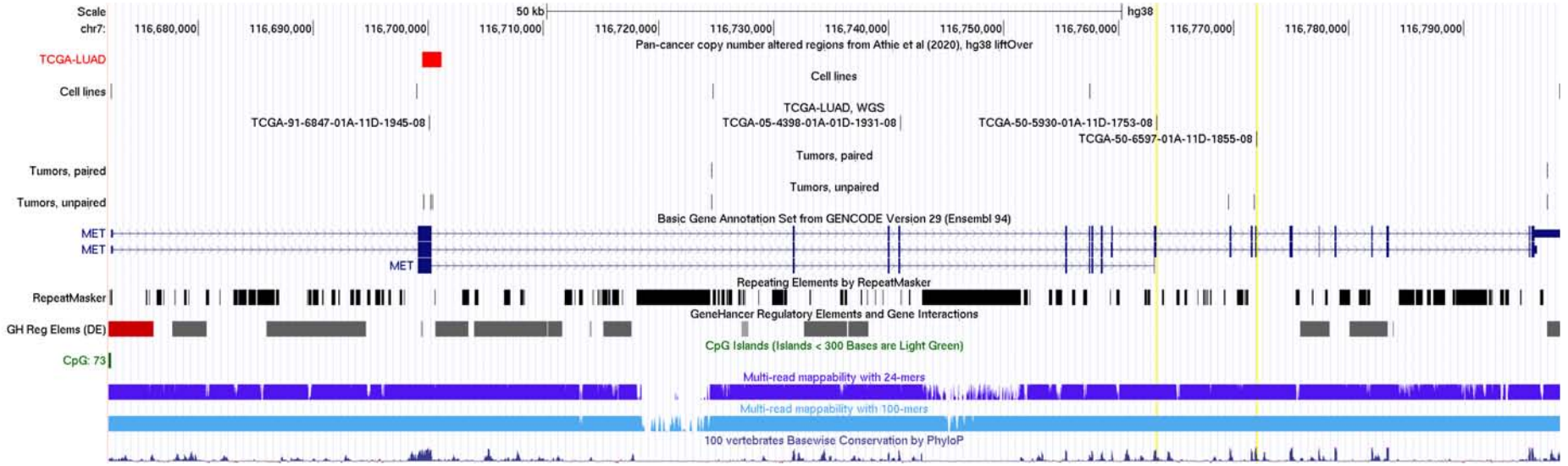
Supplementary Figure 4. Genomic features of the ZEB2-AS1 locus. *The positions of the variants predicted to have high functional impact are highlighted in yellow.*



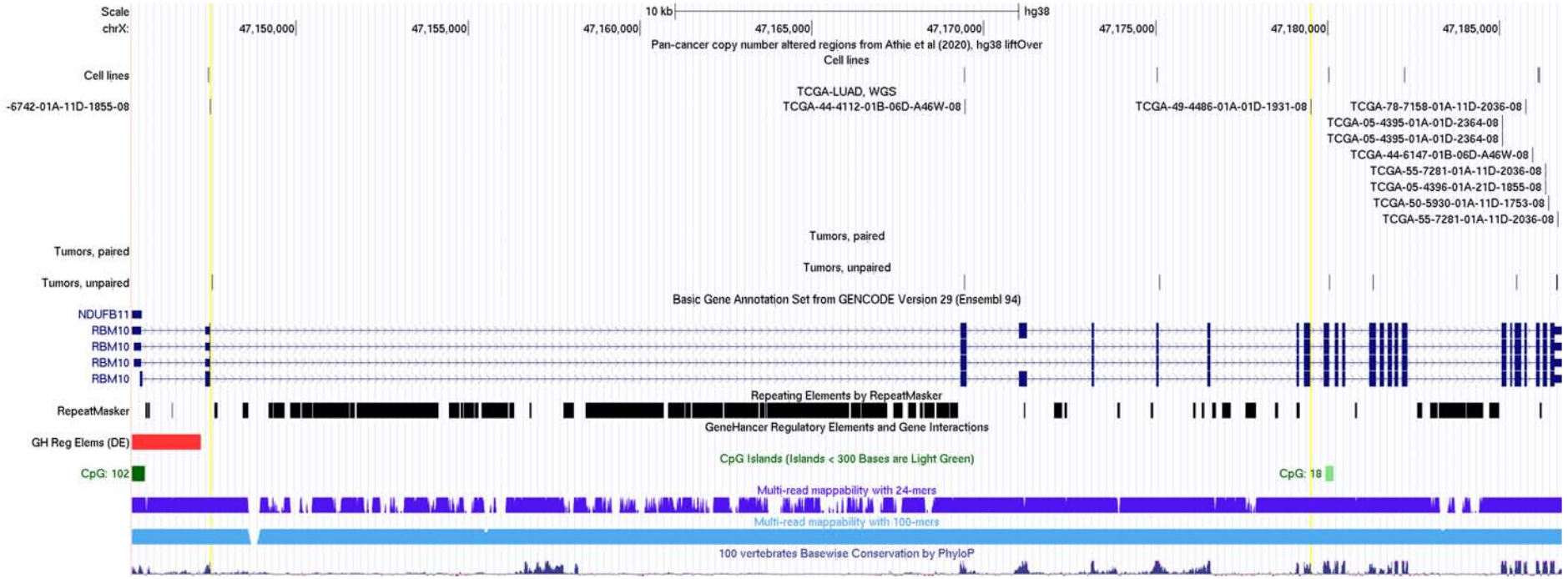
Supplementary Figure 5. Genomic features of the variants in splice regions of STK11. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



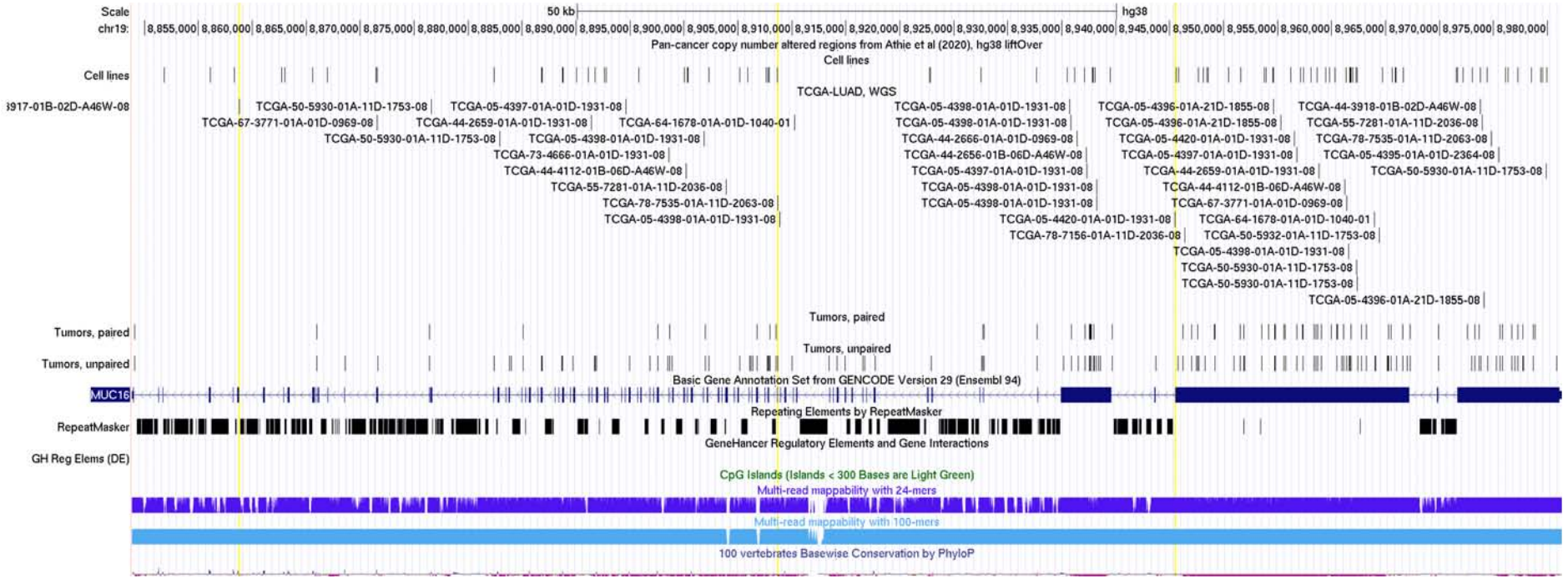
Supplementary Figure 6. Genomic features of the variants in splice regions of NF1. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



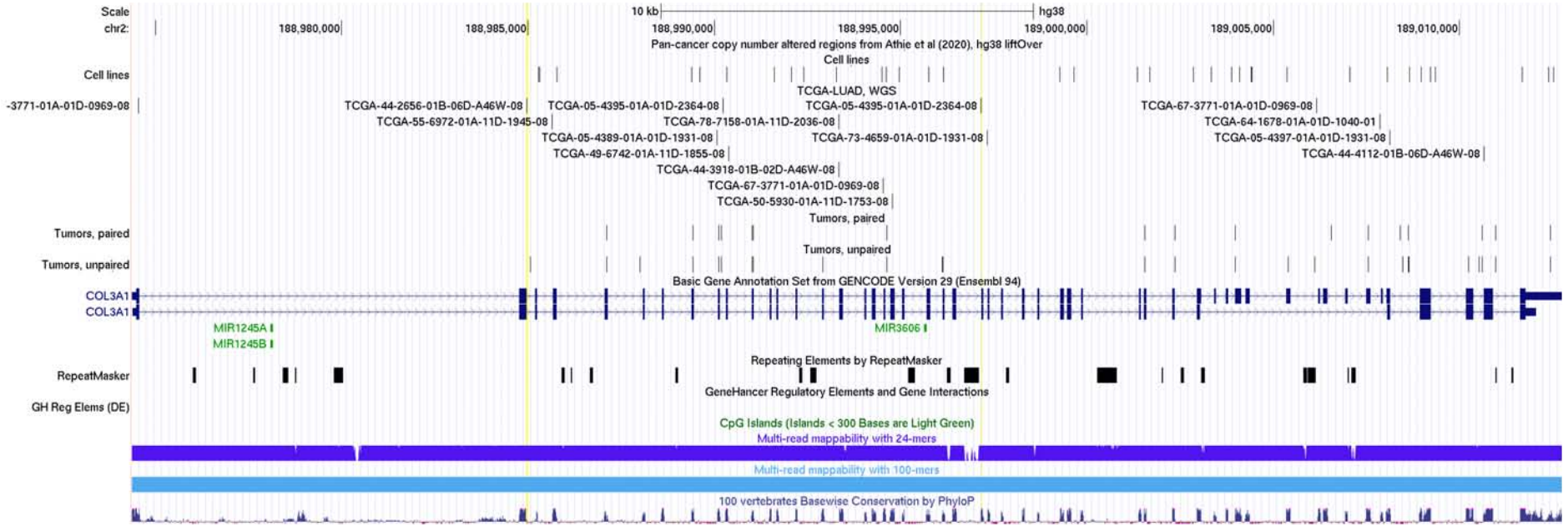
Supplementary Figure 7. Genomic features of the variants in splice regions of MET. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



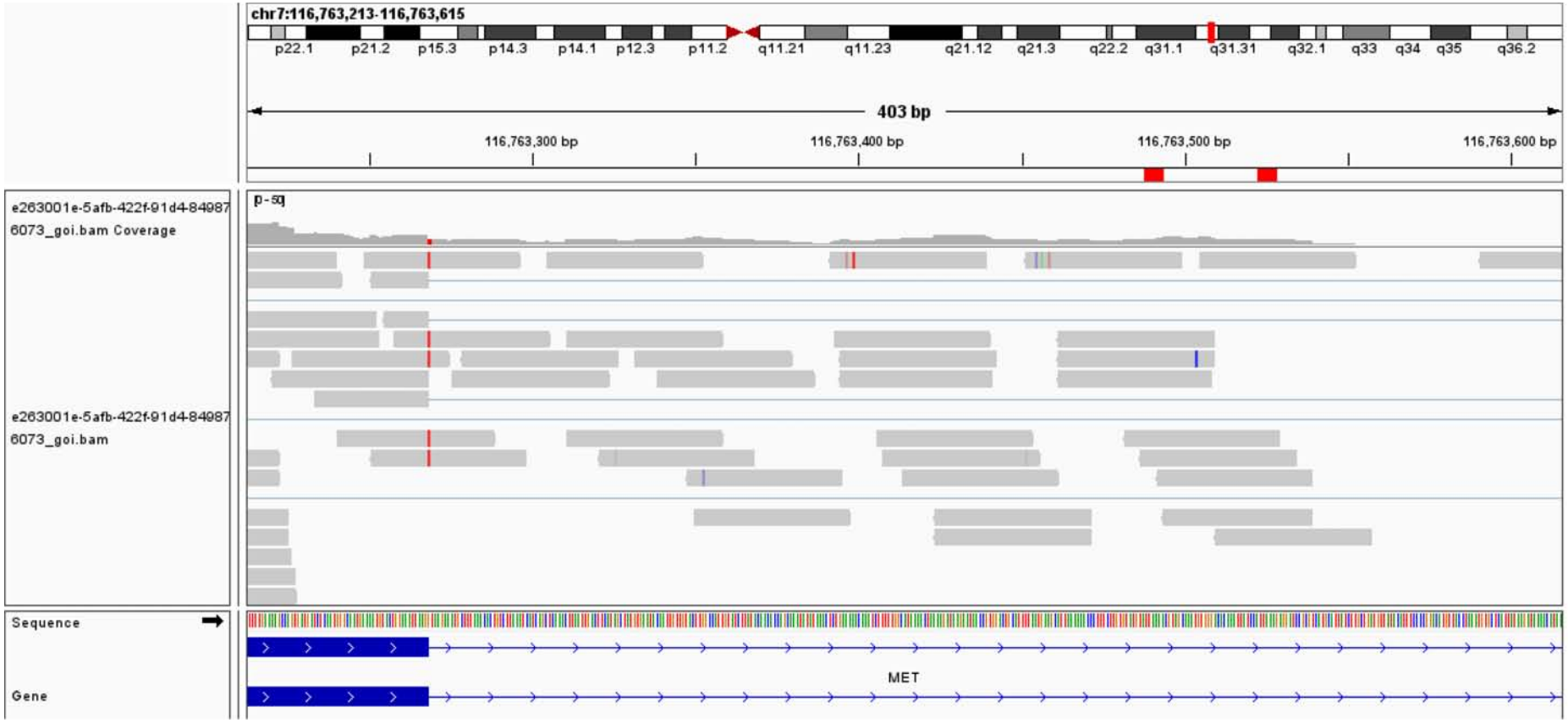
Supplementary Figure 8. Genomic features of the variants in splice regions of RBM10. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



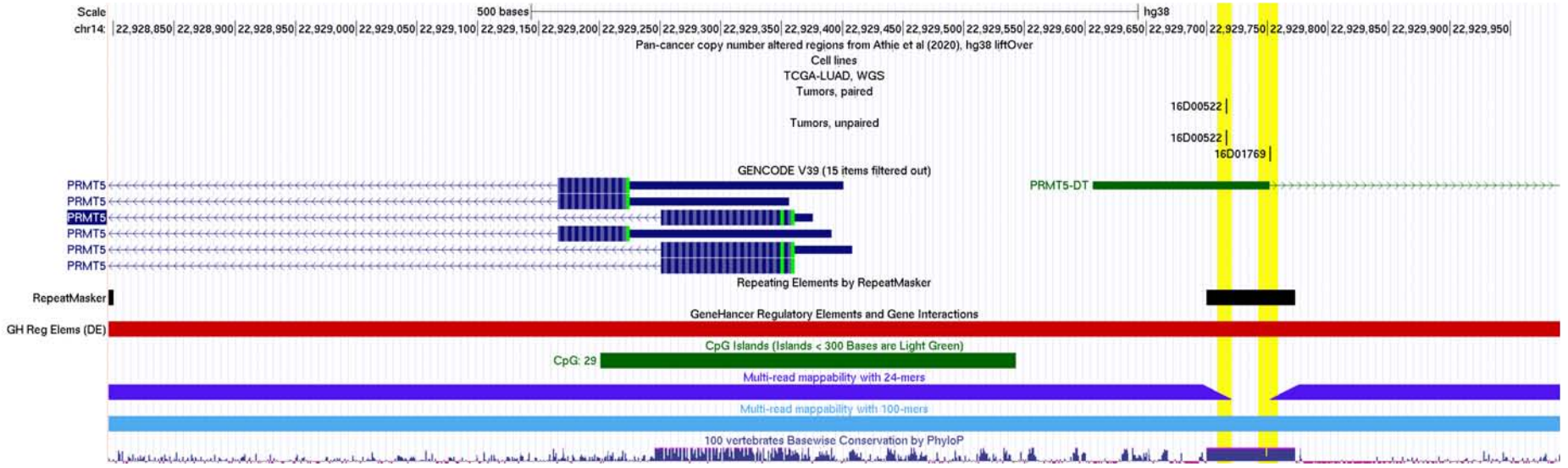
Supplementary Figure 9. Genomic features of the variants in splice regions of MUC16. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



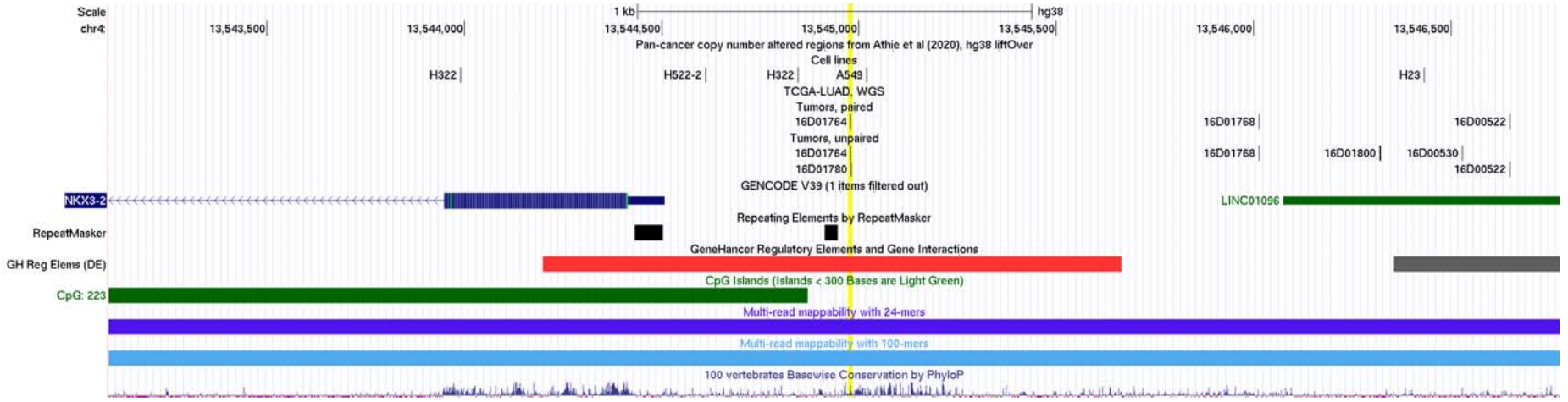
Supplementary Figure 10. Genomic features of the variants in splice regions of COL3A1. *The positions of the variants in splice regions in whole-genome sequencing samples from TCGA-LUAD are highlighted in yellow.*



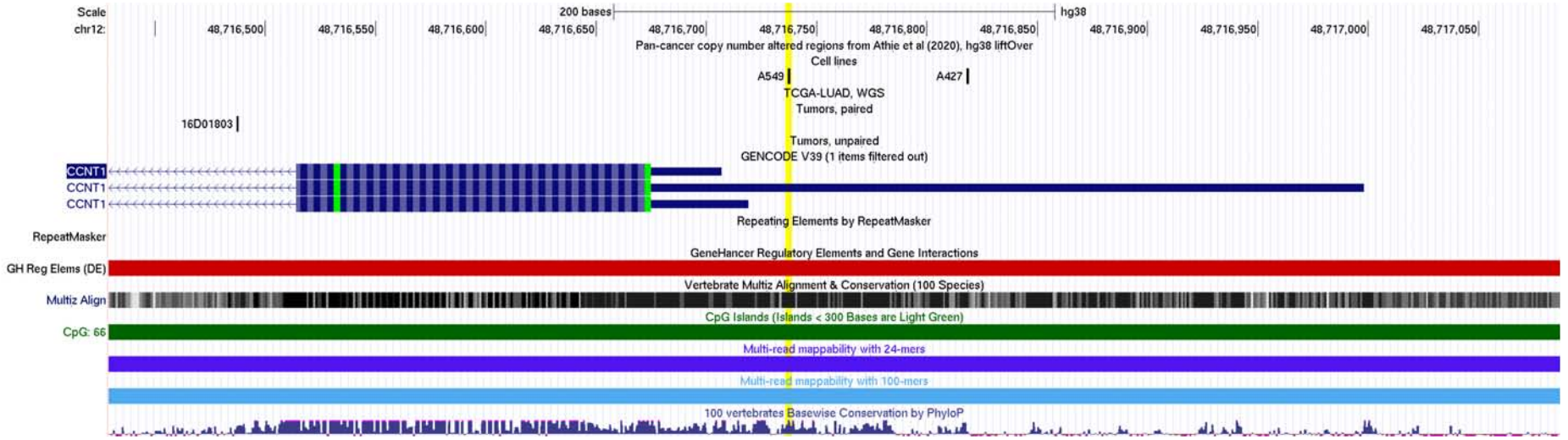
Supplementary Figure 11. Putative intronic polyadenylation in exon 11 of MET. *Reads containing a splice site variant (red rectangles at the splice junction) span the first intronic nucleotides. Coverage remains roughly constant for the first ~300 intronic nucleotides, and then drops to nearly zero. Less than 100 bp before the drop, two AUUAAA polyadenylation signals were found (red rectangles between the genomic coordinate axis and the coverage graph). Snapshot generated using Integrative Genomics Viewer.*



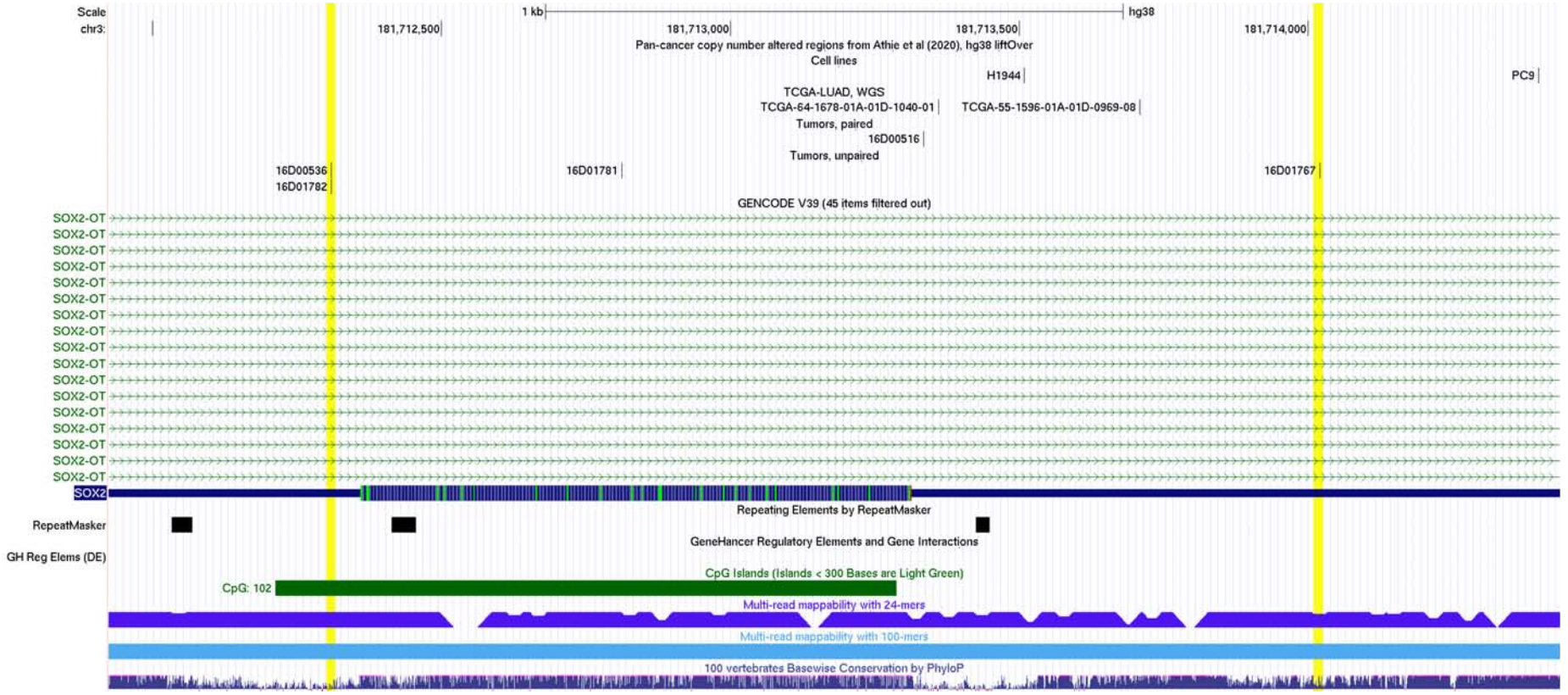
***Supplementary Figure 12. Genomic features of the PRMT5* promoter.** *The positions of the variants predicted to have high functional impact are highlighted in yellow.*



***Supplementary Figure 13. Genomic features of the NKX3-2* promoter.** *The position of the variant predicted to have high functional impact is highlighted in yellow.*



***Supplementary Figure 14. Genomic features of the CCNT1* 5'-UTR.** *The position of the variant predicted to have high functional impact is highlighted in yellow.*



***Supplementary Figure 15. Genomic features of the SOX2* UTRs.** *The positions of the variants predicted to have high functional impact are highlighted in yellow.*

Supplementary Tables

Supplementary Table 1. Features of candidate driver lncRNAs from OncoDriveCLUSTL.

Dataset	lncRNA	Genomic features	Variant features
Cell lines	*GAS6-AS1*	Antisense overlapping *GAS6* PCG. Variants are mostly within a *GAS6* intron.	Low FI. Recurrent variants in a low-mappability position (chr13:113842105).
Tumors, unpaired	*AIRN*	Antisense overlapping *IGF2R* PCG. Two isoforms: (1) has three exons and (2) is a single ~4kb exon fully overlapping (1). Variants affect intron 1 of the (1) isoform.	Low FI.
Tumors, unpaired	*AC087741.1*	Antisense overlapping *CARD14* PCG. Four isoforms, one of which is a single exon. Variants affect a region unique to the single-exon isoform.	Low FI.
Tumors, unpaired	*HEIH*	Antisense to *LINC00847* lncRNA. Two isoforms: (1) does not overlap *LINC00847*, and (2) does. Variants affect a sequence unique to (2).	Low FI.
Tumors, unpaired	*TUSC7*	Intergenic lncRNA. Variants affect exons 1 and 4.	One variant in exon 4 has high FI.
Tumors, unpaired	*AL132780.1*	Antisense (non-overlapping) to *PRMT5* PCG, and overlapping *HAUS4* PCG.	Low FI.

Dataset	lncRNA	Genomic features	Variant features
Tumors, unpaired	*CAMTA1-DT*	Antisense (non-overlapping) to *CAMTA1* PCG.	Low FI.
Tumors, unpaired	*ZFHX2-AS1*	Antisense overlapping *ZFNHX2* PCG. Variants affect a region that is intronic in most *ZFHX2-AS1* isoforms.	Low FI.
Tumors, unpaired	*LINC01391*	Sense (non-overlapping) to *FOXL2* PCG.	Low FI.
Tumors, unpaired	*AC079313.1*	>100 kb. Antisense overlapping multiple PCGs. Two variants, which affect its last exon, which overlaps intron 2 of *GTSF1*.	Low FI.
Tumors, unpaired	*AL135905.1*	Antisense overlapping *PTP4A1* PCG. Single-exon ~1400 bp lncRNA. Variants overlap with an intron of *PTP4A1*.	Moderate-low FI.
Tumors, unpaired	*SOX2-OT*	>800 kb. Sense overlapping *SOX2* PCG. >20 isoforms, inconsistent annotation of 3' end.	Four high-FI variants near 3' end of some isoforms. A MNV may have biased cluster detection.

Dataset	lncRNA	Genomic features	Variant features
Tumors, unpaired	*ZEB2-AS1*	Antisense overlapping *ZEB2* PCG. Variants are scattered along the lncRNA gene, overlapping *ZEB2* intron 1, 5'-UTR, and promoter.	Moderate-high FI for variants overlapping *ZEB2* intron 1, 5'-UTR, and promoter.
Tumors, unpaired	*DHRS4-AS1*	Antisense overlapping *DHRS4* PCG.	Low FI.
Tumors, unpaired	*NEAT1*	Intergenic lncRNA. Variants are scattered along the whole gene.	Moderate-low FI.
Tumors, unpaired	*HOTAIR*	Antisense overlapping *HOXC11* PCG. Variants affect its last exon.	Low FI.
Tumors, unpaired	*EGFR-AS1*	Antisense overlapping *EGFR* PCG. Variants affect introns 19 and 20 of *EGFR*.	Low FI. Localized hypermutation may have biased cluster detection.
TCGA-LUAD	*C5orf66*	>300 kb. Antisense overlapping *PITX1* and *MACROH2A1* PCGs.	Low FI.

Hits for which evidence of high functional impact (FI) was detected are highlighted. FI was assessed by phyloP, CADD, and FATHMM-MKL. p.+unp.: the lncRNA was a hit in both the paired primary tumors and the unpaired primary tumors. PCG: protein coding gene. lncRNA: long non-coding RNA. TSS: transcription start site. MNV: multi-nucleotide variant.

Supplementary Table 2. Seed mutations in miRNAs in our cohorts affecting conserved positions or detected in external cohorts.

Mutation	miRNA	Dataset	Sample	phyloP	FATHMM-MKL	CADD	COSMIC?	TCGA-LUAD?
chr6:52148992_G>T	miR-133b	Tumors, paired	16D00524	7.23	0.99	19.42	Colon (COSN187839)	No
chr17:77089472_A>C	miR-6516-3p	Tumors, unpaired	16D00528	5.93	0.97	21.4	No	No
chr19:10817486_A>C	miR-199a-5p	Tumors, unpaired	16D01798	5.74	0.95	17.30	No	No
chr19:53712230_G>T	miR-518b	Cell lines	H1734	0.32	0.07	2.81	Lung (COSN8606333)	No
chr19:6416449_C>T	miR-3940-3p	Cell lines	A427	0.13	0.15	2.31	Pancreas (COSN7451488)	No

All genomic coordinates use the hg38 reference genome. "phyloP": phyloP 100-way conservation score (higher values mean higher conservation); "FATHMM-MKL": pathogenicity score from FATHMM-MKL (range 0-1, higher scores mean higher pathogenicity); CADD: pathogenicity score from CADD (phred scale, higher values mean higher pathogenicity); "COSMIC?": is the miRNA mutant in independent samples from COSMIC?; "TCGA-LUAD?": is the miRNA mutated in TCGA-LUAD (WES or WGS)?

Supplementary Table 3. Mutations in DROSHA processing motifs in our cohorts.

Dataset	Sample	Mutation	miRNA	Motif (Effect)	Method	phyloP	TPM (Gil.)	RPM (TCGA)
Cell lines	H1734	chr19:51693346_C>G	mir-125a	CNNC (Disrupt)	Both	0.66	10084.54	616.37
Cell lines	H2087	chr19:53787660_G>A	mir-371b	CNNC (Disrupt)	Positional	-1.23	1.25	0.06
Cell lines	H650	chr11:60209153_C>G	mir-6503	UG (Disrupt)	Positional	0.26	1.50	0.44
Cell lines	H1975	chr15:88611851_C>G	mir-7-2	mGHG (88->4.9)	-	3.87	389.72	2.94
Cell lines	H1648	chr7:67114329_G>T	mir-4650-1	mGHG (17->2)	-	0.54	0.38	0.00
Cell lines	H1573	chr7:74191206_C>T	mir-590	mGHG (43->62)	-	1.26	50.27	16.63
Cell lines	H1734	chr7:128207946_G>A	mir-129-1	mGHG (34->23)	-	1.89	2.53	1.73
Cell lines	LC319	chr9:83969839_G>C	mir-7-1	mGHG (65->2.3)	-	9.26	389.72	2.94
Cell lines	H1568	chrX:147272406_C>A	mir-510	mGHG (86->28)	-	-0.76	0.49	0.02
Tumors	16D01787	chr19:53680031_C>G	mir-519e	CNNC (Disrupt)	Both	-0.40	0.43	0.02
Tumors	16D01765	chr14:101049652_C>T	mir-655	CNNC (Disrupt)	Both	0.48	13.72	1.08
Tumors	16D01771	chrX:147259730_C>G	mir-509-3	UG (Disrupt)	Positional	-0.52	4.24	1.73
Tumors	16D01771	chr11:72615129_A>C	mir-139	mGHG (86->6.9)	-	5.68	97.36	40.93

All genomic coordinates use the hg38 reference genome. "Effect": for motifs other than the mismatched GHG (mGHG), whether the mutation creates or disrupts a motif; for mGHG, change in mGHG score. "Method": was the motif predicted by the "positional" method, the "structural" method, or both? "phyloP": phyloP 100-way conservation score for the mutated position. TPM (Gil.): transcripts per million in tumors from Gillette et al's cohort. RPM (TCGA): reads per million in tumors from TCGA-LUAD.

Supplementary Table 4. Mutations in DROSHA processing motifs in TCGA-LUAD.

Dataset	Sample	Mutation	miRNA	Motif (Effect)	Method	phyloP	TPM (Gil.)	RPM (TCGA)
WGS	TCGA-05-4432	chr9:136746931_C>T	mir-6722	UGUG (Disrupt)	Structural	0.30	0.09	0.00
WGS	TCGA-05-4420	chr18:39676783_G>A	mir-5583	mGHG (29->33)	-	-0.94	0.41	0.00
WES	TCGA-91-6848	chr14:101065538_C>A	mir-412	CNNC (Disrupt)	Structural	2.93	4.13	2.03
WES	TCGA-L4-A4E5	chr17:59151127_G>T	mir-301a	CNNC (Disrupt)	Structural	6.93	10.39	12.74
WES	TCGA-55-8616	chr19:53697623_C>CA	mir-525	CNNC (Disrupt)	Both	0.28	1.01	0.09
WES	TCGA-93-8067	chr14:101560306_G>A	mir-1247	CNNC (Disrupt)	Positional	1.14	21.01	10.39
WES	TCGA-86-8585	chr19:53686504_T>A	mir-519c	UGUG (Disrupt)	Both	0.27	0.41	0.05
WES	TCGA-55-8089	chr19:53716633_G>T	mir-521-2	UGUG (Disrupt)	Both	-0.63	0.31	0.07
WES	TCGA-55-7907	chr19:53787707_C>G	mir-371a	UGUG (Create)	Structural	-0.82	0.34	0.20
WES	TCGA-MP-A4TF	chr4:20528325_C>T	mir-218-1	UGUG (Create)	Positional	2.27	356.05	22.26
WES	TCGA-55-A490	chr19:53690881_C>G	mir-520a	UG (Create)	Structural	-0.67	0.39	0.16
WES	TCGA-78-7220	chrX:145994352_C>A	mir-890	UG (Create)	Structural	-0.74	0.59	0.01
WES	TCGA-17-Z014	chr10:54607881_G>T	mir-548f-1	mGHG (39->8.1)	-	0.33	0.48	0.03
WES	TCGA-17-Z059	chr11:64341909_A>C	mir-7155	mGHG (49->12)	-	0.30	0.92	0.22
WES	TCGA-95-7039	chr11:122099766_C>T	mir-125b-1	mGHG (46->32)	-	6.91	8778.97	239.52
WES	TCGA-55-8506	chr14:101044275_C>A	mir-1185-2	mGHG (51->17)	-	-0.87	1.89	0.14

Dataset	Sample	Mutation	miRNA	Motif (Effect)	Method	phyloP	TPM (Gil.)	RPM (TCGA)
WES	TCGA-55-8089	chr14:101065372_G>T	mir-409	mGHG (18->2.4)	-	2.17	155.50	13.83
WES	TCGA-97-7937	chr15:31065056_T>A	mir-211	mGHG (8->1.2)	-	5.84	1.45	0.18
WES	TCGA-44-6777	chr19:53713355_T>A	mir-519d	mGHG (75->37)	-	0.28	0.42	0.06
WES	TCGA-44-2656	chr19:53716672_C>A	mir-521-2	mGHG (75->42)	-	-1.95	0.31	0.07
WES	TCGA-55-A48X	chr3:168551937_G>T	mir-551b	mGHG (59->10)	-	5.93	1.17	2.46
WES	TCGA-55-8094	chr7:128207871_T>A	mir-129-1	mGHG (34->3)	-	0.31	2.53	1.73
WES	TCGA-44-A4SU	chrX:134170272_G>A	mir-106a	mGHG (11->17)	-	4.59	189.02	10.31
WES	TCGA-44-6779	chrX:146027800_G>A	mir-891a	mGHG (31->41)	-	-0.39	1.43	3.11
WES	TCGA-17-Z010	chrX:152392226_C>A	mir-105-1	mGHG (22->4.5)	-	0.95	3.38	2.18

WES: whole exome sequencing. WGS: whole genome sequencing. All genomic coordinates use the hg38 reference genome. "Effect": for motifs other than the mismatched GHG (mGHG), whether the mutation creates or disrupts a motif; for mGHG, change in mGHG score. "Method": was the motif predicted by the "positional" method, the "structural" method, or both? "phyloP": phyloP 100-way conservation score for the mutated position. TPM (Gil.): transcripts per million in tumors from Gillette et al's cohort. RPM (TCGA): reads per million in tumors from TCGA-LUAD