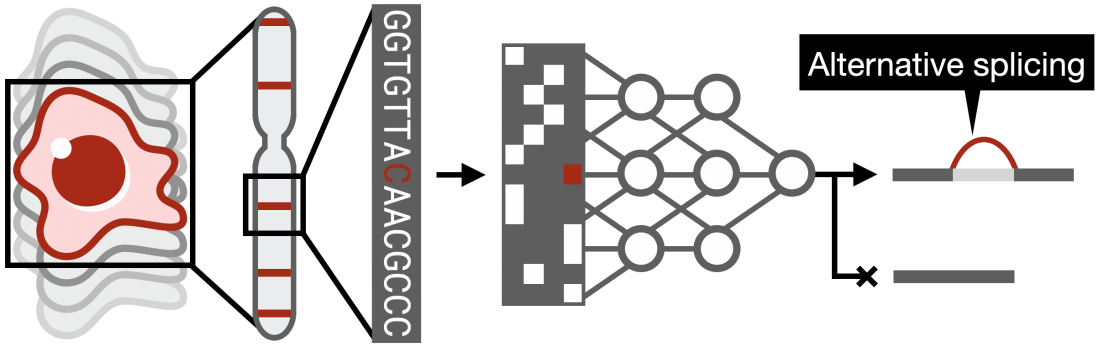

Systems biology of deregulated splicing in cancer

A pan-cancer analysis of dysfunctional splicing machinery and alternative splicing events

Master's thesis in Biotechnology

LETICIA CASTILLON

DEPARTMENT OF BIOLOGY AND BIOLOGICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

Systems biology of deregulated splicing in cancer

A pan-cancer analysis of dysfunctional splicing machinery and alternative splicing events

LETICIA CASTILLON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Biology and Biological Engineering
Division of Systems and Synthetic Biology
Nielsen Lab
Chalmers University of Technology
Gothenburg, Sweden 2021

Systems biology of deregulated splicing in cancer
A pan-cancer analysis of dysfunctional splicing machinery and alternative splicing events
LETICIA CASTILLON

Supervisor: Angelo Limeta, Department of Biology and Biological Engineering
Examiner: Christer Larsson, Department of Biology and Biological Engineering

Master's Thesis 2021
Department of Biology and Biological Engineering
Division of Systems and Synthetic Biology
Nielsen Lab
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Graphical workflow for the use of a deep learning algorithm to predict the splice-altering variants in cancer.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Systems biology of deregulated splicing in cancer
A pan-cancer analysis of dysfunctional splicing machinery and alternative splicing events
LETICIA CASTILLON
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

The deregulation or disruption of the splicing process has been shown to play a role in the onset, development, and even response to treatment of some malignancies. Partly due to our incomplete understanding of the mechanism and regulation behind splicing and alternative splicing, the importance of aberrant splicing in oncogenesis is not yet understood. In this project, we set out to perform a systematic analysis of aberrant splicing events in cancer cell lines from two perspectives: deregulated splicing because of dysfunctional splicing factors and the appearance of de novo splicing events because of splice-disruption mutations in the DNA sequence.
All the analyses have been performed using data publicly available in the DepMap portal. For the analysis of de novo splicing events caused by mutations in the DNA sequence, we used the deep-learning tool SpliceAI. To the extent of our knowledge, SpliceAI has not been previously used to analyze RNA-sequencing from cancerous samples - cell lines nor human tumour samples. In this project, we decided to test the deep learning algorithm in data from the Cancer Cell Line Encyclopedia collected by the DepMap consortium and evaluate its performance to detect splicing alterations that may affect oncogenesis. In order to try to evaluate the clinical relevance of our findings, we used the MSK-IMPACT sequencing panel to narrow down the analysis to actionable genes - genes that can be targeted by drugs.

Keywords: multi-omics, splicing, alternative splicing, deep learning, cancer, transcriptome.

Acknowledgements

Now that I am almost done with this MSc thesis, it is time to look back at the journey. I would like to say thank you first and foremost to Angelo, for allowing me to participate in this project and all the help and incredibly useful guidance. I also want to thank Christer Larsson for agreeing to be my examiner; Jens Nielsen, for letting me carry out this work in his lab at Chalmers; and both Jens and Francesco Gatto for the valuable feedback they have provided during the project.

Working in Sysbio has been an incredible experience thanks to the very nice people that are part of the Department. So, a big thank you to all the students, researchers and administrative stuff that make this place such a friendly and pleasant environment to be a part of.

I would also like to take a moment here to acknowledge the Chalmers Centre for Computational Science and the Swedish National Infrastructure for Computing for providing the access to Hebbe, the computational cluster that has made possible some the computationally demanding analysis in this project.

Finally, a shoutout to all my friends, in Sweden, Spain, and other parts of the world, that have patiently listened to me explaining what this project was about – even when they did not always ask for it. And to my family, that always supports me no matter what. Love you.

Leticia Castillon, Gothenburg, June 2021

Contents

List of Figures

List of Tables

Acronyms & abbreviations

ICI	Immune Checkpoin Inhibition
WHO	World Health Organization
HGP	Human Genome Project
PCR	Polymerase Chain Reaction
NGS	Next Generation Sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
EGFR	Epidermal Growth Factor Receptor
TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
PARP1	poly (ADP-ribose) polymerase-1
ER	estrogen receptors
PR	Progesterone receptors
HER2	Human Epidermal growth factor Receptor 2
shRNA	small hairpin RNA
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DepMap	The Cancer Dependency Map
CCLE	The Cancer Cell Line Encyclopedia
RNAi	RNA interference
PRISM	Profiling Relative Inhibition Simultaneously in Mixtures
CTPR	Cancer Therapeutics Response Portal
AI	Artificial Intelligence
TCGA	The Cancer Genome Atlas
MHC	Major Histocompatibility Complex
ASDs	Autism Spectrum Disorders
MSK-IMPACT	Integrated Mutation Profiling of Actionable Cancer Targets
MAF	Mutation Annotation Format
VCF	Variant Calling Format
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing
TPM	Transcript Per Million
snRNP	small nuclear ribonucleoproteins
GSA	Gene Set Analysis
NMD	Nonsense mediated decay
SNP	Single Nucleotide Polymorphisms
PCA	Principal Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
GSA	Gene Set Analysis
TPR	True Positive Rate
FPR	False Positive Rate
ROC curve	Receiver Operating Characteristic curve
PR curve	Precision-Recall curve
AUC	Area Under the Curve
FDR	False Discovery Rate

1
Introduction

Cancer is arguably one of the most widely known and feared diseases of our time. In spite of the mammoth efforts - economical, scientific, and medical - that have been invested in finding a cure, an efficient treatment for all instances of cancer remains elusive. Although this may seem puzzling, the truth is that cancer is a complex disease about which we knew practically nothing until the last century.

Cancer converges to an array of distinguishing characteristics - namely cells that are able to grow uncontrollably, to adapt to their environment better than normal cells and eventually, to spread and invade other tissues. However, it would be wrong to consider cancer as a single entity, since the molecular mechanisms driving each cancer are exceedingly unique and influence factors such as the aggressivity of the tumor, the response to treatment or the likelihood of relapse. Moreover, this shape-shifting disease displays a twisted mirror of our own biology, usually being caused by small variations in the biology of the cell that disrupt the delicate mechanisms that maintain the homeostasis of the body. Yet the changes are so small that it is exceedingly hard to find a treatment that obliterates cancerous cells while leaving healthy ones undisturbed.
Cancer has shown to be an insidious disease that takes multiple forms and hitchhikes our own biology to its advantage, eluding the efforts to completely eradicate it for over a century. Its insidiousness and stubbornness – relapsing often after oncologists believed they had ridden the body from the disease – have caused some researchers to coin it as the emperor of all maladies [113].
In order to understand where we stand in our battle against cancer, it is important to define the term itself.

1.1 Defining cancer

The term cancer comprises many diverse and complex neoplastic diseases [63] that nevertheless share several characteristics, such as uncontrolled cell growth and the ability to invade - and eventually destroy - adjacent tissues or organs [31], [48]. Cancer is caused by the accumulation of mutations and/or epigenetic changes in the DNA that lead to alterations in the biology of the cell [1]. Therefore, cancer is caused by the interaction of both internal factors, such as mutations, hormones, or immune

[1]Some cancers can be caused by viruses, as first demonstrated by Peyton Rous [134]. Human papillomavirus (HPV) is a prominent example of such a virus, being the leading worldwide cause of cervical cancer [36]

conditions, and environmental/acquired factors, namely tobacco, radiation or diet [3]. However, only 5-10% of all cancers are due to inherited genetic characteristics, whereas the rest can be attributed to mutations linked to environmental factors [3].

Before addressing the molecular characteristics of cancer, it is good to clarify some of the jargon. The terms tumor and cancer are usually used indistinctly, but they do not necessarily denote the same objects. A tumor is a mass of cells that start growing uncontrollably. A tumor may, however, be or not be cancerous, depending on its ability to rapidly grow and invade other tissues. A benign tumor, although still able to grow, is usually encapsulated and lacks the ability to invade surrounding tissue - although it may still cause problems depending on its size and location. On the other hand, a tumor is malignant or cancerous if it is invasive, meaning that it can spread to adjacent tissues, or if the tumor can metastasize, meaning it spreads to other parts of the body [1]. Therefore, not all tumors are cancers, and not all cancers form tumors - for example leukemias, most lymphomas, and some types of myelomas, which are cancers of the blood [1]. For the sake of clarity, when using the word tumor in this work I will be referring to malignant tumors unless the contrary is stated.

1.2 A brief history of cancer

Figure 1.1: The most ancient description of a neoplasm was found in a papyrus coming from ancient Egypt, dating back to 2625 BC. The medical text described what may have been a breast tumor.

Cancer is, in fact, an instance of evolution. Following an alteration that disrupts the delicate balance of cellular proliferation and death, cancer usually arises because a cell is able to grow faster and adapt better than its neighbors [113]. Therefore, cancer is inherent to human biology and it becomes more prevalent as we age, since the molecular mechanisms that prevent mutations to set in and protect cellular homeostasis begin to fail. For this reason, even though there is evidence that cancer has always accompanied the human race - Imhotep, the famous Egyptian physician who lived around 2625 BC, already described a case of breast cancer; and cancerous tissues have been found in mummies, the most ancient of these dating of 2000

BC [113] - the disease became more evident as advances in medicine furthered our life expectancy and plummeted the prevalence of previously deadly diseases such as tuberculosis or smallpox. In the 1940s, cancer made its way to the second most common cause of death in the US, only surpassed by heart disease [113]. In the beginning of the century, cancer was listed seventh on that list [113].

Although cancer was present in the early days of human history, its prevalence was indeed rare and therefore the disease remained nameless until - approximately - 400 BC, when Hippocrates gave it the name karkinos, coming from the Greek word for "crab" [113]. Illness was explained by ancient Greeks as an imbalance between the four humors that formed the body: blood, black bile, yellow bile and phlegm. Following this philosophy, cancer - back then only diagnosable when already forming noticeable tumors - was explained as an excess of "trapped" black bile. Therefore, the first treatment for cancer came in the form of medicines that attempted to cleanse the excess of black bile [113]. However, in the seventeenth century it became clear that the humoral theory of disease did not hold, hence opening the door for a different course of treatment for cancer: the extraction of the tumor [91]. It was the dawn of surgical oncology, propelled forward by the discoveries of antisepsis and anesthesia - before of which it had been only possible to extract very small and superficial tumors without an extreme risk for the patient [113]. In the twentieth century, many primary tumors - not yet metastasized - could be removed by surgery. As of today, surgical extraction still remains the standard treatment of localized tumors, often combined with adjuvant therapy [16], [149]. However, in spite of surgical efforts to remove every bit of malignant tissue, some cancers kept coming back. Although it was impossible to know back then, some of the tumors surgeons were excising had already spread to other parts of the organism - the cancers had metastasized.

Later, the discovery of X-rays unveiled yet another form of treatment. X-rays can directly attack the DNA or produce toxic chemicals that are harmful to the molecule. In response to this damage, cells usually stop dividing or even die - especially those cells that divide quickly [113]. Soon enough, doctors and researchers realized the potential of the new technology for treating cancer. In 1826 X-ray irradiation was used for the first time to treat cancer; it was the birth of radiation oncology. Unfortunately, this treatment presented similar shortcomings than surgery, and was of limited utility on tumours that had already metastasized. Moreover, as we well know nowadays, radiation produces cancers by the same mechanism that destroys the tumors: the DNA damage can induce cancer-driving mutations in genes.
It became evident that locally attacking a tumor was not sufficient to cure cancer. The surgeon Willy Meyer put this realization into words in 1932: "If a biological systemic after-treatment were added in every instance, we believe the majority of such patients would remain cured after a properly conducted radical operation [2]" [113]. Oncologists and researchers came to the conclusion that cancer was, after

[2] A radical operation was a surgery in which they not only excised the tumor, but also surrounding tissue, muscle and sometimes even lymphatic nodes in order to ensure that no cancerous cells remained - in an attempt to avoid cancer relapse.

all, a systemic disease. Therefore, the efforts to find a cure took a new direction: finding a drug that would be potent enough to kill malignant cells, yet specific enough to spare healthy ones. This concept was known as chemotherapy and was not restricted to cancer, but also to other diseases caused by microorganisms. Paul Ehlrich, considered one the father of chemotherapy, found chemicals able to cure the sleeping sickness and syphilis - microbial diseases - and called them magic bullets [113]. In spite of his efforts, he could not find such a chemical to treat cancer: cancerous cells were too similar to healthy ones, and hence difficult to target.

Later attempts were more successful - Sidney Farber's antifolates (1946) and Hitchings and Elion's 6-mercaptopurine (tried on patients in 1950) [113] - but only temporarily. Similarly to surgery and radiotherapy, cancers kept relapsing. Combinations of drugs - cocktails of up to four cytotoxic drugs were tried in the 60s - proved to be more prosperous, but to a great cost for the patients. Eradicating the tumour without killing the patient while doing it was the largest challenge of oncology - and in many cases it still is. Even in the most successful clinical trials, only a small percentage of patients both survived the treatment and remained cancer free for more than one year [113]. The first targeted drug for cancer was tamoxifen, an estrogen antagonist that was first used to treat breast cancer in 1969. The trial was a success - in a proportion of the patients, the tumors almost immediately shrank, although they would eventually relapse - but more importantly, it was an important proof-of-concept: it was possible to target a specific pathway in a cancer cell. The molecular mechanism behind the action of the drug was deciphered in 1973, being the first time a molecular logic could be laid for a drug, its target and a cancer cell [113]. It was also the first hormonal therapy used to treat cancer - practically before the function of hormones in the body was completely understood. Notwithstanding the success, the problem persisted that cancerous cells eventually become resistant to both chemotherapy and hormonal treatment - even if only after years or decades. It did highlight, however, the heterogeneity of cancer. At a genetic level, only a certain type of breast cancer, which still possessed estrogen receptors, was responsive to treatment with tamoxifen. At an anatomic level, some cancers remained local while others had a tendency to metastasize [113].

Besides the eventual development of resistance, the major setback of cancer treatments is that we lack a therapeutic strategy incisive enough to distinguish between cancerous and healthy cells, with current chemotherapy and radiation treatments effectively poisoning malignant and normal cells alike. In 1997, it was observed that cancers could present tumour-specific antigens - generated by point mutations, for example - able to elicit an immune response from the patient [22]. In some cases, such an immunological response could eliminate the tumor. With this realization the development of a new treatment strategy began: immunotherapy, which aims to harness the immune system in order to directly treat cancer [162], while sparing the rest of the organism, since many of the targets used for immunotherapy are tumour specific. Since the early days of immunotherapy, several strategies have appeared: vaccines against cancer, cellular immunotherapies, antibodies or recombinant proteins [47]; with immune checkpoint inhibition (ICI) therapy being one of the most promising advances in cancer treatment [128] in the recent years. However, in spite of the clinical efficacy of approaches such as ICI, most tumours are virtually un-

responsive to this approach (i.e. pancreatic cancer, colorectal tumors and prostate tumors) and even in responsive cancers, relapses usually occur [37].
In spite of the tremendous research efforts, it is evident that an efficient and durable treatment for cancer remains elusive, with most cancers eventually able to become resistant to therapy. Cancer is still the second cause of death worldwide, with an incidence of 442.4 per 100,000 people per year [15] and mortality of 158.3 per 100,000 people per year in the US [17]. Besides the human cost, cancer also supposes a high economic burden, costing around $150.8 billion in the US in 2018 [15]. Worldwide statistics indicate that by 2040, the yearly number of new cancer cases will rise to 29.5 million and the associated deaths will also increase to 16.4 million [15], [55].

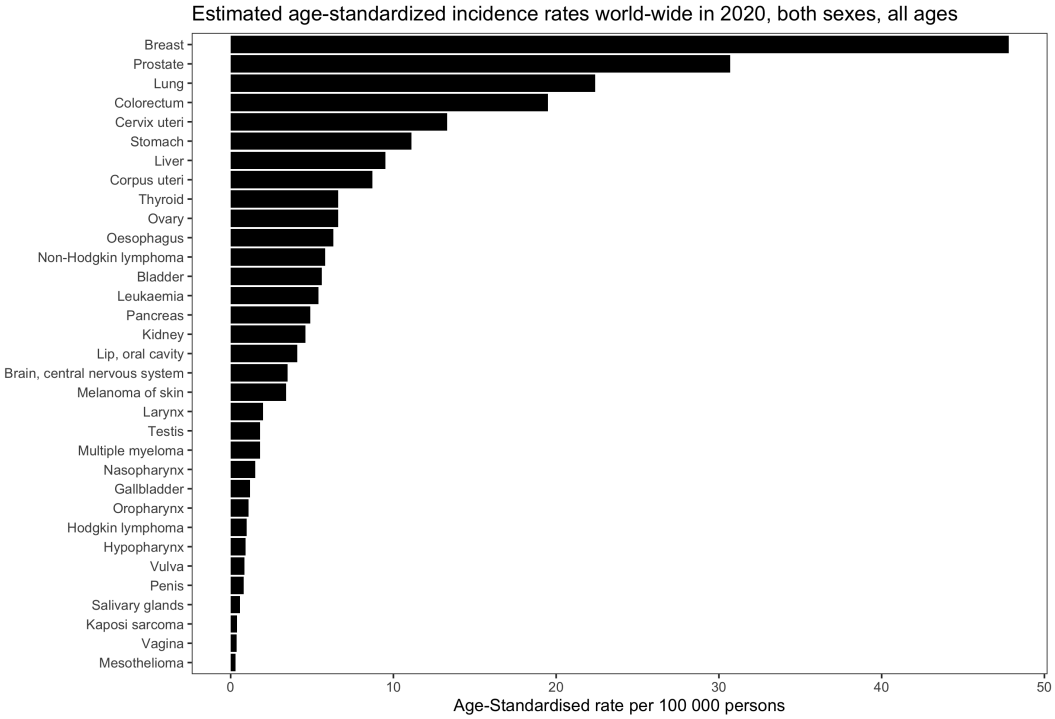

Figure 1.2: Age-standardized incidence rates world-wide in 2020, across different cancer types. Data source: GLOBOCAN 2020 & WHO; International Agency for Research on Cancer 2021.

In order to be able to envision efficient treatments, it seems clear that we need to better understand the molecular mechanisms by which cells become cancerous and the microevolutionary process occurring in the tumor that leads to the development of resistance to chemotherapy or immunotherapy.

1.3 How do cells become cancerous?

As stated before, cancer is a process of microevolution: a cell will acquire a mutation or epigenetic change that will give it a slight advantage over neighboring cells, making it more suited for survival. Therefore, several rounds of genetic changes and natural selection lead to tumor progression [1]. The consensus is that most cancers

arise from a single cell, which is known as the monoclonal origin of tumors. However, there is evidence supporting a different theory, namely the multiclonal origin of tumors that states that cancers arise from the interaction of many different clones of cells [120].

To better understand the diversity and complexity of the disease, several molecular events have been defined as the hallmarks of cancer. These hallmarks can be thought of as the capabilities that a cell needs to acquire to become malignant, regardless of the molecular mechanism behind said capability [62]. In other words, the hallmarks explain how tumoral cells can thrive and proliferate uncontrolled in the organism, for example escaping immunological surveillance [62]. This does not mean that every single tumor will check all of the hallmarks, but rather provides a framework to understand the effects of the different changes a cell suffers when undergoing a malignant transformation.

The hallmarks of cancer have been defined such as follows:

- *Ability to alter energy metabolism*: since cancer cells need to sustain an increased rate of growth, they have increased energy needs. To sustain it, cancerous cells can limit their energy metabolism to glycolysis, even in the presence of oxygen, which is known as the Warburg effect [1], [62].
- *Genomic instability*: healthy cells have mechanisms in place to keep the mutation rate low. In cancerous cells, these mechanisms are typically impaired, facilitating the accumulation of random mutations over time [1], [62].
- *Ability to avoid replicative cell senescence*: normal cells can only divide a limited number of times, which is determined by the length of the telomeres, which shorten with every cell division. By overexpressing the protein telomerase, cancer cells avoid the shortening of the telomeres and therefore the programmed senescence [1], [62].
- *Inducing angiogenesis*: tumors need to create new blood vessels to be able to uptake oxygen and nutrients and remove metabolic byproducts and carbon dioxide [62].
- *Activating invasion and metastasis*: cancer cells are capable of breaking the constraints that keep other cells in the place where they are supposed to be. The process of metastasis is not completely understood yet but involves a process that starts with the local invasion of adjacent tissue, followed by intravasation to the blood and lymphatic vessels and the transport of cancer cells to distant tissues [1], [62].
- *Ability to evade immune destruction*: usually the immune system is able to eliminate those cells that have suffered aberrant changes. However, some cancer cells are able to evade the immune surveillance [62] through different mechanisms, for example producing proteins that bind to and deactivate immune cells [159].

Knowing and understanding the characteristics of cancerous cells is important to find genetic signatures or biomarkers that can alert us of the presence of the disease,

ideally in its early onset -for example liquid biopsies - as well as to find efficient/effective treatments [62], [13]. It is also important to understand the genetic mutations that may be driving the disease or cancer dependencies.

1.4 Towards a molecular understanding of cancer

In the battle against cancer, it soon became evident that in order to find an efficient treatment, we needed to understand the molecular mechanisms driving oncogenesis. One of the first theories to emerge was the somatic mutation theory, first stated in 1914 [157]. Originally, its main premise was that cancer is derived from a somatic cell that has accumulated multiple DNA mutations in genes that control cell proliferation, making them more active [147], and resulted from the observation that exposure to DNA-damaging chemicals often induced the disease and supported by research using model systems such as cancer cells and oncogenic viruses [38]. The study of heritable cancers uncovered different gene alterations that rendered some genes inactive [147]. Since being formulated, the theory has grown to account for the many different roles that genes have in oncogenesis, directly and indirectly, and remains the paradigm of cancer research [157]. The realization that the cause for cancer may be buried within our genes propelled the search of other genes responsible for oncogenic events such as metastasis [38] and provided one of the reasons to embark on the Human Genome Project (HGP), an ambitious research effort seeking to sequence and map of the genes of our species that began in 1990 and was completed in 2003 [151].

At the time the HGP commenced, sequencing was far from what it is today. The birth of recombinant DNA technology in 1973 made isolating individual genes possible, but sequencing was extremely laborious. The first generation DNA sequencing was developed by Sanger and colleagues in 1965 [137], [65]. The technique, later known as Sanger sequencing, was subsequently improved and other protocols, like Maxam and Gilbert's [107] appeared. However, it was in 1977 when Sanger's chain-termination technique revolutionized DNA sequencing [65] and made it possible to elucidate 300 to 500 bp of DNA sequence in a single experiment [[172]. When the HGP started, Maxam-Gilbert sequencing and Sanger sequencing were used. Automation tools and high-throughput sequencing technologies had yet to be developed and the cost of sequencing a single DNA base was of approximately $10 and involved the construction of clones for sequencing and PCR amplification [41]. The HGP - together with other institutions, such as the Welcome Trust in the UK - also propelled the development of high-throughput sequencing technology, or next-generation sequencing (NGS).

The first large success of the second generation DNA sequencing, or next-generation sequencing, was pyrosequencing. This method used a two-step luminescent technique that made it possible to observe the synthesis in real-time and avoided alteration of the nucleotides with chemical labels. Moreover, it does not need a cloning step of the DNA fragment to be sequenced which, together with other characteristics, allows for massive parallelization, significantly increasing the amount of DNA

that could be sequenced in a single experiment [65], [175]. Nowadays, the most successful sequencing platform is owned by Illumina, ensconced in a near monopoly and probably the most worldwide used NGS technology [60]. The advent of NGS technologies was of especial importance for functional genomics. The development of RNA-Sequencing in 2007 [43] allowed for profiling of the transcriptome using deep-sequencing technologies - providing far more precise measurement of transcript and isoform levels than previously used technologies, such as microarrays [171], [175]. Before RNA-sequencing, the transcriptome was already studied using microarrays or Sanger sequencing, far more laborious techniques that also lacked the accuracy and sensitivity of RNA-Seq. The profiling of the transcriptome with RNA-sequencing and the fallouts of this technology are discussed more profoundly in the Theory section.

Understanding the transcriptome is important for understanding the functional elements of the genome, and thus for increasing our understanding of development and disease [171]. RNA-sequencing experiments have allowed for quantifying the levels of expression of the genes across tissues, development stages or in disease; or to decipher the transcriptional structure of the genes - for example the splicing patterns [171]. Other technologies also make use of NGS to study other events; for example ChIP-Seq is used to discover interactions between the DNA and proteins [138].

The advancement of NGS - together with the completion of the HGP - also revolutionized cancer research, giving birth to cancer genomics [14], a discipline that analyses the DNA and RNA of a tumor and classifies it based on its molecular characteristics - among other features, the presence or absence of mutations.

Deciphering the molecular events in the genome and the transcriptome driving cancer progression has had a huge impact in choosing the right treatment for each cancer [115]. For example, non-small-cell lung cancer, a cancer type with poor response to chemotherapy, was further classified into subtypes depending on its driver mutation. One of these subtypes of lung cancer was characterized by mutations in the epidermal growth factor receptor (EGFR) gene. Mutated EGFR, in turn, usually responds to inhibitors of epidermal growth factor receptors. In fact, two studies have proven that EGFR-positive non-small-cell lung cancer treated with an EGFR inhibitor shows longer progression-free survival than when treated with standard chemotherapy [119]. Thus, non-small-cell lung cancer was one of the first examples of using the genetic and molecular framework of the disease to guide its therapy strategy [53]. However, in spite of the surge of genomic cancer information, genome-based treatments have not evolved accordingly, probably because of the complexity of the disease pathogenesis, which goes beyond mutations in the genome [173]. The next logical step is to combine every piece of information in a multi-omics approach - adding the information from the epigenome, transcriptome, proteome and metabolome - to draw a better understanding of the events driving cancer, improve the precision of treatments, and discover new biomarkers of disease [10], [160]. Such an integrative approach will also improve our knowledge of those malignancies with a non-mutated but dysregulated genome or transcriptome. The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are two

of the consortiums that set out to fully characterize the genome of thousands of tumors. A huge amount of data combining mutations, gene expression and protein levels across 33 cancer types has been collected [11]. The genomic characterization of malignancies has aided the management of some cancers. For example, in breast cancer it was demonstrated that the expression of poly (ADP-ribose) polymerase-1 (PARP1), is upregulated in tumors with a negative expression of estrogen receptors (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) - or negative expression of the three of them [118]. A tumor with these molecular characteristics would be susceptible to therapy with PARP inhibitors - drugs that target this particular gene. Nowadays, molecular testing is routinely used for the management of breast cancer, to establish whether the tumor is susceptible to treatment with PARP-inhibitors - or if it presents any other targetable mutation [100]. Nevertheless, the enormous amount of information does not directly translate into improved therapy - with less than a quarter of patients suffering from the most common cancers benefitting from targeted therapy [11], since we do not know how most mutations affect or perturb drug activity. Hence, beyond the molecular characterization of human malignancies, it is important to understand the function of the mutated genes, the molecules with which they interact and their molecular networks. Furthermore, to complete the picture we need to understand how perturbations in genes and other cellular elements - transcripts, proteins - connect with the cellular phenotype and therefore influence the response to a treatment. Lately, functional genomic studies are showing how screening technologies such as shRNA and CRISPR/Cas9 screening platforms can be used to identify genome or context-dependent cancer vulnerabilities [61]. One of the most ambitious projects to date is The Cancer Dependency Map [154], [179], a consortium formed by the Broad Institute and the Wellcome Sanger Institute collecting massive amounts of data to uncover the vulnerabilities of cancer cells.

Probably, the next step will be a rise in the use of in-vivo functional screening - using genetic engineering and screening tools to perform high-throughput in-vivo genome perturbation, once the limitations of CRISPR technology in in-vivo systems are overcome - which would allow for the study of the temporal dynamics of context-dependency changes during the evolution of the disease [173].

1.5 The Cancer Dependency Map - systematic functional screening to understand cancer

The Cancer Dependency Map (DepMap) Consortium was created with the aim of supporting the systematic discovery of novel dependencies - vulnerabilities of the cancer cells that arise from the mutations that cause cancer cells to grow - and the efficient identification of patient populations for target discovery programs [154], [179]. DepMap scientists are characterizing hundreds of cancer cell line models collecting genomic information and sensitivity to genetic and small molecule perturbations. To achieve these goals, the consortium is divided in three projects, each working towards a particular goal:

- The Cancer Cell Line Encyclopedia (CCLE) [54] started in 2008 as a collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research. The project aims to characterize cell line models to study the genome diversity across different human cancers, in order to devise efficient targets for precision medicine.
- The Project Achilles [154], [56], [80] systematically identifies gene essentiality across the characterized cancer cell lines. To conduct the screenings, they use genome-scale RNAi and CRISPR-Cas9 genetic perturbations to knock out or silence single genes and evaluate their effect in the survival of the cell. After perturbation of an individual gene, they measure cell growth, cell death, gene expression and cellular morphology [11], [33].
- To screen for drug sensitivity, screening methods such as PRISM (Profiling Relative Inhibition Simultaneously in Mixtures) [179], [21] and CTPR (Cancer Therapeutics Response Portal) [130], [140], [7] aim to understand the effect of drugs in the characterized cell lines, therefore evaluating the effect of the compounds in a broad genetic spectrum.

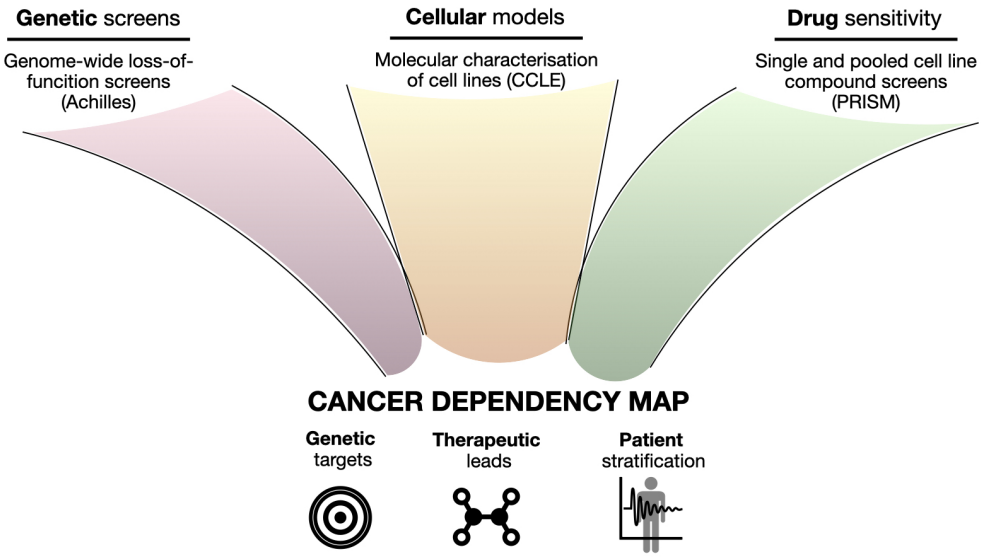

Figure 1.3: Overview of the different ongoing projects at the Cancer Dependency Map. Figure adapted from the DepMap portal.

The data generated from these experiments was processed and deposited in the DepMap portal, where it is publicly available for other scientists to use and analyse. In this thesis, I focused on the datasets provided by DepMap instead of other resources, such as the in-vivo genomic information contained in TCGA, mainly due to the functional information provided by molecular perturbation screens (i.e. Achilles). Data from these screens makes it possible to infer the mechanistic consequences that a cancer-specific molecular feature (e.g. a mutation) may have in

the cell. Although it is true that information derived from cell lines will inevitably miss some of the complexities of cancer biology, in vitro research is still undoubtedly useful to understand the mechanisms behind cancer development and even to find new therapeutic targets - with most drugs available today discovered from in vitro experiments [11].

Multi-omics and functional screening experiments usually result in large data sets whose information needs to be integrated in order to obtain a comprehensive characterization of the organism or model of interest. Such a task is too complex to be achieved by simple investigation of pairwise correlations [44]. Therefore, concomitant with the surge of data is an increased demand for new tools and methodologies that allow for an efficient analysis of these big datasets. Machine learning and deep learning are good examples of such tools, being algorithms able to detect patterns and features from the data without the need of previous knowledge. Machine learning and deep learning methodologies have been successfully applied to analyze genomic data since the demonstration of the applicability of deep learning to DNA sequence data in 2015,[98], [117], [32], [67], [141], [44], [2], [184]. The application of these methodologies has allowed for significant advances in functional genomics and regulatory genomics, for example [185], [44]. More recently, an AI-based tool was used to successfully the origin of metastatic tumors [101].

2

Theory

As I explained in the previous section, there are many different molecular mechanisms by which cancer cells can acquire the hallmarks that characterize them. This divergence is one of the reasons behind the complexity of cancer. One such mechanism is the dysregulation of the splicing process by either alterations on the DNA sequence of the spliced gene, or mutations in the spliceosome. The goal of this thesis is to use a multi-omics approach to relate alterations in the splicing process with cancer dependencies or response to drugs, using the data generated by the DepMap consortium. Before getting into the details of the research, I explain in this section the key concepts to understand the analysis and its results.

2.1 Revisiting fundamental cell biology

Molecular biology has long revolved around the central dogma of molecular biology, first coined by Francis Crick [24], [23]. The dogma broadly describes the transfer of information in the cell during DNA replication, transcription into RNA and translation into proteins [126]. We now know that the dogma is an oversimplification that overlooks many regulatory features - one such feature being the maturation of messenger RNA (mRNA), of which we will talk more about later - but nevertheless the central dogma is still useful to illustrate the basic functioning of the cell.

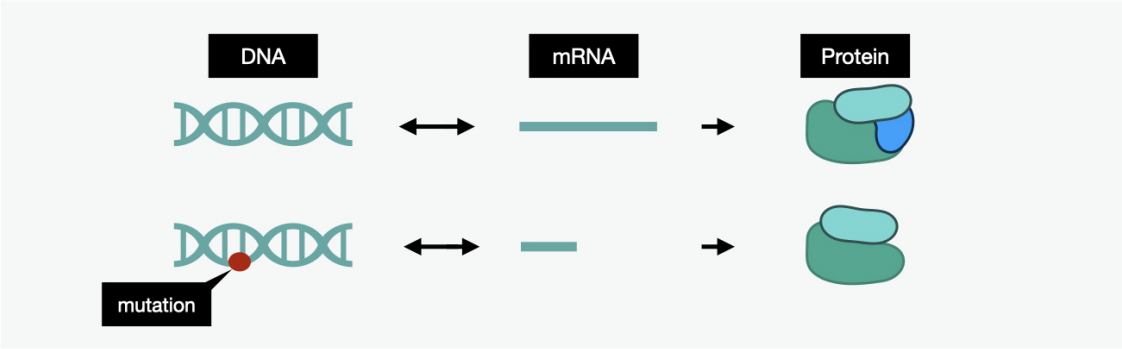

Figure 2.1: Simplified representation of the central dogma of molecular biology, illustrating the effect that a mutation in the DNA may have in the resulting protein.

In cancer, mutations or epigenetic alterations at the DNA level often interfere with this flow of information and result in altered proteins which cannot carry out their normal function anymore, disrupting the homeostasis of the cell. However, there are

many ways by which these alterations affect the resulting protein. One such way is by altering the process of splicing, by which the pre-mRNA matures into mRNA that is ready to migrate to the cytoplasm and be translated into proteins.

2.2 Splicing and alternative splicing

Cells need to fabricate proteins in order to properly function. As explained before, these proteins are encoded in the DNA of the cell, which is transcribed into different mRNAs. These mRNAs are then translated into proteins [1]. Until 1977, it was assumed that every base pair in the gene contained information transferred to the protein. A series of observations proved that this was not the case for eukaryotic organisms, and it was shown that most genes in multicellular organisms contain introns, segments of base pairs that were present in the DNA but not in the shorter mRNA that was found in the cytoplasm [143]. Thus, there is one extra step prior to translation, namely mRNA maturation. Before being translated, the mRNA suffers a catalytic reaction known as splicing, which is defined as the process of intron removal and exon ligation of the majority of the exons in the order in which they appear in a gene [167]. Almost at the same time splicing was discovered, alternative splicing was described as well. Alternative splicing is a deviation from the standard mRNA maturation process by ligating a different number or combination of exons [167]. This leads to a variety of mature mRNAs able to be translated into proteins, all of which result from the same gene. In fact, alternative splicing leads to the differences in transcript composition across different tissue types, with estimates that over 50% of all splicing events differ among tissues [165].

Alternative splicing has an important role in several processes of eukaryotic organisms [40], [84], being the major contributor to transcriptome and proteome diversification in eukaryotes. For humans, this means we can produce way more different proteins than we have genes: approximately 95% of genes undergo alternative splicing to produce different mRNA transcripts [84], [164]. In fact, it is estimated that between 15% to 50% of human disease causing mutations seem to affect splice site selection [166], [6].

2.3 The splicing code

The processes of splicing and alternative splicing modulate many critical cellular processes [6] and usually work robustly and with minimal errors. The precision of the process, however, astonished researchers, who could only find two universally conserved nucleotides: GU at the 5´ site of the intron and AG at its 3´ site, which seemed insufficient to explain the precision of splicing [5]. Later, the discovery of other sequence features (the branch point and the polypyrimidine tract) further explained the accuracy of the process, defining the exon/intron junction - or splice sites, or consensus sequences -, to which spliceosomal factors and other proteins could bind to catalyze splicing [5]. However, these features were still not enough to justify the accuracy of the reaction - usually error free -, least when taking into

account alternative splicing. The discovery of auxiliary sequences, namely cis-acting elements - which depending on their location and effect on splicing were termed enhancers and silencers - and trans-acting elements - splicing factors -, helped to better understand the regulation of the process. Yet again, none of these elements are well-conserved, and their functions are poorly understood. Enhancers and silencers can be located within exons or introns, depending on which they are considered exonic or intronic splicing enhancers (or silencers). Enhancer sequences located within an exon promote exon recognition, which in term promotes RNA splicing. A silencer, on the other hand, inhibits splicing of adjacent sequences [143]. Furthermore, other factors also influence the outcome of splicing: the RNA structure, the transcription rate and transcription factors (since splicing is synchronous to transcription [83], [143], [85]), chromatin remodelling and epigenetic modifications [5].

In short, the regulation of splicing is far from completely understood, partly due to its complexity and partly due to its degeneracy. Less than 5% of 5' splice sites are a perfect match of the established consensus sequence [5]. The degeneracy of the splicing code is probably for the sake of plasticity, since it allows for alternative splicing and in turn, for greater proteome diversity. However, this plasticity also difficults the distinction between canonical and alternative splicing of a gene, which is often context dependent [150].

2.4 Alternative (aberrant) splicing in cancer

Although we know that splicing events contribute to proteome diversity and have a functional impact, the function of many of these events remains unknown [85]. Furthermore, alternative splicing often results in products that have no function [85]. Aberrant splicing - abnormal alternative splicing - is widespread in cancer [158], [85]. First thought to be a consequence of other cancer driving mutations, now it seems clear that some of the aberrant splicing confers new properties to the cancer cell [46], [79], although a percentage of it is probably cancer-specific noise likely not relevant for the pathology of the disease [139]. An example of altered splicing contributing to tumorigenesis would be mutations in the splicing factors SRSF2, SF3B1 or U2AF1, all part of the spliceosome and involved in the recognition of the 3´splice site. These mutations would alter the splicing by affecting the recognition of the splice site, which would affect the stability of transcripts that encode proteins that promote transformation [57].

Previous analysis of data from The Cancer Genome Atlas (TCGA) highlighted the extent to which mutations that alter splicing patterns are driver mutations for cancer, showing that 119 genes that encode core spliceosome and splicing factors are suspected to be drivers for the disease across 33 tumor types [142]. Aberrant splicing has proven so relevant to oncogenesis that some researchers have stated that dysregulated alternative splicing should be considered a hallmark of cancer [87].
Beyond improving the molecular characterization of different cancer types, understanding the role of aberrant splicing in cancer could unveil novel treatment avenues. In fact, clinical trials testing spliceosome inhibitors have already been conducted, al-

beit unsuccessfully due to cytotoxic effects [164]. Nevertheless, a number of clinical trials are currently ongoing using spliceosome inhibitors that show less toxicity (i.e. H3B-8800 being tested in clinical trial NCT02841540) or inhibitors of the protein arginine methyltransferase, PRMT5, which have shown antitumor effects associated with aberrant splicing across many cancer types (i.e. the drug GSK3326595, tested in two clinical trials with NCT numbers NCT03614728 and NCT02783300) [164].
Furthermore, alternative mRNA splicing may offer a wide variety of novel target sites for cancer immunotherapy [50]. Aberrant splicing events generate a suite of novel peptides that can bind to MHC class I molecules and serve as neoepitopes [50]. Recent studies have shown that the target space of splicing-derived neoepitopes is significantly higher than those of somatic single-nucleotide variant events [50]. Furthermore, a recent study has shown that the majority of tumor-specific antigens derive from presumably non-coding regions [89], and therefore could be missed by standard alternative splicing analysis that rely on exon usage.

2.5 The spliceosome

One of the causes behind aberrant splicing are alterations of the DNA sequence to be spliced, which would affect the splice sites and the binding of splicing factors and other associated proteins. In addition to this, research has also shown that many of the observed changes reflect alterations in components of the spliceosome [4]. The spliceosome is the molecular machinery that catalyzes the splicing process. The core spliceosome and its associated factors consist of more than 300 proteins and 5 small nuclear RNAs (or snRNAs) that both perform and regulate splicing and alternative splicing [4]. The spliceosome is found in the nucleus and it forms on the pre-mRNA by recognizing active splice sites in the chain [161]. It is a very complex and dynamic construct that changes composition across the different stages of the splicing process [161]. Such dynamism is important to respond to regulation cues that may control the catalysis of canonical splicing or alternative splicing [65]. It has also been reported that the splicing of a particular type of introns, named minor or U12 introns, utilize a different spliceosome composed of different snRNPs [40], highlighting even further the plasticity of the molecular machine. Furthermore, cancerous cells across all cancer types present high levels of intron retention, which exemplifies the importance of the spliceosome and its composition in the transcription profile of human diseases [40]. Yet another example was provided by a study that showed that the alternative splicing of important metabolic genes is regulated by splicing factors, suggesting that splicing factors have an important role in tumor metabolism [181].
Moreover,the spliceosome has also been identified as a cancer vulnerability for certain tumors. As an example, MYC-driven tumors need of spliceosome components for MYC to behave as an oncoprotein [4], [69]. MYC is a transcription factor and one the most frequently overexpressed oncogenes in cancer, often driving transformation. Previous work has shown that interactions between MYC and splicing factors such as SRSF1 contribute to the oncogenic activity of MYC, sometimes even enhancing its malignant potential [27], [81]. Furthermore, research has revealed that certain spliceosomal elements are essential for the oncogenic activity of MYC, opening to

the possibility of targeting the spliceosome in order to treat MYC-driven cancers [4], [69], [81].

2.6 Machine learning & Deep learning

In order to know the role a mutation plays in the disease, we need to take into account complex regulatory relationships that are not fully understood yet. Nevertheless, the DNA sequence contains most of this regulatory information - with the exception of epigenetic mechanisms, such as DNA methylation [102], [112], [109]. Because of this, the use of machine learning algorithms that are able to learn and recognize patterns from raw data is suitable to understand or predict the effect that a mutation can have on the phenotype of a cell or disease .

The term machine learning refers to algorithms that are able to extract patterns and learn features directly from data, and to improve their performance with experience or, in other words, with the amount of data fed to the algorithm - the more data the algorithm has, the more it will learn [98], [58]. However, the performance of most simple machine learning algorithms relies on the pieces of information that they are given - each piece of information known as a feature of the data [58]. When applying machine learning to genomics, it is usually hard to know what features should be given to the algorithm - if each nucleotide in the DNA sequence is a feature, it is nearly impossible to know how each individual base pair will affect the phenotype of the cell, partly due to our incomplete knowledge of some molecular mechanisms and the regulation behind some processes. Such a problem - also encountered in computer vision, for example, where the goal is that the algorithm recognizes the different elements of an image - was solved with the development of an approach known as representation learning, by which the algorithm is able to discover the representation of the data on its own, meaning it is able to find a good set of features with minimal human intervention [58]. Deep learning defines methods that are able to implement representation learning by using multi-layer algorithms that build complex features out of simpler features in successive elementary operations [58]. In this way, it can extract useful information from the given data without the need of explicitly providing an appropriate representation. For example, in classifying a tumour as malign or benign, a deep learning algorithm can take into account not only cell counts - which is the information that would be manually provided by researchers - but also visual features like cell morphology or location, and identify patterns in these features that are relevant for the classification problem [44]. Similarly, using deep learning algorithms in genomics or multi-omics data means that the algorithm is able to evaluate the complete context of the raw DNA sequence and learn patterns and features embedded in the nucleotide chain - without the need of feature selection. Not needing feature selection is extremely important in the context of genomics and other omics, since our knowledge of the molecular events that may influence the different phenotypes is incomplete. Hence, if we were to give features as a representation of the data, this representation would most likely be flawed and affect the performance of the algorithm.

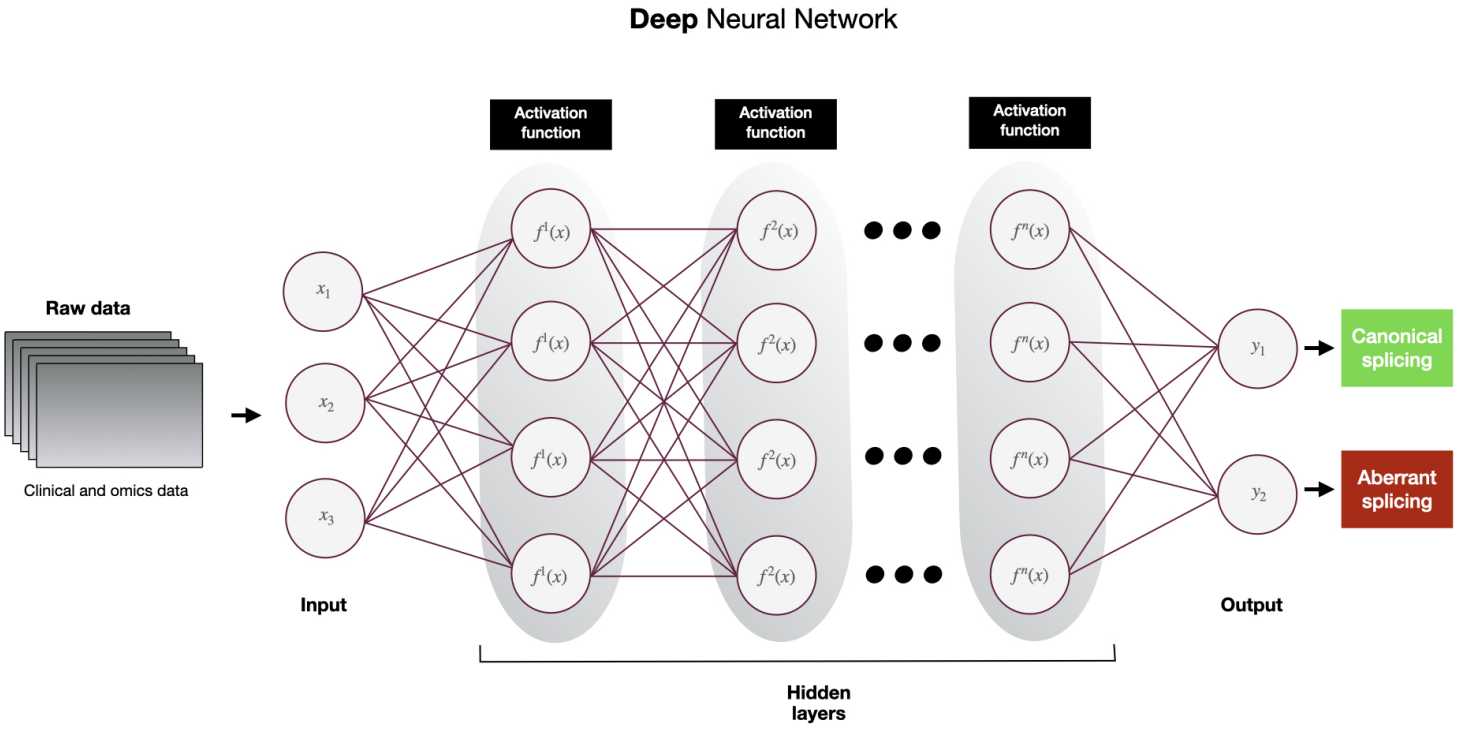

Figure 2.2: Simplified overview of a deep neural network. The example takes raw clinical and omics data as input and eventually predicts whether a specific sample presents the molecular signatures typical of canonical or aberrant splicing.

Since dysregulated alternative splicing or mutations in the spliceosome play such an important role in the oncogenesis of some cancers, being able to predict whether a mutation will affect the splicing process would help to better understand the impact of the variant, eventually allowing for more informed therapeutic decisions.
Traditionally, alternative splicing has been studied using RNA-sequencing, which enables transcriptome-wide profiling of the process [182]. Although RNA-sequencing has allowed for unprecedented characterization of alternative splicing, the technique presents some problems. One of the issues of using RNA-sequencing to analyze alternative splicing is the reliance of the sequencing method on high sequencing coverage, which means that it is not reliable to detect splicing in genes with low expression [182]. Some probabilistic methods such as Cufflinks [153] and MISO [77] perform relatively well even for low-coverage, but do not completely solve the challenge [70]. Moreover, low-cost RNA-sequencing methods that rely on only capturing the 3' tail of transcripts are unfit for analysis of splice variants. However, since it was proven that splicing (in bulk cells) can be reliably predicted from sequence-associated features [70], [176] many machine learning models have been developed to predict splicing levels both in bulk tissues and single cells [153], [94]. One such method is the deep-learning tool **SpliceAI** [72]. As explained before, deep learning algorithms are a subtype of machine learning that uses representation-learning methods - features do not need to be fed by the user - with multiple levels of representation - or many

layers [42], [92]. Each layer performs a nonlinear transformation on the previous layer, building up a complex representation that makes them suitable to understand complex patterns [146], [145]. Deep learning is particularly useful to study the process of alternative splicing, since alternative splicing is a highly regulated and complex biological event whose regulation and mechanism are not completely understood yet. Even though we do not know exactly how splicing is regulated, the splicing code is written in the DNA sequence. Therefore, deep learning approaches that do not require feature selection are useful to extract patterns from this primary sequence.

2.7 Using deep-learning to predict alterations in the splicing process

There are many diseases caused by aberrations in the splicing process, such as the Prader Willi syndrome, Treacher-Collins syndrome or prostate cancer [20]. Many of these diseases may be caused by so-called cryptic splice variants, mutations occurring outside the consensus GT and AG splicing nucleotides [72], [20], [75]. Identifying such cryptic variants in the clinic is hard, mainly because of our incomplete knowledge of the splicing code [72], [170]. Furthermore, even though RNA sequencing has proved to be a useful tool for detecting splicing aberrations [20], [30], its capacity to detect random novel mutations is limited because of the extremely large genomic space in which splice-altering mutations can take place [20]. To overcome these limitations, Jaganathan et al. developed a deep-learning algorithm able to predict splicing events from any given pre-mRNA sequence, evaluating whether each nucleotide in the transcript is a splice donor, splice acceptor or neither [20]. SpliceAI is a deep residual neural network - an approach developed to tackle the loss of accuracy when increasing the depth (or the number of layers) of an algorithm [64].

SpliceAI deep learning algorithm spans 32 dilated convolutional layers that are able to identify sequence determinants separated by very large genomic distances [20]. Convolutional layers are those that apply a convolution operation, which merges two sets of information and eventually serves to detect specific features from the input [34]. The algorithm considers a window of 10,000 nucleotides of the flanking context sequence of the evaluated position [20], which is important considering that the different splicing sites could be located very far from each other in the genome. The tool was developed to aid with the diagnosis of rare genetic diseases, where cryptic splice variants have been found to play a critical role [20]. To evaluate its performance, the accuracy of the algorithm was compared to those of three other classifiers referenced in the literature for the diagnosis of genetic rare diseases: GeneSplicer [124], MaxEntScan [177] and NNSplice [131]. SpliceAI outperformed these tools in terms of specificity, probably because the other classifiers focus on local motifs and do not take the whole context of the sequence - i.e. long-range specificity determinants - into account [20]. SpliceAI was tested on data from patients with autism spectrum disorders and severe intellectual disabilities (using data from the Deciphering Developmental Disorders (DDD) study, which studies severe neurodevelopmental

disorders [108]; and data from the Simons Simplex Collection and the Autism Sequencing Consortium, both containing data from individuals with autism spectrum disorders (ASDs) [30], [136], [155]. The analysis successfully showed that splice-disrupting mutations are enriched in intellectual disability and in autism spectrum disorders [20]. Even though SpliceAI has shown to be able to predict splice-altering mutations with high accuracy and specificity, the tool presents some limitations. The algorithm only considers variants within genes defined by the gene annotation file. Therefore, its prediction power is somewhat constricted by the quality of the reference genome and the associated annotation file, and limited to the genes that have been annotated there. Additionally, SpliceAI does not work on variants that are close to chromosome ends (in 5 kilobases on either side of the genome) , on deletions of length greater than twice the input parameter, or on variants that are inconsistent with the fasta file used as reference [71].

3

Methods

3.1 Datasets

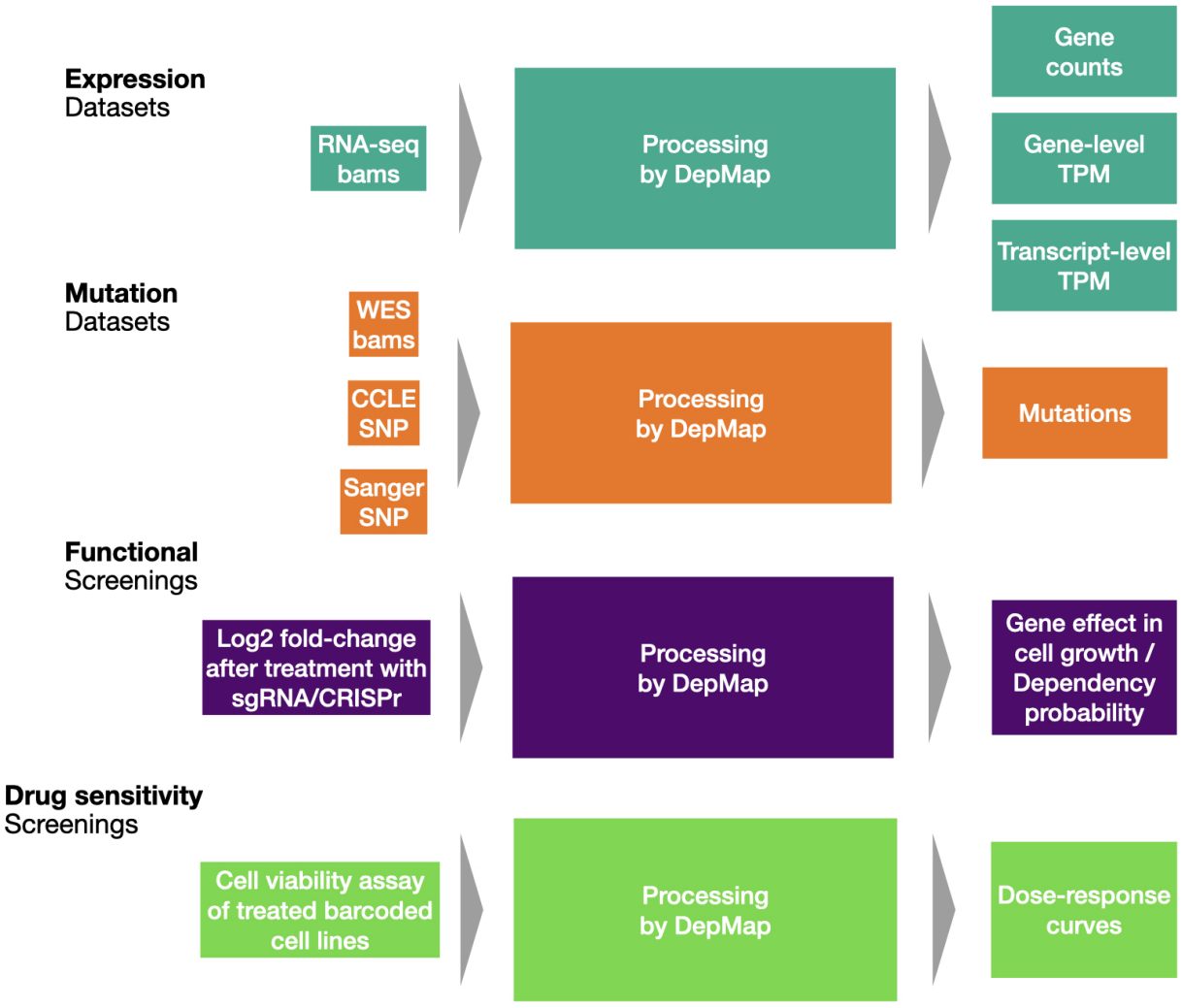

Figure 3.1: Overview of the processing steps used by DepMap to generate the datasets used in the analysis.

3.1.1 Mutation datasets

The mutations dataset was collected by the DepMap consortium and downloaded from the DepMap portal. The version used in this project was the release 21Q1,

uploaded in the portal in February 2021. The file contains the genomic characterization of 1747 cell lines comprising 35 different cancer types. The data was generated as part of the Cancer Cell Line Encyclopedia (CCLE) project [54]. The file is written in Mutation Annotation Format (MAF), aggregated from annotated Variant Calling Format (VCF) files. It is important to note that the VCF files often report variants on multiple transcripts whereas the MAF files generated from the VCF only report the most critically affected one.
In these datasets, the mutations annotated as splice site mutations are those aberrations occurring exactly at consensus annotated splice site sequences, and do not include mutations that may affect splicing elsewhere. DepMap generated these annotations using Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) data and the tool Oncotator [129].

3.1.2 Expression datasets

The expression dataset used was generated by the DepMap consortium and downloaded from the DepMap portal. The analysed dataset in this study was version 21Q1, released in February 2021.
The file *CCLE_expression.csv* contains RNAseq TPM gene expression data for protein coding genes - not accounting for the different isoforms. The transcript quantification was done by the DepMap team using RSEM [96]. The reporter expression levels were log2 transformed, using a pseudo-count of 1.
The file *CCLE_expression_full.csv* contains RNAseq TPM gene expression data for all genes. Transcript quantifications were also performed using RSEM and expression levels reported with a log2 transformation using a pseudo-count of 1.
The file *CCLE_RNAseq_transcripts.csv* contains RNAseq transcript TPM data - which means that expression levels are reported for the different isoforms present in the cell. The transcript quantification was done by the DepMap team using RSEM. The reporter expression levels were log2 transformed, using a pseudo-count of 1.

3.1.3 Functional screening - Achilles knock-out screening

For a functional analysis of the mutations, we used the data generated by the Achilles project, also stored in the DepMap portal [33], [110]. The analysed data set is the version uploaded in the release 21Q1.
The file *Achilles_gene_effect.csv* contains the results of the pipeline CERES. The data was processed and scale so that the median effect of nonessential knocked-out genes is 0 and that the median effect of essential knocked-out genes is 1.
The file *Achilles_gene_dependency.csv* reports the probability that knocking-out a particular gene has a real depletion effect. This metric is calculated from the gene effect dataset.
Interpretation of the gene effect data (*achilles_gene_effect.csv*) should be done as follows: a negative score corresponds to an essential gene in a cancer type, meaning that if the gene is lost, the cancer cells will not survive - which means the gene may be a potential oncogene. A positive score points in the opposite direction - identifying the gene as a potential tumor suppressor. However, according to DepMap

small positive scores are likely to be noisy and CRISPR data is not as reliable for identifying tumor suppressors as it is for identifying tumor dependencies. The difference between both files is that the "gene effect" contains the corrected CERES score, which measures the effect size of knocking-out a gene and normalizes this effect against the distributions of non-essential and pan-essential genes. The dependency scores, on the other hand, assess how likely a certain gene is to belong to the non-essential distribution or the common essential distribution in the studied cell line.

3.1.4 Drug sensitivity

The dataset was generated as part of the PRISM repurposing project, which aims to treat molecularly barcoded cell lines with non-oncology drugs and monitor their effect in the cell line growth [21]. The file was uploaded to the DepMap portal as part of the release of 19Q4, in December 2019. The dataset screened 4686 compounds in 578 cell lines.

The file *primary-screen-replicate-collapsed-logfold-change.csv* contains the log-fold change values relative to controls grown in DMSO . In this dataset, the more negative the reported value is, the more sensitive is the cell line to that compound. Values larger than 0 suggest that the treated cells grow more than control cells, but this is usually an artifact.

3.1.5 List of spliceosome & splicing factors genes

The list of genes coding for elements of the spliceosome machinery and other splicing factors to analyze was extracted from the research work carried out by Seiler et al. [142]. The list contains 404 genes categorized as part of the core spliceosome elements or not by Hegele et al. [66]. Some of the splicing factors have not been located in any of these two categories, and they have been labelled as "other" for the purposes of this analysis. According to this study, those genes labelled as "core" genes were present in high abundance and known fucntion, or for being associated with any of the parts of the spliceosome (i.e. with any of the snRNPs). The non-core proteins, on the other hand, include mRNA binding and regulatory proteins [66]. A comprehensive list of the splicing factors included in the analysis can be seen in the Table A.1 in Appendix A.

Category	Number of genes
Core	141
Non-core	103
Other	225

Table 3.1: Summary of the categories in which the genes coding for spliceosome proteins and other splicing factors have been placed

3.2 Hardware

3.2.1 Local analysis

Most of the analysis was performed locally using a MacBook Pro (2018) with 32 GB of memory and a 2,6 GHz 6-Core Intel Core i7 processor, provided by the Systems and Synthetic Biology Division at the Department of Biology and Biological Engineering at Chalmers University of Technology.

3.2.2 High-performance computing environment

We used the deep learning tool SpliceAI to predict whether a mutation reported by the CCLE project could potentially disrupt the splicing process by modifying splice donor or acceptor sites - or by creating new ones. SpliceAI computations were performed in the PC-cluster Hebbe, a high-performance computing environment. The access to the cluster was enabled by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE).

3.3 Software & tools

Unless otherwise stated, the bulk of the analysis was carried out in the R programming language, version 1.4.1106.

3.3.1 SpliceAI

As mentioned before, SpliceAI is a deep residual neural network employing a network architecture consisting of 32 dilated convolutional layers. The input for the deep learning network was the mutation datasets from CCLE, formatted into a VCF file (see Supplementary information for more details).
SpliceAI was run in Hebbe, a local computing cluster at Chalmers University of Technology. Access to the computing facilities was provided by C3SE (see below). In order to run the software, we created another conda environment in the cluster in which we installed both the deep learning tool (from bioconda) and tensorflow - an open-source platform for machine learning containing tools necessary for the analysis. The reference genome fasta file that was used for the analysis was the assembly GRCh37[1] since it is the genome used by DepMap for annotation in their experiments [15].
To facilitate running the pipeline and the parallelization in the cluster, the mutation file was divided by chromosome and SpliceAI was run for each chromosome separately. The output VCF files were then parsed together into a single file, ready for analysis. The output of SpliceAI consists of 4 different scores that reflect the probability that a particular mutation eliminates or creates a splice donor or splice acceptor site.

[1] GRCh37 reference genome was downloaded from the NCBI portal

3.3.2 Code availability

The code written to do all the analysis reported in this manuscript can be found in the GitHub repository SplicingInCancer.

3.4 Mutational landscape & Gene expression patterns in spliceosome deficient cancers

In order to understand the context of the mutations on the splicing factors, we began by exploring the amount of mutations in the spliceosome suffered by different cancers, and identifying those splicing factors bearing most mutations.

3.4.1 Dimensionality reduction

All the analysed datasets span thousands of features which makes the interpretation virtually impossible. Therefore, we applied the linear and non-linear dimensionality reduction techniques Principal component analysis (PCA) [123], [73] and t-distributed stochastic neighbor embedding (t-SNE) [103], respectively, for visualization of our high-dimensional data. We used these techniques in the exploratory analysis to get a sense of the distribution of the data and see if these algorithms could detect patterns in the different datasets - expression, knock-out screening and drug sensitivity - between cell lines with intact splicing factors and those with mutated spliceosome machinery.
PCA is an unsupervised learning method able to find patterns without the need of previous knowledge regarding the sample characteristics. The algorithm works by reducing the dimensionality of the data by geometrically projecting the data points onto lower dimensions - or principal components (PCs) - with the goal of finding the minimum number of PCs that best summarize the variance of the data [95].
Although t-SNE is also a method for dimensionality reduction, both algorithms work in different ways. The main difference is that while PCA tries to maximize the variance, t-SNE preserves local similarities. In other words, where PCA tries to maintain the global properties of the dataset, t-SNE plots neighbors close to each other, allowing for visualization of both the local and global structure of a data set.

3.5 Transcriptional differences related to spliceosome deficiencies in cancer

3.5.1 Differential expression

The expression analysis aims to characterize the transcriptome: to quantify and measure the relative frequency of all expressed transcripts in the cell at a certain time [96]. We analyzed the differential expression on transcript data for both coding and non-coding genes. As mentioned before, the reads are reported in the form of transcripts per million (TPM), with a pseudocount of 1 (i.e. $\log(\text{TPM} + 1)$).

Transcripts per million is a measurement of relative expression: the fraction of transcripts of the transcriptome produced by a certain gene or isoform [96].

$$\tau_i = \frac{v_i}{l_i}(\sum_j \frac{v_j}{l_j})^{-1} \tag{3.1}$$

Reporting the reads for each gene isoform as TPM is slightly problematic to perform typical differential expression analysis - standard tools like edgeR do not work well on TPM reads. TPM is a measurement of the proportion of transcripts in the sample. Therefore, this measurement allows for comparison of transcript with different lengths - since each transcript is normalized by length - such as different isoforms or different genes within a sample. However, it is worth noting that this measure can be biased by expression levels of other transcripts. Such bias makes it impossible to correct for library composition and therefore TPM expression values are not suited for differential expression analysis using tools such as edgeR. To overcome this problem, we tested whether there is a significant relationship between the expression level of isoforms (log(TPM) data) and the presence of mutations in splicing factors. When considering mutations in all splicing factors, the amount of cell lines that remained "wild-type" - thus not presenting mutations in any splicing factor - was quite low compared to the number of cell lines presenting a mutation. In an attempt to balance the number of cell lines in both categories, we only took into account cancer types presenting both cell lines with and without mutations - reducing the number of cell lines with mutations in the splicing factors being considered for the analysis. This is also important because we need to include the variation across disease in the analysis. Hence, we only kept those cancer cell lines for which at least 3 wild-type samples were available.

To test the relationship, we used a linear regression model and evaluated the relationship between expression levels and the presence of mutations in the spliceosome. To correct for multiple testing, we applied the Benjamini-Hochberg procedure to adjust the p-values. We considered significant those p-values below 0.05. As a result of the linear regression analysis,we obtained p-values for each transcript - from here on referred as "transcript-level p-values". However, extracting biological insights from raw p-values is a daunting - nearly impossible - task. Thus, we decided to perform a gene set analysis in order to facilitate the interpretation of the results. However, most tools require gene-level p-values for gene set analysis.

3.5.2 Lancaster method for the aggregation of transcript-level p-values

RNA-seq data is mostly studied at the gene level, among other reasons because the analysis seems to be more robust and because results are easier to verify experimentally [178]. However, mutations in the splicing factors may affect the splicing process or other steps in the maturation of pre-mRNA. These changes may not be reflected in the expression of the gene, but by alterations in the relative expression of isoforms of that gene.

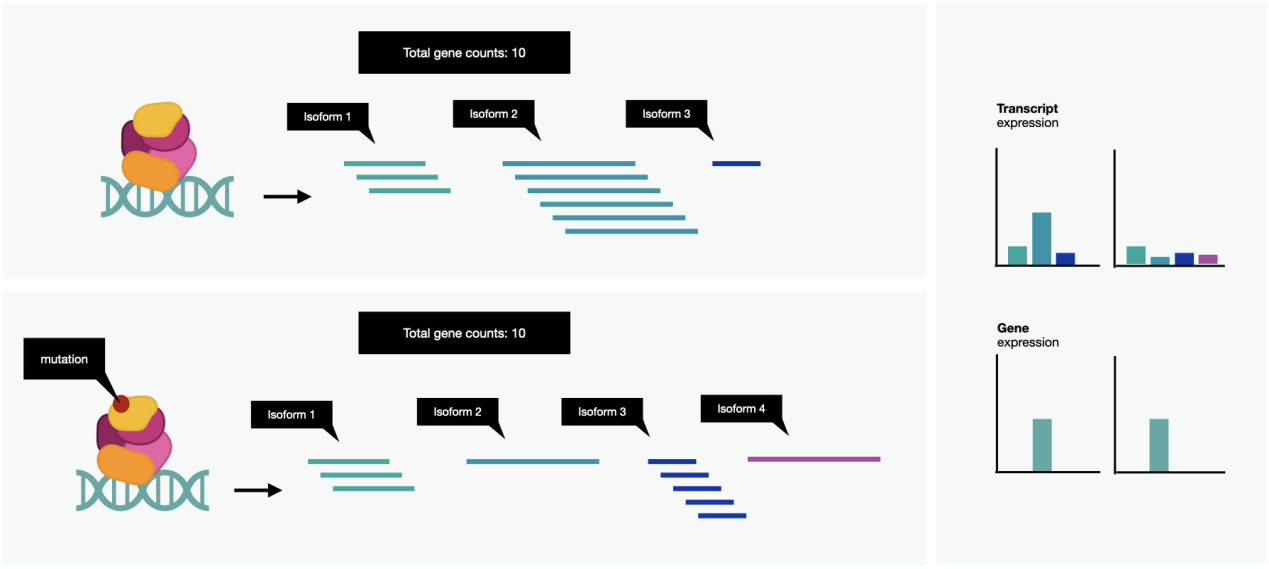

Figure 3.2: The image depicts the reason why often using gene-level p-values may overlook significant changes in the transcriptome.

Therefore, to study the effect that mutations in the splicing factors may have in the transcriptome, studying the relative expression of transcripts is a more appropriate approach. Nevertheless, it is problematic to use transcript-level p-values with tools designed to obtain biological insights from statistical information. Most tools to perform gene set analysis need gene-level p-values. Thus, we needed to aggregate the transcript p-values to obtain gene-level p-values that would reflect the changes seen in the transcriptome. *Yi et al (2018)* [178] suggested a new framework to translate transcript p-values to gene p-values using the Lancaster method [88]. This method proposes the weighted aggregation of transcript-level p-values, weighting the p-value according to the expression level of the transcript.

$$T = \sum_{i=1}^{K} -2log(p_i) \tag{3.2}$$

$$T = \sum_{i=1}^{K} -\phi_{wi}^{-1} log(p_i) \tag{3.3}$$

In order to weigh the transcript-level p-values generated from the linear model, we calculated the mean expression for each transcript for both cell lines with defects in the splicing factor and cell lines with intact splicing factors. Then we used the Lancaster method to aggregate the transcript-level p-values from the same gene, generating the gene-level statistics.

3.5.3 Gene-set analysis

Gene set analysis (GSA) is a method used to incorporate existing biological knowledge into expression data [156]. In GSA, the genes are clustered into gene sets based on shared properties that are usually defined by a reference database [106]. The gene sets are analyzed as a whole to determine whether such properties are of interest for the phenotype under study [29]. In other words, it tries to identify sets of biologically related genes involved in, for example, complex human diseases [106], [97]. It facilitates interpretation since it reduces the dimensionality - from genes to aggregations of genes (gene sets) - and it facilitates the detection of patterns and effects that arise from coordinated smaller changes.
To perform the GSA, we used the software tool PIANO [156]. PIANO is an R package that offers a platform to perform gene set analysis using several statistical methods [156]. The tool uses the gene-level statistics derived from the linear model and Lancaster aggregation: the unmodified (not corrected) aggregated p-value and log fold-change, not including information about directionality. Therefore, significant results indicate that the gene set is affected by differential expression, not specifying whether the gene is up- or down-regulated [156]. The p-values do not need to be adjusted before running the GSA because the algorithm includes a step performing FDR correction. To define the gene sets to analyze, we downloaded several curated gene sets from KEGG [74] and the hallmark gene sets from MSigDB [99]. We used the hallmark gene sets because they represent well-defined biological states or processes. These gene sets were generated using a computational methodology that identifies gene set overlaps and retains those genes that display coordinated expression, therefore reducing noise and redundancy and providing a better delineated biological space for GSA [99].

3.5.3.1 Gene statistics methods

We performed the GSA using two different statistical methods:

- Fisher's test [156], [49] In this methodology, for a given gene set i, which contains n genes with p-values p_1, p_2, ... p_n, the gene set statistic S_i is calculated according to the following equation:

$$S_i = 2\sum_{j=1}^{n} -ln(p_j) \tag{3.4}$$

  In this case, for our gene set of n genes significance was calculated by gene sampling. Gene sampling methods aim to identify those genes enriched with differential expression signals [114]. In gene sampling, the significance (S_i) of the gene set score is evaluated by comparing it to the scores of randomly assembled sets. The random sets are subsets of all genes under study [105]. The method presents the advantage of not being dependent on the number of samples, therefore being suitable for GSA of small samples [105]. However, a stringent correction for false positives is necessary because genes within a gene set are not sampled independently: many of the genes included in a gene set take part in the same biological pathways, which means that they are corre-

lated to each other to some extent [114], [105].

- Reporter features [156], [122] Given a gene set i, containing n genes with p-values p_1, p_2, ... p_n, the algorithm calculates the gene set statistic S_i as follows:

$$S_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\theta^{-1}(1 - p_j)) \tag{3.5}$$

  Where θ^{-1} is the inverse normal cumulative distribution. The inverse cumulative distribution function gives the value associated with a specific cumulative probability. The gene set p-value is calculated using the normal cumulative distribution function:

$$p_{S_i} = 1 - \theta(S_i) \tag{3.6}$$

  The significance estimation is calculated by normalizing the gene set statistic S_i with the mean and standard deviation of the background gene set statistics for that particular gene set. The background gene set statistics are calculated by randomizing the genes labels and recalculating the gene set statistic many times. The gene set p-value can then be calculated back from the normalized Z-scores by using the normal cumulative distribution function.

$$p_{S_i} = 1 - \theta(\frac{S_i - \mu_n}{\sigma_n}) \tag{3.7}$$

  This algorithm reports back all directionality classes. The non-directional class is obtained by using all the unmodified p-values, which means that a mix of up- and down- regulated genes can be considered as a significantly changed gene set, as long as the gene set as a whole is significantly changed.

Since we are working with aggregated p-values generated with the Lancaster aggregation, we lack a proper log fold-change value. Therefore, we can calculate the significance of the gene set but we cannot generate information regarding the directionality of the changes. To address this, we selected some gene sets of interest, used the labelled cell lines according to presence of mutations in the spliceosome machinery and aggregated the transcripts into two categories: alternative and canonical. The canonical transcript was extracted from the ENSEMBL database for each gene, using the suite BioMart [39].

3.5.4 Shannon Entropy

A limitation to classical differential expression is its inability to detect altered levels of expression, namely changes in the expression profile with low or no change in mean expression [167].
Shannon entropy (SE) and coefficient of variation (CV) are metrics that measure the variability of numerical data [168]. Differential variability analysis has been applied before to the study of gene expression in human disease [68]. As an example, increased entropy in gene expression has been observed as a characteristic of cancer, [174], [9] with entropy levels correlating with different cancer stages [19]. The goal

of these metrics is to detect changes in the expression profile even if the mean change in level expression is too low to be detected by differential expression [167]. The difference is that both entropy and variation deal with normalized expression variability, while differential expression detects differences in the mean expression [167]. We used the package EntropyExplorer to calculate the differential Shannon entropy (DSE) and differential coefficient of variation [168]. The measurement of SE allows for the quantification of information content associated with the likelihood that a given node has a given connectivity value in the network of interest [19]. The formula to calculate the Shannon entropy can be seen below [19].

$$H = -\sum_{k=1}^{n} p(k) log_2(p(k)) \tag{3.8}$$

3.6 Predicting novel splice-altering mutations using SpliceAI

3.6.1 Running the deep learning algorithm

We used the DepMap dataset containing somatic mutations as input for the deep learning network. Due to the large file size, the dataset was split according to the chromosome in which the mutation occurs, to facilitate uploading the data to Hebbe. Before running the deep neural network, the dataset was parsed into a VCF file to meet the requirements of the algorithm. More details can be found in the Appendix B.

3.6.2 Evaluating the performance of SpliceAI

To estimate the performance of the deep learning algorithm, we measured its ability to classify the mutation in two classes: "splice-altering" or "not splice-altering". As mentioned before, SpliceAI outputs four probability scores for a mutation, evaluating the likelihood of the event to alter either the donor or acceptor site by creating new ones or destroying existing sites. The four scores are: Acceptor gain, Acceptor loss, Donor gain and Donor loss.
According to Jaganathan et al. [72], a mutation can be confidently considered as splice-altering when its SpliceAI score is above 0.5 - and this is the score they recommend as a cutoff. Therefore, we considered putative splice-altering mutation all the variants for which SpliceAI gave a score of 0.5 or above. To evaluate the performance of the algorithm, we compared its classification of mutations in splice-altering or not splice-altering with the annotations made by DepMap in their somatic mutations file. It is noteworthy that we can only estimate the number of true positives and false negatives, since often SpliceAI finds that a mutation is likely to affect a splice site downstream or upstream of the somatic mutation that was reported by DepMap, and we lack experimental data for such positions. For the evaluation, we used a Receiver Operating Characteristic (ROC) curve. A ROC curve is built using a confusion matrix or contingency table with four categories [28]:

- True Positives: cases correctly labelled as positives. In this case, it would be mutations correctly considered to alter the splicing process by SpliceAI.
- False Positives: negative cases incorrectly labelled as positives. Here, it would be mutations classified as splice altering that in reality have no influence on the process.
- True Negatives: negatives correctly labelled as negative.
- False Negatives: positve cases incorrectly labelled as negative.

Using the contingency table, a ROC plot is drawn by plotting the false positive rate, or inverted specificity (in the X-axis) against the true positive rate, or sensitivity (in the Y-axis) for a number of candidate decision thresholds between 0 and 1 [12]. The sensitivity, or true positive rate (TPR), and specificity are calculated as follows:

- $Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives}$;
  where the TPR measures the fraction of positive cases being correctly labeled [28].
- The false positive rate (FPR) can be calculated as the inverted specificity, where specificity:
  $Specificity = \frac{True\ negatives}{True\ negatives + False\ positives}$;

  and therefore FPR:

  $False\ Positive\ Rate = 1 - Specificity.$
  The FPR measures the fraction of negative examples misclassified as positive [28].

ROC curves present the limitation of being overly optimistic evaluating the algorithm's performance when the class distribution of the data is highly skewed [28]. In our case, due to the nature of the data, it is fair to assume that the distribution will be skewed toward negative cases - mutations that will not affect splicing. To overcome this problem, we can evaluate the performance of SpliceAI with yet another tool: precision-recall (PR) curves. PR curves are also built by using a confusion matrix, but the metrics to plot differ. In the PR space, one plots what we call Recall - which is the same as TPR - on the X-axis and Precision on the Y-axis.

- $Recall = \frac{True\ positives}{True\ positives + False\ negatives}$
- $Precision = \frac{True\ positives}{True\ positives + False\ positives}$;

  where precision measures the fraction of cases being classified as positive that are truly positive [28].

However, plotting these curves does not give an immediate measurement of the performance of the classifier. To obtain an accurate evaluation, we can use the Area Under the (ROC/PR) Curve (AUC) as a metric of the performance. The AUC accurately represents the probability that a randomly chosen positive example of the evaluated data is correctly rated - or ranked - above a randomly chosen negative example [12].

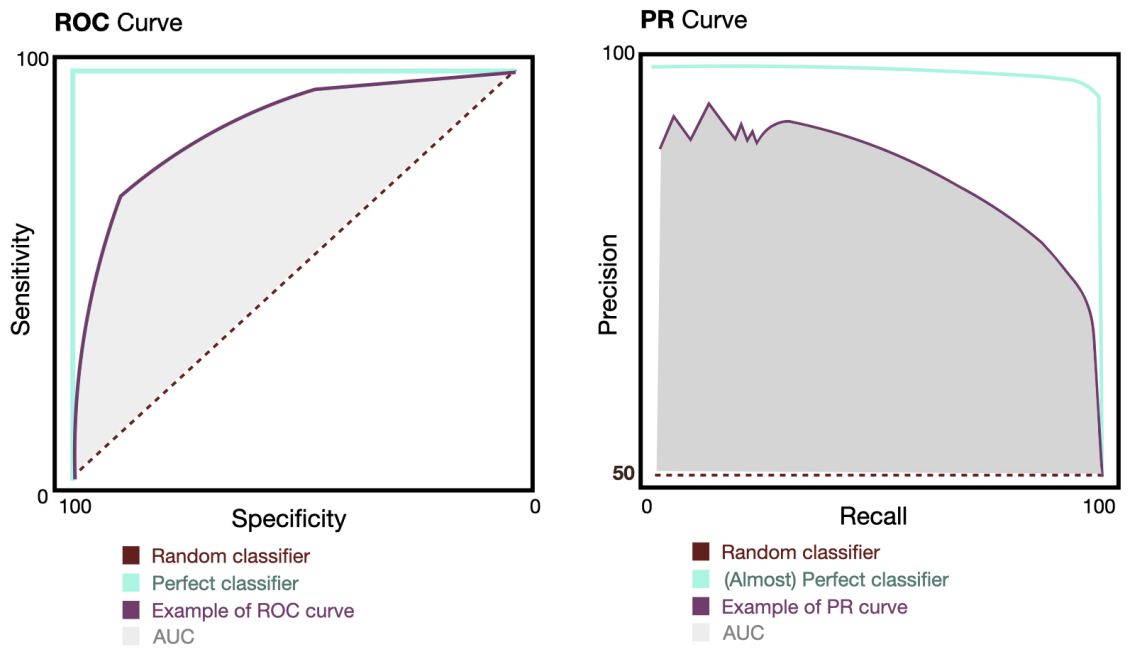

Figure 3.3: Example of a ROC and a PR curve. The curves do not correspond to the data from SpliceAI. The figure depicts a curve corresponding to a perfect classified (blue line), a curve corresponding to a classifier with no prediction power, or random classifier (dotted red line), and an example of what a typical ROC and PR curve with some prediction power may look like (purple line). The gray area represents the Area Under the Curve (AUC) score.

The evaluation of SpliceAI's performance using ROC and PR curves was done using the R packages pROC [133] and PRROC [78].

3.6.3 MSK-IMPACT

MSK-IMPACT stands for *Integrated Mutation Profiling of Actionable Cancer Targets*, a research project carried out by the Memorial Sloan Kettering Cancer Center. The researchers developed a hybridization capture-based NGS panel that performs an integrated mutation profiling of actionable - meaning that they can be targeted with drugs - cancer targets. In total, MSK-IMPACT analyzes 505 genes that play a critical role in the development and behavior of tumors. The test can also detect all protein-coding mutations, copy number (CN) alterations and some promoter mutations and structural rearrangements in cancer-associated genes [180]. The overarching goal of the panel is to facilitate the identification of patients for enrollment in genomically guided clinical trials [180].
In this project, we used the list of clinically relevant somatic mutations generated by the MSK-IMPACT project to narrow down the analysis of functional screening datasets to genes that are clinically relevant. The list of targeted genes generated by the study [180] was retrieved from the cBio Cancer Genomics Portal [17], [52] where it can be accessed under the name MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017) [2].

[2]MSK-IMPACT data can be accessed here

4

Results

4.1 Mutational landscape of splicing factors

We used the spliceosome and splicing factors list compiled by Seiler et al. [142] to explore the mutation landscape of the splicing factors in the somatic mutation dataset provided by DepMap. The accumulated mutations per cancer type are normalized for the number of cancer cell lines studied for each disease.

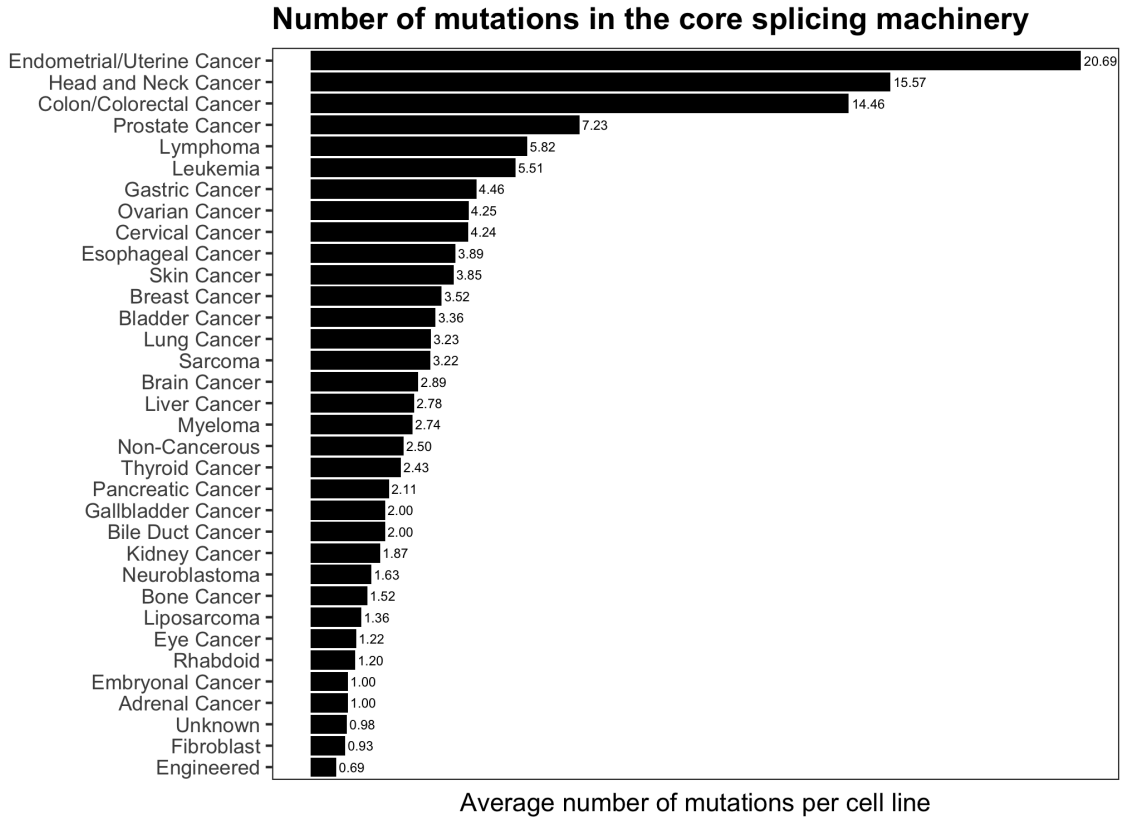

Figure 4.1: Number of mutations in the core spliceosome across different cancer types normalized for the number of cell lines studied.

In general, we see that the spliceosomal mutational load varies greatly across cancer types, and that it correlates well with the number of mutations across the entire genome (total mutational load) - not only splicing factors - per cancer (see Appendix A). We also created a binary heatmap to show the distribution of the deleterious mutations in the splicing factors across different cell lines - for details about the heatmap construction see Section A.3.

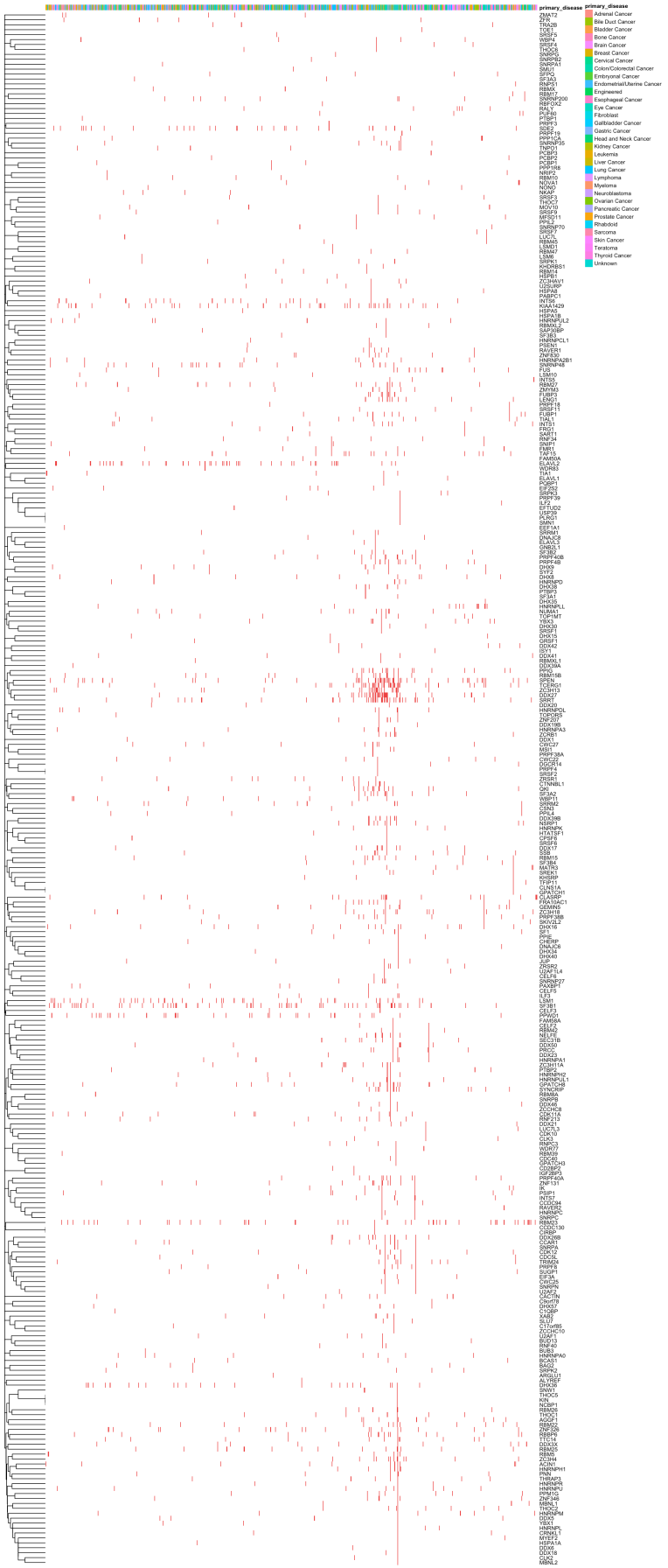

Figure 4.2: Heatmap showing the deleterious -not SNP- mutations occurring on the splicing factors across the studied cell lines.

4.2 Gene expression patterns in spliceosome deficient cancers

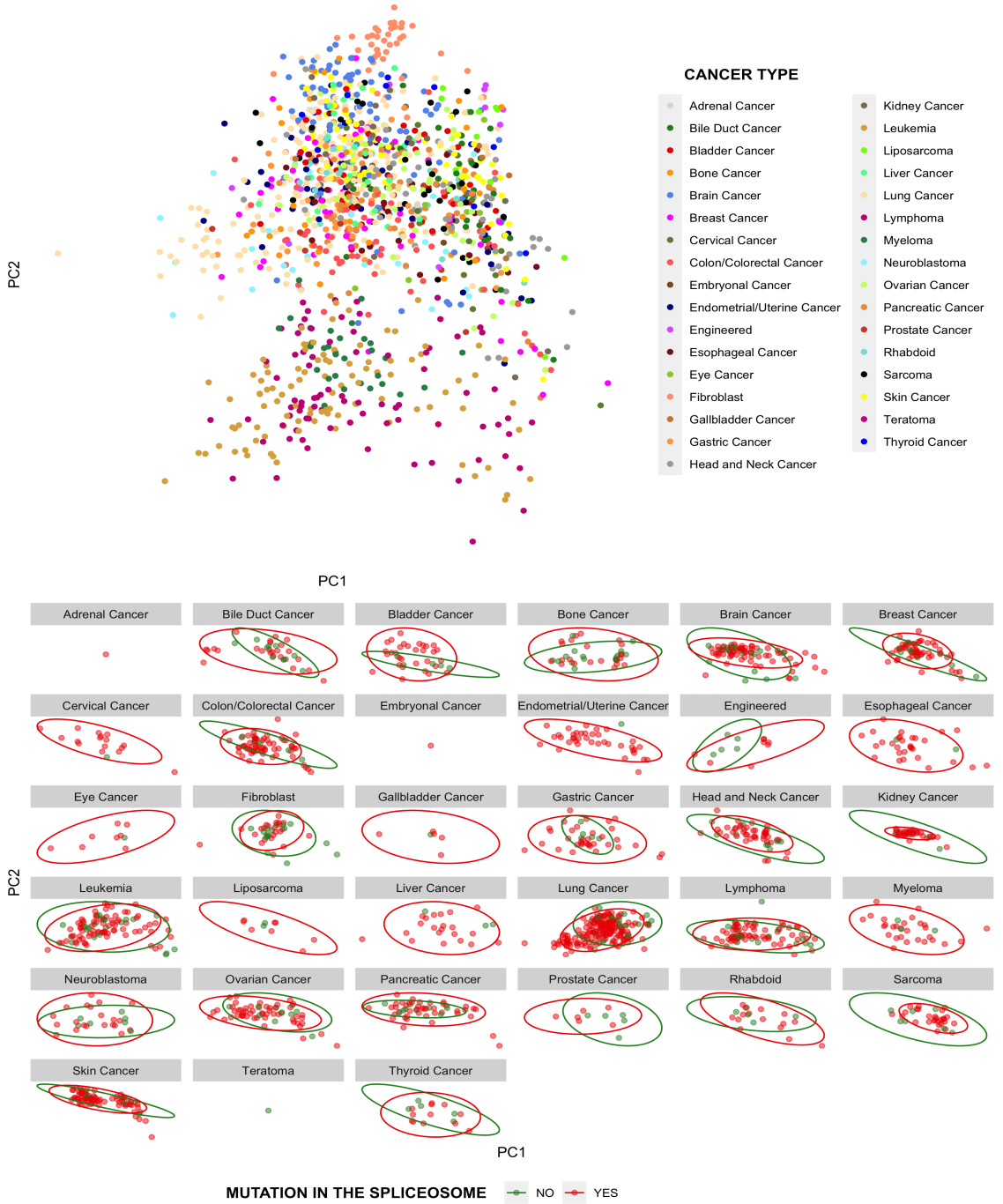

Figure 4.3: PCA performed on the gene-level expression data (TPM), labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the core spliceosome machinery

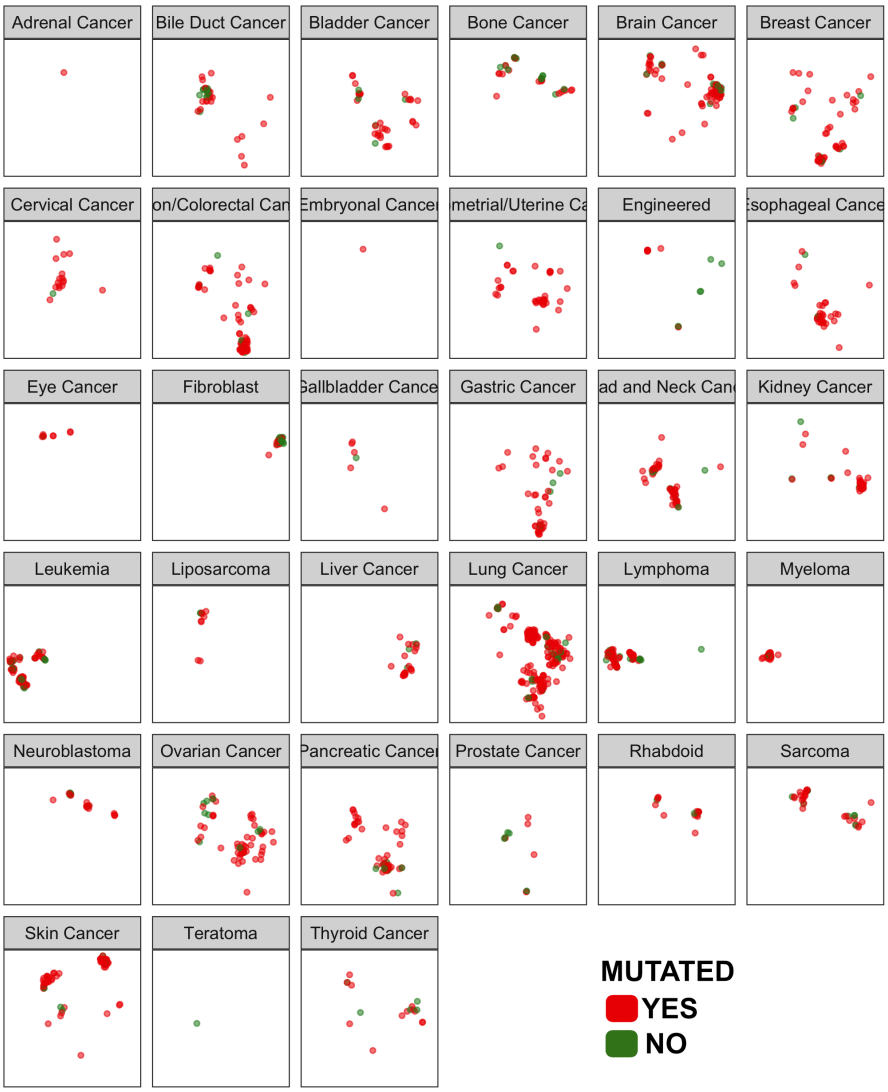

Figure 4.4: tSNE performed on the gene-level expression data (TPM), labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the core spliceosome machinery

From the PCA, we can see that there is not a clear separation of cell lines based on cancer type (top part of figure 4.4). When applying the dimensionality reduction technique to each cancer type labelling the cell lines according to the state of their spliceosome, we can see a slight separation between both classes for some of the malignancies - for example for lung cancer, bladder cancer or breast cancer. However, this difference is not clear enough to conclude that mutations in the spliceosome are a major source of variability in the expression data. In the t-SNE plot, we can see separation in subgroups for some of the malignancies (such as pancreatic cancer, ovarian cancer, bile duct cancer, or sarcoma), but this separation is again not driven by the mutations in the spliceosome. The application of these dimensionality re-

duction techniques for the cell lines with mutations on the splicing factors from the other categories show similar results. The plots can be found in the Appendix A. Overall, it seems that there are no large differences at the gene level expression between cell lines with an altered spliceosome and cell lines that maintain the complex and regulatory elements intact.

4.3 Transcriptional differences related to spliceosome deficiencies in cancer

4.3.1 Differential expression

Results for the GSA performed using PIANO on the aggregated p-values obtained using the Lancaster method. We performed different analysis for cell lines with mutations in the core spliceosome, cell lines with mutations in the non-core spliceosomal proteins and splicing factors and for the cell lines with mutations in genes categorized as "other".

Mutations in core spliceosome proteins - Hallmarks gene set
Gene Set Analysis - Fisher

	Genes	p-value	adj. p-value
G2M CHECKPOINT	196	0.0001	0.0012
E2F TARGETS	200	0.0001	0.0012
MYC TARGETS V1	200	0.0001	0.0012
OXIDATIVE PHOSPHORYLATION	198	0.0001	0.0012
UNFOLDED PROTEIN RESPONSE	111	0.0011	0.0110
PI3K AKT MTOR SIGNALING	83	0.0032	0.0267
MYC TARGETS V2	58	0.0058	0.0414

p-values adjusted for multiple testing using Bonferroni correction

Mutations in non-core spliceosome proteins - Hallmarks gene set
Gene Set Analysis - Fisher

	Genes	p-value	adj. p-value
UNFOLDED PROTEIN RESPONSE	110	0.0001	0.0017
E2F TARGETS	200	0.0001	0.0017
MYC TARGETS V1	200	0.0001	0.0017
G2M CHECKPOINT	196	0.0003	0.0037
OXIDATIVE PHOSPHORYLATION	198	0.0004	0.0040
MYC TARGETS V2	58	0.0006	0.0050

p-values adjusted for multiple testing using Bonferroni correction

Mutations in spliceosome components not classifed as core nor non-core - Hallmarks gene set
Gene Set Analysis - Fisher

	Genes	p-value	adj. p-value
MYC TARGETS V2	58	0.0007	0.0350

p-values adjusted for multiple testing using Bonferroni correction

4.3.2 MYC targets

The gene set MYC targets consistently appears as differentially expressed across the three different categories. Both MYC targets gene set contain a subgroup of genes regulated by MYC. MYC is a transcription factor that regulates the processes of cell growth, differentiation, metabolism and death, and appears deregulated in more than 50% of human cancers [81]. Furthermore, such deregulation has been associated with poor prognosis and patient survival. Although it may seem an obvious therapeutic target, due to its structure MYC has been rendered "undruggable" [51].

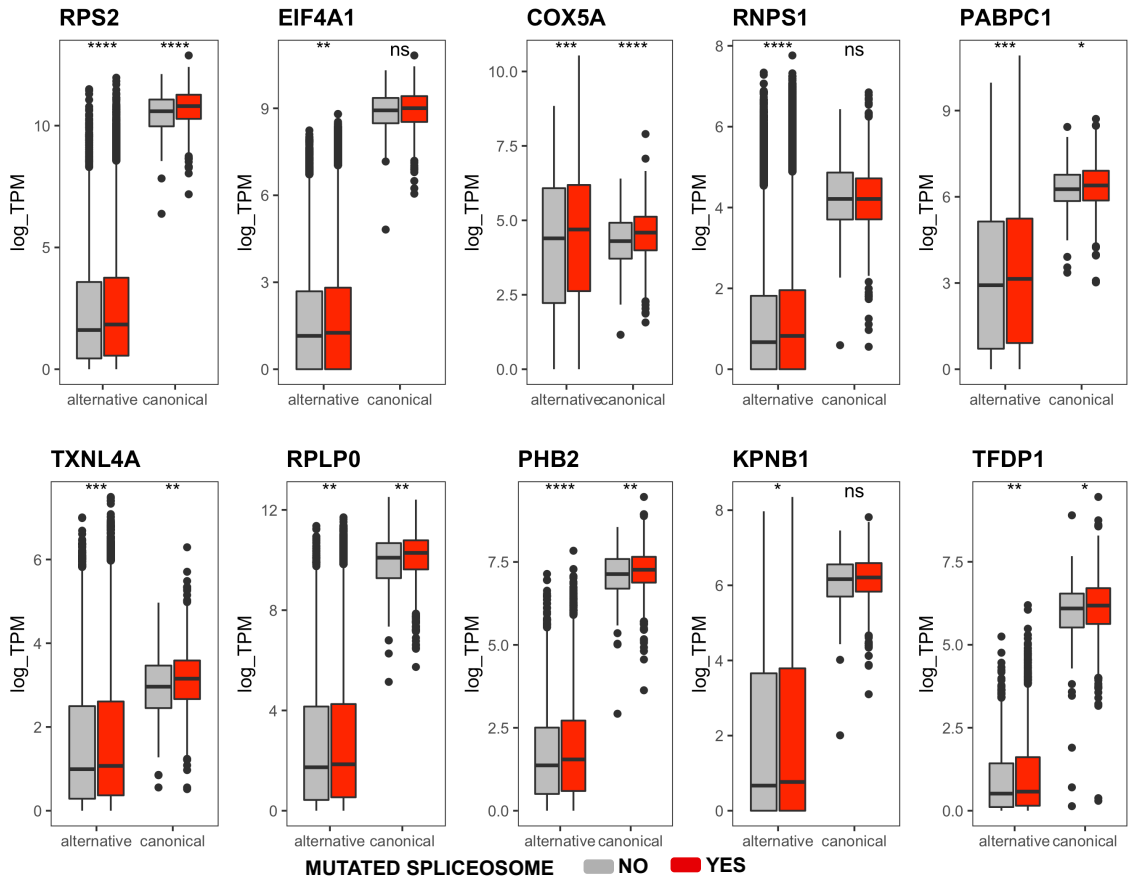

Figure 4.5: Ten most significant differentially expressed genes when considering cell lines mutated for core splicing factors vs. wild-type cell lines in the MYC targets gene set

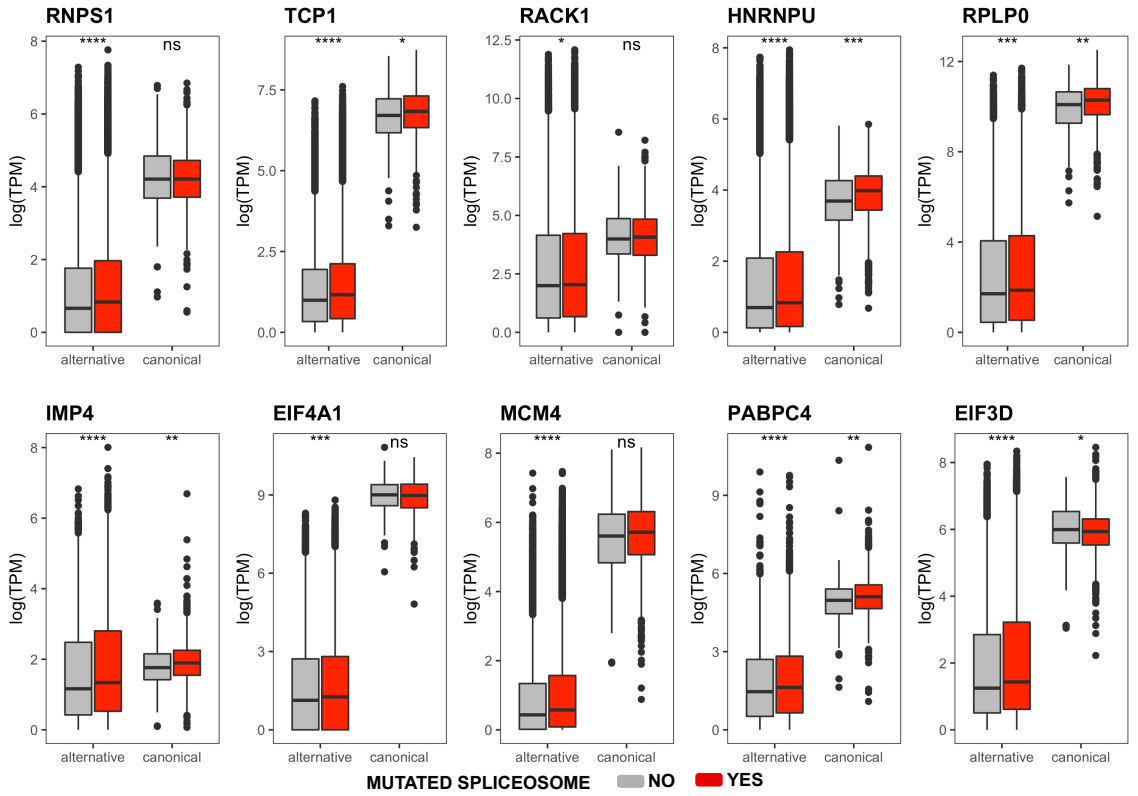

Figure 4.6: Ten most significant differentially expressed genes when considering cell lines mutated for non-core splicing factors vs. wild-type cell lines in the MYC targets gene set

The plots shows an aggregation of the transcripts into two categories: alternative and canonical transcripts as depicted in Subsection 3.5.3. We can see a general trend of increased expression for those transcripts in cell lines with a mutated spliceosome. In general, this trend is sharper for the alternative transcripts than for the canonical one. Many of these genes are related to the processing of mRNA in different ways. For example, IMP4 is necessary for pre-18S ribosomal RNA processing [59] and EIF3D is necessary for the initiation of translation [93]. Some of these genes are splicing factors as well: RNPS1 and PABPC1 are part of the non-core spliceosome machinery. RNPS1 is part of a post-splicing complex that initiates nonsense-mediated mRNA decay (NMD), a process that clears truncated transcripts [135]. PABPC1 binds to the poly(A) tail and promotes ribosome recruitment and the initiation of translation, and is also required for the first steps of NMD [86].

4.3.3 Shannon Entropy

The Shannon Entropy was calculated for each transcript using the transcript expression data. The results consistently show higher entropy for the expression of transcript in mutated cell lines. For this calculation we considered the mutations in any splicing factor or spliceosomal gene regardless of category.

Increased entropy in the expression of mutated cell lines

***

SYMBOL	MUTATED	WILD-TYPE
Gene 1 (Transcript 1)	0.93	0.80
Gene 1 (Transcript 2)	0.90	0.82
Gene 2 (Transcript 1)	0.89	0.77
Gene 3 (Transcript 1)	0.95	0.81

Figure 4.7: Example of calculation of the Shannon Entropy metric - note that the values are only representative and do not correspond with the actual results.

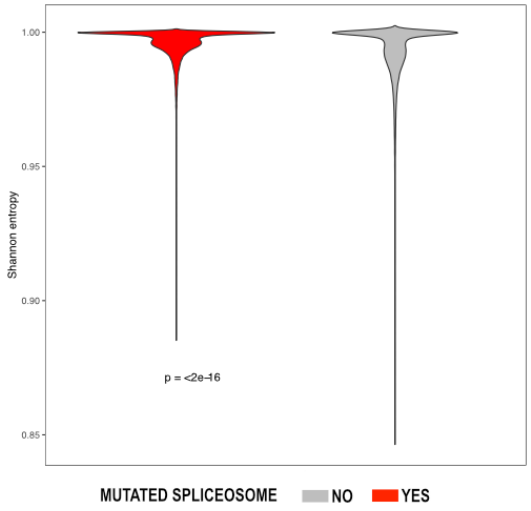

Figure 4.8: Comparison of the distribution of the calculated Shannon Entropy per gene between mutated and wild-type cell lines. Significance assessed using Wilcoxon test. p-value $<$ 2e-16.

The higher entropy for cell lines with dysfunctional or mutated spliceosome indicates that a wider variety of transcripts are expressed in those cells than compared to spliceosome wild-type cell lines. This is consistent with the results observed from the differential expression analysis at transcript level, which show that there are transcripts that are significantly up- or down-regulated between the two types of cell lines (mutated and wild-type) and strengthens the idea that mutations in the spliceosome may drive subtle but network-wide changes that are not detectable with classical differential expression approaches.

4.4 Predicting individual mutation-associated splicing events using SpliceAI

4.4.1 Performance of SpliceAI

SpliceAI calculates the probability that a particular mutation will affect the splicing process. More specifically, the deep learning tool predicts four probabilities, depending on the likelihood of the considered mutation to disrupt or create a donor or acceptor site. Since we are only interested in knowing whether the algorithm considers the mutation able to alter splicing, we extracted the maximum probability of the four that are calculated and used it as a predictor variable. If the maximum probability is 0.5 or larger, the mutation is considered to have an effect on the splicing process. To understand the prediction capability of SpliceAI, we decided to compare the prediction with the classification that CCLE had given to the mutation. CCLE classifies mutations occurring in known splice consensus sites as "Splice_site". Therefore, we decided to compare the predictor variable with the given annotation to test the classification performance of SpliceAI. Since we are facing a binary classification problem, we used ROC and PR curves as metrics to evaluate the algorithm.

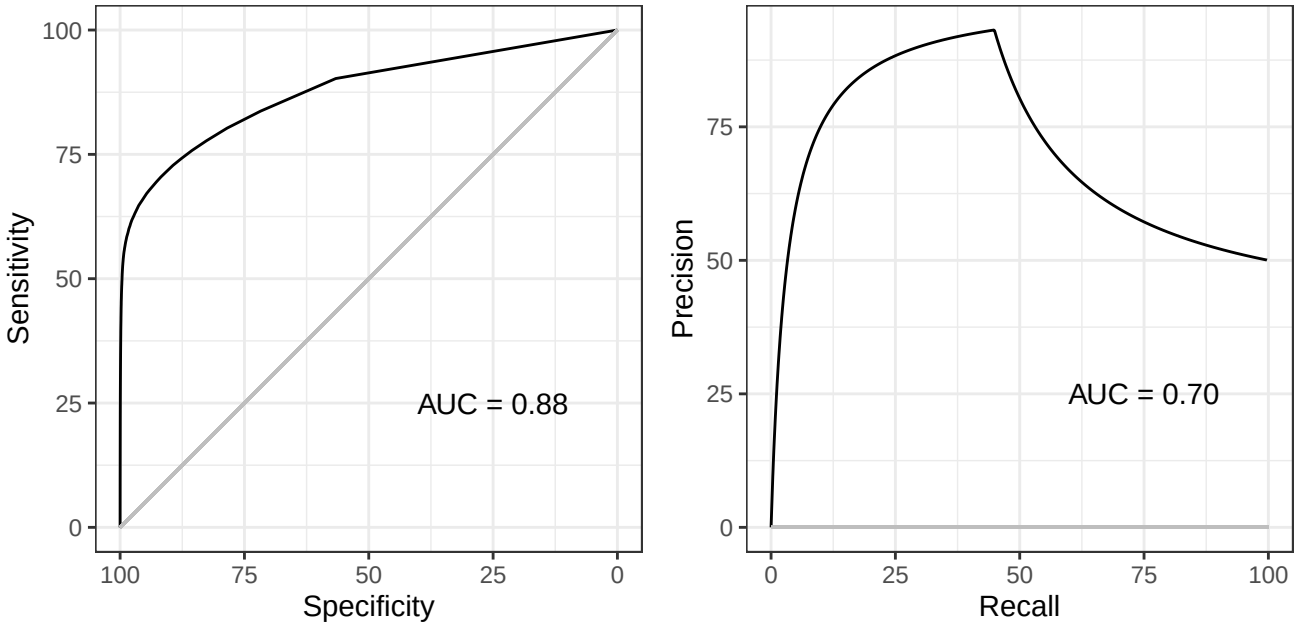

Figure 4.9: Left graph displays the ROC curve when comparing the classification of mutations made by SpliceAI with the annotation provided by CCLE. Right graph shows the PR curve on the same data. AUC calculated for both curves.

In addition to the curves, we calculated the AUC as an additional metric of performance. In both cases, the AUC is above 0.5 - an AUC (or a ROC or PR curve) with value 0.5 or similar would mean that the algorithm is not better at classifying the mutations than randomly assigning a class to them. In our case, as mentioned in

Section 3.6.2, the data is imbalanced: we expect a larger number of mutations to not affect the splicing than to have an effect on this process. Therefore, the PR curve is a better metric since to a certain extent is able to correct for such imbalance. It is worth mentioning that the predictive capability of SpliceAI is probably underestimated with this approach, due to the nature of the data from DepMap. From their annotations, we can only consider as splice-altering those mutations actually occurring in the splice site. However, SpliceAI considers a large window of nucleotides around the somatic mutation and can identify disrupting events that may occur at a different position. Hence, some of the mutations predicted to be splice-altering by SpliceAI that were not annotated as so by DepMap are probably not False Positives (as we are considering them when calculating the ROC and PR curve), but having an effect elsewhere in the genome. However, with the information that we possess at the moment it is impossible to assess these cases.
To obtain a visual representation of the classification, we can observe the distribution of the probabilities calculated by SpliceAI for the mutations that were given the category of Splice Site (assumed to be splice altering mutations) by DepMap and those that were not.

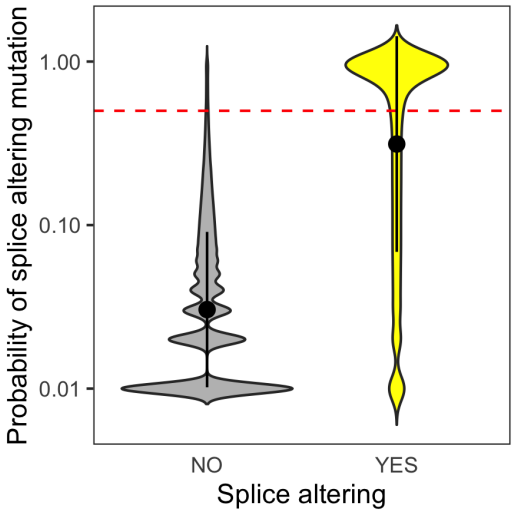

Figure 4.10: Distribution of the probabilities calculated by SpliceAI in mutations labeled as being splice altering or not. Y-axis is displayed in log-scale. The dashed line shows the probability score of 0.5 calculated by SpliceAI.

In the Figure 4.10 we can see that most of the mutations labelled as not splice-altering (gray-colored in the figure) are given a low probability - some of these mutations that have a probability above 0.5 are explained by the splice-altering capability of some mutations that do not occur in splice sites and therefore were not annotated as such by DepMap, as has been mentioned above. In the case of the mutations that are classified as occurring in a splice site by DepMap (yellow-colored in the figure), some of them are clearly below the 0.5 threshold. This may be because the mutations may be silent, meaning that they do not have an effect on the phenotype of the cell. Lastly, it is clear that the performance of the algorithm varies across cancers - for example, the classification power for teratomas seems worse than

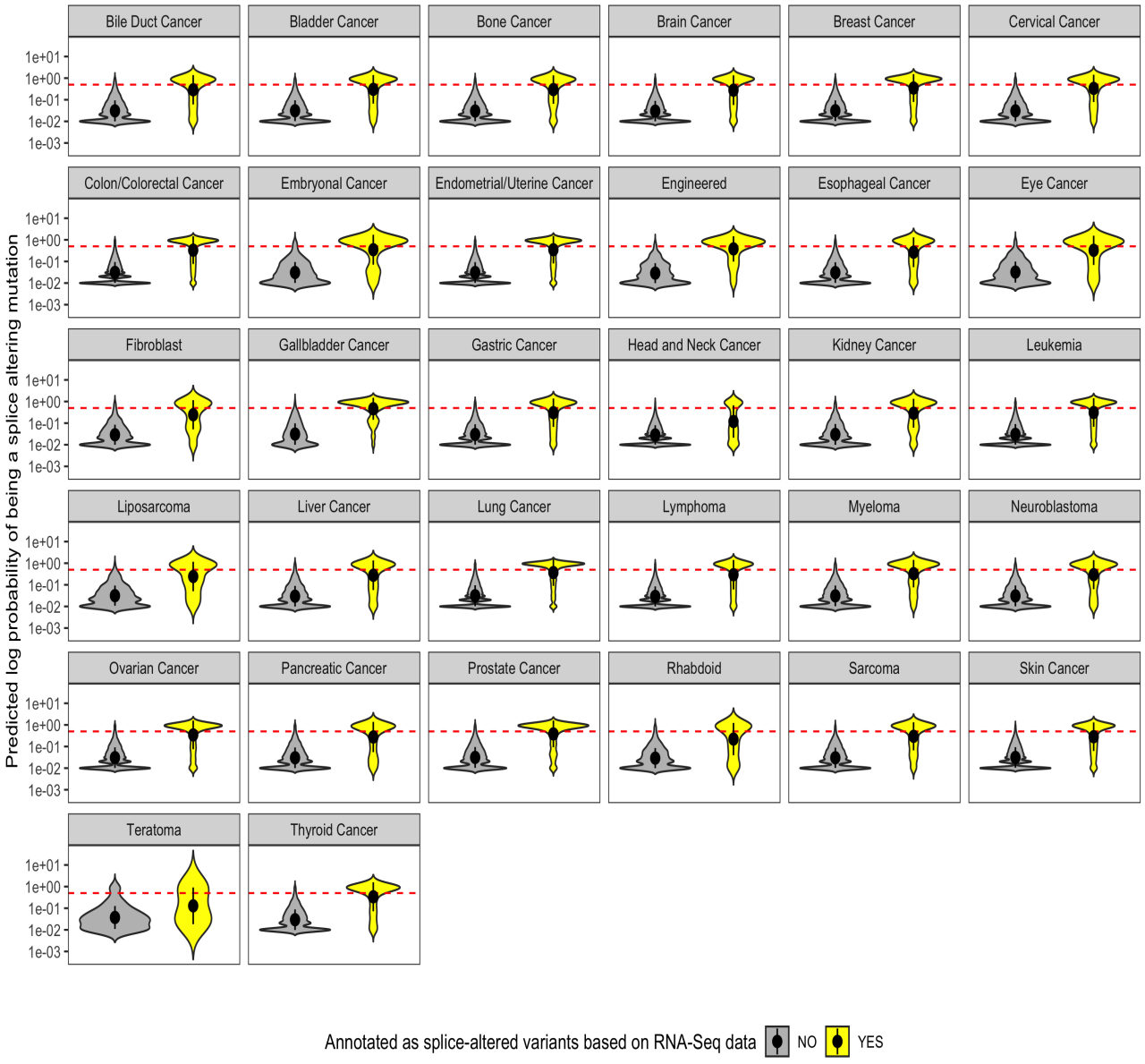

Figure 4.11: Distribution of the probabilities calculated by SpliceAI in mutations labeled as being splice altering or not, per cancer type. Y-axis is displayed in a log-scale. The dashed line shows the probability score of 0.5 calculated by SpliceAI.

for colorrectal cancer cell lines. Likely, this is due to the amount of cell lines studied for each of the malignancies. Only one cell line for teratoma was analyzed in the studied dataset, whereas for colorrectal cancer 79 cell lines are available (see Figure A.1).

4.4.2 Predicted variants

We fed over 1 million somatic mutations to SpliceAI. Of these, the neural network predicted slightly over 2% to alter the splicing process of a gene. The numbers can be seen in the Table 4.4.

Events considered	1293157
Events predicted by SpliceAI	28982
Off-target events	1665

Summary of the mutation events used as input for the algorithm and the output of the predictions.

Interestingly, we observed that in some cases, SpliceAI predicts that the variant will occur in a different gene than the one carrying the mutation causing the alteration - or in both the gene carrying the mutation and another one. We called these events off-target splicing. Almost a 6% of the mutations were predicted to cause off-target splicing by the algorithm (total number in Table 4.4). A complete list of the reported off-target variants can be found in the table C.1 in Appendix C. Given that some of the trans-splicing factors participating in the regulation of splicing can be located in genomic regions far from the actual splicing site, it is not surprising that some mutations may alter splicing events occurring elsewhere in the sequence.

The plot shows that most of the predicted splicing events do occur relatively close to the somatic mutation causing the event. However, we can see that the distribution is highly skewed in both tails, which shows that a minority of events occur in different areas of the genome.

To better understand the possible effects of the predicted splicing alteration, we wanted to know if the event would fall in an exon or an intron. For this, we used the transcript list from CCLE (all the transcripts that suffer somatic mutations) to extract the genomic positions of exon boundaries using BioMart databases [39]. Then, we used the package GenomicRanges [90] to create intervals for the exons and for the mutations predicted by SpliceAI (we simply add +1 to the position where splicing is predicted to be altered). The function findOverlaps detects when the alteration predicted by SpliceAI falls within one of the exons.

Most of the splice variants predicted by SpliceAI lie within intronic regions, see Table 4.5. Splicing mutations in deeper intronic positions are vastly under-reported [132], hence it is positive to see that SpliceAI is able to predict variants laying within introns as well.

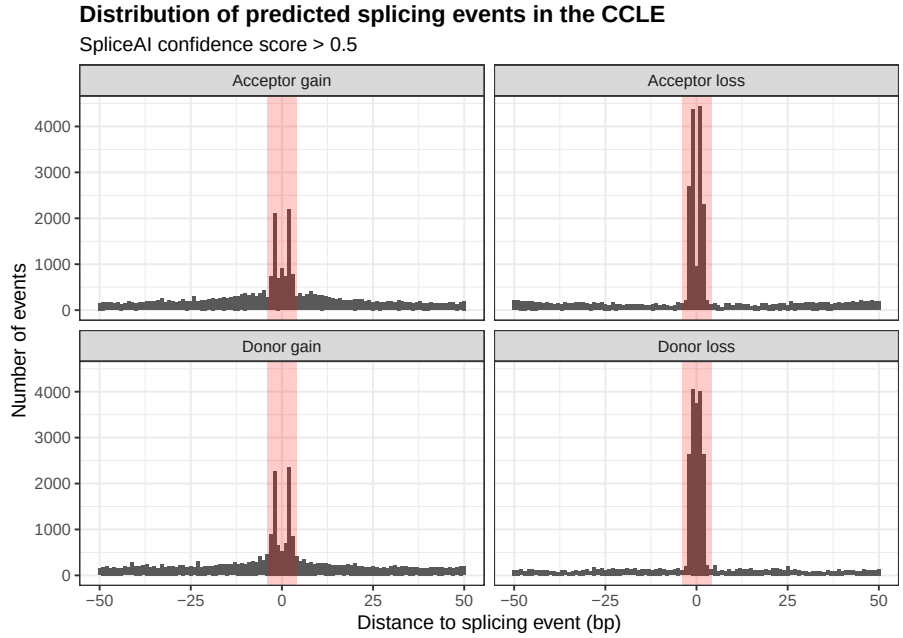

Figure 4.12: Plot showing the distance between the position in the genome were the somatic mutation reported by CCLE occurs and the position where the splice-altering event is predicted to happen according to SpliceAI.

Variants in introns	28117
Variants in exons	865

The table summarizes the number of SpliceAI predicted variants to fall within introns or exons.

4.5 Functional impact of the splice-altering variants

We were interested in assessing the impact of the altered splicing events in the function of the gene and the sensitivity of the cell line to drugs. To understand the effect of the splicing alteration in the phenotype of the cell line, we used the information contained in the CRISPR knock-out screening dataset (see Subsection 3.1.3) and the PRISM drug sensitivity screening dataset (see Subsection 3.1.4). To narrow down the study, we decided to prioritise the analysis of genes that have been shown to have clinical relevance in cancer. To achieve this, we used the MSK-IMPACT panel - see Section 3.6.3. For clarity, from here we will be referring as intact genes to those genes that were not predicted to suffer an alteration by SpliceAI. However, bear in mind that this does not mean that these genes are wild-type, since most of them are carrying a somatic mutation as reported by CCLE. The objective of this section is to compare the effect that a splice-altering mutation has vs SNPs or other type of mutations. For this analysis, we excluded those mutations that have been classified as insertions or deletions by CCLE, since these alterations are so disruptive that they usually result in loss of function of the gene.

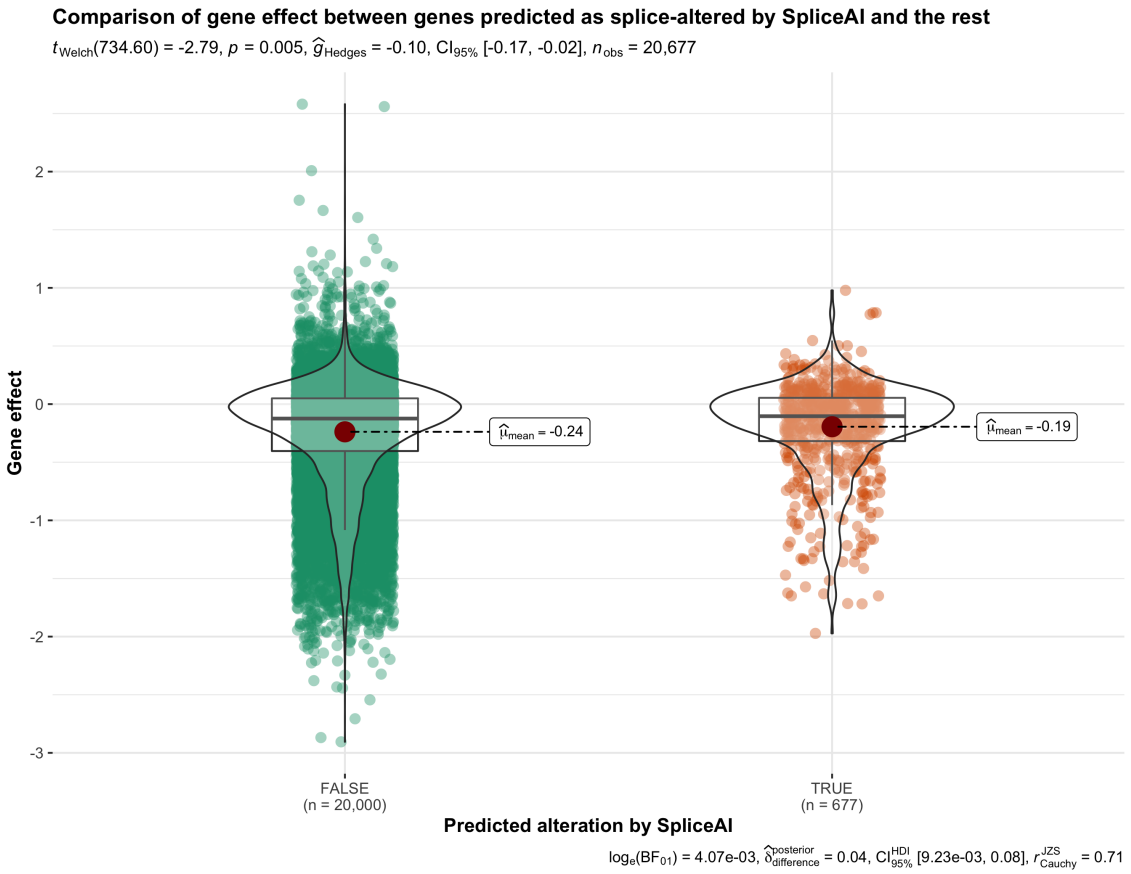

Figure 4.13: Comparison of gene effect for genes in the MSK-IMPACT panel that were predicted to suffer a splice-altering mutation vs. those that were not. We took a random subsection of the events to facilitate plottig and visualization and we excluded damaging mutations. The statistical significance was calculated using Welch's t-test. The p-value is 0.005.

In order to produce the plot 4.13, we first filtered the output of SpliceAI to only keep those events happening in genes present in the MSK-IMPACT panel. Then, we used the resulting list to label the CRISPR knockout dataset depending on whether the evaluated gene carried a splice-altering event or not. Since the resulting dataset was still considerably large, we randomly sampled the rest of the events and retained 20000 "intact" events and all the splice-altering mutations according to SpliceAI. The reason for this sampling was to reduce the number of events to plot while retaining a similar proportion of both events to the one in the complete dataset. The plot was created using the R package *ggstatsplot* [121].

The plot shows the effect on cell-line growth of knocking out a certain gene. A negative score means that depleting the gene has a negative impact on the growth, whereas a (relatively large)[1] positive score means that knocking out that gene improves growth. In the plot, we can observe that for the genes carrying splice altering events, the scores tend to be closer to 0 - we get very few positive scores and very few scores below -1.5. The scores' tendency to be closer to 0 could be an effect of the function of such genes being already impaired due to aberrant splicing events. Consequently, since these genes would lack their function or have their activity decreased, the effect of knocking them out would be less evident or negligible. Due to the nature of the data, we cannot simply compare the means - both means seem to be fairly similar, but the reason for this could be that the altered genes also present less positive gene effects, which lowers the mean value. This is consistent with the hypothesis that the splice-altering mutations have already disrupted the gene function. Thus, we compared the difference in gene effect between the "intact" and the altered genes using Welch's t-test for unequal variance, which gave a p-value of 0.005. It seems that the effect of knocking out the altered genes is significantly smaller than knocking out genes that have not suffered an altered splicing event, which could mean that these splice-altering mutations are more disruptive than mutations that do not alter the splicing patterns.
To confirm this, we looked at the gene effect of splice-altering mutations in individual genes - still focusing on those included in the MSK-IMPACT panel. The results can be seen in the Figure 4.14. We decided to show only those genes for which a predicted variant was present in more than 10 cell lines - in order to obtain more robust statistics. Interestingly, all the genes are well-known tumor suppressors [76], [8], [35], [183], [82], [125]. The general effect that we had observed in Figure 4.13 is repeated for the individual genes, and the differences are significant in all cases except for CDKN2A. The gene effect tends to be closer to 0 for those genes with mutations that affect the splicing, and in all cases (again with the exception of CDKN2A) we see that the positive effect in growth that the "intact" cell lines get from knocking out the gene is missing on those carrying a splice-altering mutation. This observation reinforces the hypothesis that these splice altering mutations disrupt the function of the gene. If we take TP53, arguably the best studied tumor suppressor gene, we see that upon knocking it out some of the cell lines without a splice-altering variant get a clear positive effect in growth. This effect is missing in those cell lines with

[1]Due to the screening pipeline and how CCLE/DepMap processes this kind of data, a slightly positive core is most likely attributed to noise

altering mutations, which could indicate that the variants of TP53 carrying these mutations were already unable to act as tumor suppressor genes.
Similarly, we were interested to see if there would be differences in the sensitivity to drugs between genes carrying variants that alter the splicing pattern and genes with splice-unrelated variants. For these analysis, we used the drug sensitivity data generated by the PRISM repurposing project, see Subsection 3.1.4. To facilitate the visualization, we decided to only show in this section the three cases for which a significant difference in sensitivity can be observed (Figure 4.15). The complete figure can be found in the Appendix C, Figure C.2. The three cases showed here are two drugs targeting the gene MET and another one targeting the gene PARP1. PARP1 is a gene involved in many cellular processes and usually overexpressed in human malignancies [169]. Furthermore, drugs inhibited PARP1 have been approved or are under clinical investigation, showing the potential of the gene as a therapeutic target [169]. In this case, the mutation seems to render the gene insensitive to the drug. On the other hand, MET is a well-known proto-oncogen, with some alterations and gene amplification events playing an important role in the onset of cancer [111], [152]. For this gene, mutations also seem to diminish the sensitivity of the cell line to the drug. Although the results seem indicate that splice-altering mutations can have an effect on the sensitivity of the cell line to drugs, the small number of cell lines with splice-altered mutations make a rigorous estimation difficult, and the small effects that we observe may disappear when taking into account a larger number of mutated cell lines.
Considering that some of the observed responses to drugs may be driven by other differences between cancer type, we also considered differences in sensitivity separating the cell lines by malignancy. The results can be seen in the Appendix C in Figure C.3. Here, we can see that the sensitivity seems to vary greatly from one cancer type to the other, but the main observation that alterations seem to lower the sensitivity to the drug remains. Here, we can also observe a clear case of misleading statistics: for Rhabdoid cell lines, we obtain a significant p-value but when observing the data points, it does not seem that there is actually a biologically relevant difference in the distribution - the result is likely a fluke due to the smaller number of cell lines considered.

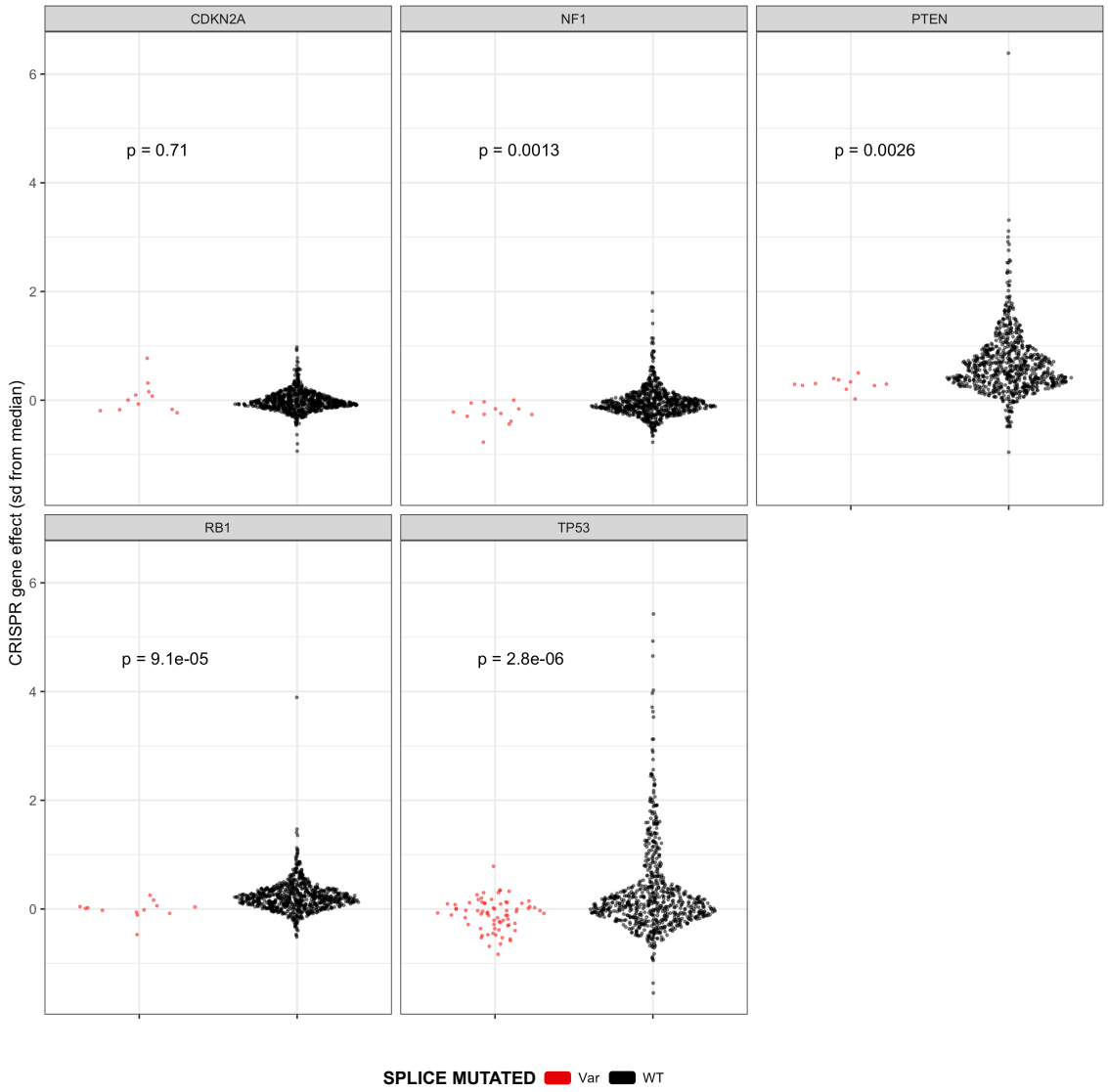

Figure 4.14: Comparison of the gene effect (calculated from CRISPR knockout screens) in cell lines carrying a predicted variant from SpliceAI vs cell lines that do not. Considering only genes present in the MSK-IMPACT panel. We considered genes that carried a predicted variant in at least 10 cell lines. In the context of this analysis WT means that the gene does not carry a variant predicted by SpliceAI. The significance was assessed using the Wilcoxon rank test.

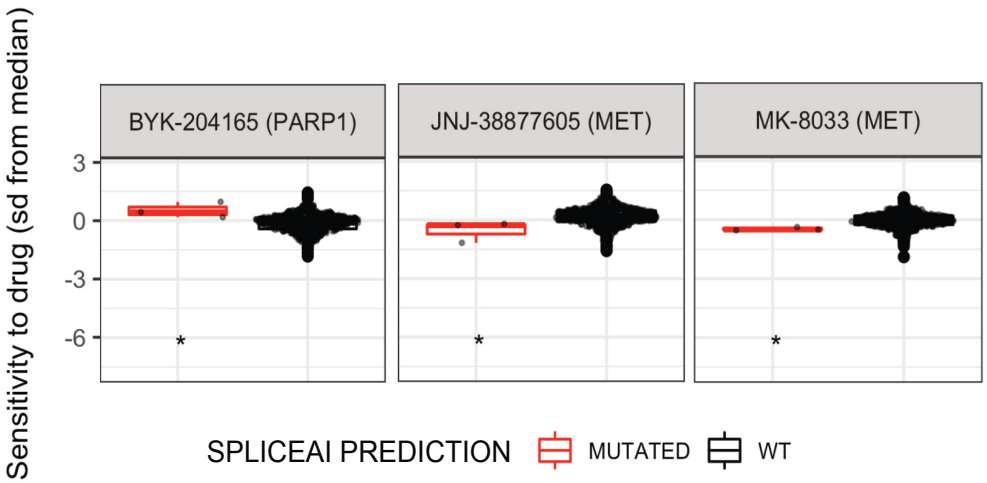

Figure 4.15: Difference in sensitivity to drugs from PRISM per gene comparing cell lines where the gene is predicted to be splice-altered according to SpliceAI vs. cell lines where SpliceAI did not predict a mutation. Figure showing the drug-gene pairs where a statistically significant difference was found (FDR corrected p-value $<$ 0.05). The significance assessed using Welch's t-test. In the context of this analysis WT means that the gene does not carry a variant predicted by SpliceAI.

5

Conclusion

Initially, we attempted to study the changes in expression due to mutations in the spliceosome performing a classical gene expression analysis. However, no significant changes in expression appeared. Considering the important role of the spliceosome and splicing factors in the process of splicing, alternative splicing and transcription in general, it seemed suspicious that mutations in these genes would not affect expression. We then considered that it was likely that changes in these processes would yield an altered expression in the individual transcript variants rather than differences in the mean expression of the genes. To test our hypothesis, we studied differential expression from the transcript level, using a linear model together with Lancaster aggregation to obtain p-values that would reflect the changes in transcript expression. To support this, we also calculated the Shannon Entropy for each gene to measure the variability in expression, since previous research has shown that an altered spliceosome correlate with higher entropy levels. Both analysis revealed that there are significant differences in the transcriptome expression levels when comparing cell lines with a defective spliceosome with cell lines with an intact one. Therefore, when studying the mutational landscape of spliceosome and splicing factor genes, it seems that the presence of mutations leads to changes in the transcriptional program of the cell. These changes seem to be subtle and span widely through the transcriptional network rather than being large, localized changes - therefore not always being obvious when using classical gene expression analysis techniques that rely on mean change of the gene expression. Our results show that the changes translate in an altered composition of the expressed transcripts, especially increasing the expression of alternative transcripts compared with the canonical ones. Nevertheless, the alterations caused by changes in the spliceosome machinery can be detected by adapting differential expression workflows to allow for analysis of transcriptional levels. Doing so, we were able to perform a GSA that revealed that several important pathways presented differentially expressed genes. Several of these pathways play an important role in oncogenesis. For example the PI3-AKT-MTOR signaling pathway modulates cell growth, survival, metabolism and angiogenesis, among other processes [144]. We decided to focus on the study of the differences in the genes part of the MYC targets gene set. The reason for this was that the gene set appeared significant even when considering only those splicing factors that were not defined as part of the core or the non-core spliceosome machinery so it seemed that members of these gene set were more consistently affected by mutations in any of the splicing factors. Furthermore, MYC has been shown to be widely deregulated in human malignancies [51], [104], [163] and it is hard to target therapeutically [18], [26]. When studying the most significantly differentially

expressed genes in the gene set, we observed that some (RNPS1 and PABPC1) were part of the non-core splicing machinery but also have a role in the process of nonsense-mediate mRNA decay (NMD), see Figures 4.5 and 4.6. NMD is a process that plays an extensive role in gene expression regulation, and its disruption has been shown to lead to many pathologies, such as cancer [116], [148]. PABPC1 acts by binding to the poly(A) tail of mRNAs and, when this binding occurs at specific positions, it can repress NMD. Therefore, hyperactive or promiscuous isoforms of PABPC1 could hinder NMD and allow for aberrant transcripts to be expressed [116]. On the other hand, the up-regulated expression of factors involved in NMD may actually indicate that this process is more active than usual, which could be a side-effect of aberrant spliceosome machinery producing high amounts of aberrant transcripts. In such case, inhibition of NMD may appear as a therapeutic strategy by increasing the load of aberrant transcripts in the cell and allowing the expression of neoantigens in cancer - therefore enhancing the effect of immunotherapy [148]. The actual role of NMD in cancer is context dependent [127] and therefore a more detailed analysis would reveal in which cases targeting NMD (perhaps by targeting genes such as PABPC1 or RNPS1) could be beneficial. Furthermore, we also tested for significant differences in drug sensitivity between cell lines with a deficient spliceosome vs cell lines that retained a wild-type spliceosome machinery. The results were not included in this report since we did not find any statistical significant difference for the drugs part of the PRISM project. Nevertheless, the spliceosome has proven to be a possible therapeutic vulnerability for cancers [69],[45], and there are some molecules, such as H3B-8800, that have shown promise in clinical trials [142]. Unfortunately, such drug was not available in PRISM repertoire - as PRISM is a drug re-purposing project, and most of the tested chemicals were not originally intended to treat cancer. Most likely, the lack of positive results in this case is simply because there is no drug in PRISM specifically targeting the spliceosome.

Regarding the prediction of new splicing variants using the deep learning tool SpliceAI, the performance of the algorithm is outstanding considering that it has not been specifically trained for cancer data. Furthermore, the methods we have to assess its performance have limitations - we can only compare the predictions against mutations annotated by CCLE, and we know they only annotate a mutation as altering the splicing process when it happens precisely in the consensus sequence. Therefore, it is likely that we are actually underestimating the accuracy of the algorithm. In order to improve our estimates we would need data annotated in more detail. An alternative would be to obtain access to the raw RNA sequence data ad process it in order to manually detect exon-intron splice junctions and differences in transcript expression due to mutations. However, this would be a complex and computationally demanding process. Nevertheless, the deep learning tool performs well and it can be used to confidently predict splice altering variants from RNA sequence data in the context of malignancies. More interestingly, the tool has shown to be able to predict variants outside the genes were the mutation is originally located, showing that is able to take into account surrounding context and revealing putative variants that could not be detected through the analysis of the genome, for example using whole genome sequencing (WGS) or similar techniques.

We did not set to estimate the relevance of these variances due to the relatively small number of off-target variables available - we need to keep in mind that only a small proportion of the studied mutations was predicted to be splice-altering by SpliceAI and an even smaller percentage showed to produce off-target events. Undoubtedly, further studies with more available data - or perhaps data collected specifically for the studies of these variants could better assess the clinical relevance of these events.

A further assessment of the capability of SpliceAI was provided by the analysis of the functional effect of the splice altering mutations in both effect on the cell lines growth (provided by the CRISPr screenings) and the sensitivity to drugs (estimated through the PRISM repurposing dataset). Specially in the analysis of the gene effect, it seems clear that those genes carrying a mutation altering splicing are rendered not functional more often than those carrying another kind of mutation, as evidenced by the lower variance of the gene effect and the values closer to zero (see Figures 4.13 and 4.14). This also indicates that SpliceAI is indeed able to predict splice altering mutations, since it seems clear that at least some of the predicted mutations have a functional consequence and that the effect of such consequence seems to be larger than for other types of mutations. More interestingly, all the genes that carry a variant that is predicted to alter splicing in more than 10 cell lines appeared to be important tumor suppressor genes. This is, however, to be expected since the function of tumor suppressor genes is to prevent cells from becoming carcinogenic. Therefore, it is common for carcinogenic cells to present aberrant isoforms that allow for maintenance of the abnormal carcinogenic rhythms [181]. On top of this, hereditary cancer genes, such as NF1 and RB1 have been previously reported to accumulate splice-altering mutations [132].

Regarding the response to drug treatment, it is hard to establish whether there is indeed an effect from the mutations that leads to a lower sensitivity to the drugs. Although Welch's t-test indicates some significant statistical differences in the deviation to the mean for some of the drug-gene pairs, the number of cell lines carrying a splice altering mutation is too low to calculate robust statistics. It would not be surprising that the splice-altering mutations do have an effect in the response to certain drugs, and therefore it is a venue definitely worth exploring. Probably better results would be achieved from collecting data specifically for this purpose - enriching for cell lines with splice altering mutations.

Lastly, the work carried out in this thesis shows that well-maintained and prolific public databases are very valuable resources for driving research, and the many insights that can be derived from these collections. Clearly, the ambitious project of the DepMap portal will go a long way in helping in the battle against cancer.

Bibliography

[1] Bruce Alberts. *Molecular biology of the cell*. English. 2015. ISBN: 978-0-8153-4432-2 0-8153-4432-5 978-0-8153-4464-3 0-8153-4464-3 978-0-8153-4524-4 0-8153-4524-0.

[2] Babak Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". en. In: *Nature Biotechnology* 33.8 (Aug. 2015). Number: 8 Publisher: Nature Publishing Group, pp. 831–838. ISSN: 1546-1696. DOI: 10.1038/nbt.3300. URL: https://www.nature.com/articles/nbt.3300 (visited on 05/24/2021).

[3] Preetha Anand et al. "Cancer is a Preventable Disease that Requires Major Lifestyle Changes". en. In: *Pharmaceutical Research* 25.9 (Sept. 2008), pp. 2097–2116. ISSN: 1573-904X. DOI: 10.1007/s11095-008-9661-9. URL: https://doi.org/10.1007/s11095-008-9661-9 (visited on 05/04/2021).

[4] Olga Anczuków and Adrian R. Krainer. "The spliceosome, a potential Achilles heel of MYC-driven tumors". In: *Genome Medicine* 7.1 (2015). Publisher: Genome Medicine, pp. 1–4. ISSN: 1756994X. DOI: 10.1186/s13073-015-0234-3. URL: http://dx.doi.org/10.1186/s13073-015-0234-3.

[5] Marco Baralle and Francisco Ernesto Baralle. "The splicing code". en. In: *Biosystems*. Code Biology 164 (Feb. 2018), pp. 39–48. ISSN: 0303-2647. DOI: 10.1016/j.biosystems.2017.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0303264717303210 (visited on 04/22/2021).

[6] Yoseph Barash et al. "Deciphering the splicing code". In: *Nature* 465.7294 (May 2010). Publisher: Nature Publishing Group, pp. 53–59. ISSN: 00280836. DOI: 10.1038/nature09000. URL: https://www.nature.com/articles/nature09000.

[7] Amrita Basu et al. "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules". eng. In: *Cell* 154.5 (Aug. 2013), pp. 1151–1161. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.08.003.

[8] Chiara Bazzichetto et al. "PTEN as a Prognostic/Predictive Biomarker in Cancer: An Unfulfilled Promise?" In: *Cancers* 11.4 (Mar. 2019). ISSN: 2072-6694. DOI: 10.3390/cancers11040435. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6520939/ (visited on 06/03/2021).

[9] Regina Berretta and Pablo Moscato. "Cancer biomarker discovery: The entropic hallmark". In: *PLoS ONE* 5.8 (2010). Publisher: Public Library of Science. ISSN: 19326203. DOI: 10.1371/journal.pone.0012262.

[10] "Big picture oncology through multi-omics". en. In: (). URL: https://www.nature.com/articles/d42473-020-00119-7 (visited on 04/24/2021).

[11] Jesse S. Boehm et al. “Cancer research needs a better map”. en. In: *Nature* 589.7843 (Jan. 2021). Number: 7843 Publisher: Nature Publishing Group, pp. 514–516. DOI: 10.1038/d41586-021-00182-0. URL: https://www.nature.com/articles/d41586-021-00182-0 (visited on 04/26/2021).

[12] Andrew P. Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. en. In: *Pattern Recognition* 30.7 (July 1997), pp. 1145–1159. ISSN: 0031-3203. DOI: 10.1016/S0031-3203(96)00142-2. URL: https://www.sciencedirect.com/science/article/pii/S0031320396001422 (visited on 05/17/2021).

[13] Sinisa Bratulic, Francesco Gatto, and Jens Nielsen. “The Translational Status of Cancer Liquid Biopsies”. en. In: *Regenerative Engineering and Translational Medicine* (Nov. 2019). ISSN: 2364-4141. DOI: 10.1007/s40883-019-00141-2. URL: https://doi.org/10.1007/s40883-019-00141-2 (visited on 05/24/2021).

[14] *Cancer genome research and precision medicine - National Cancer Institute*. en. cgvArticle. Archive Location: nciglobal,ncienterprise. Apr. 2015. URL: https://www.cancer.gov/about-nci/organization/ccg/cancer-genomics-overview (visited on 04/22/2021).

[15] *Cancer Statistics - National Cancer Institute*. URL: https://www.cancer.gov/about-cancer/understanding/statistics (visited on 01/03/2021).

[16] “Cancer Treatment Options”. en. In: (), p. 2.

[17] Ethan Cerami et al. “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data”. eng. In: *Cancer Discovery* 2.5 (May 2012), pp. 401–404. ISSN: 2159-8290. DOI: 10.1158/2159-8290.CD-12-0095.

[18] Hui Chen, Hudan Liu, and Guoliang Qing. “Targeting oncogenic Myc as a strategy for cancer treatment”. en. In: *Signal Transduction and Targeted Therapy* 3.1 (Feb. 2018). Number: 1 Publisher: Nature Publishing Group, pp. 1–7. ISSN: 2059-3635. DOI: 10.1038/s41392-018-0008-7. URL: https://www.nature.com/articles/s41392-018-0008-7 (visited on 05/31/2021).

[19] Alessandra J. Conforte et al. “Signaling Complexity Measured by Shannon Entropy and Its Application in Personalized Medicine”. In: *Frontiers in Genetics* 10 (Oct. 2019). Publisher: Frontiers Media S.A., p. 930. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00930. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00930/full.

[20] Thomas A. Cooper, Lili Wan, and Gideon Dreyfuss. “RNA and Disease”. In: *Cell* 136.4 (Feb. 2009), pp. 777–793. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.02.011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2866189/ (visited on 05/25/2021).

[21] Steven M. Corsello et al. “Non-oncology drugs are a source of previously unappreciated anti-cancer activity”. en. In: *bioRxiv* (Aug. 2019). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 730119. DOI: 10.1101/730119. URL: https://www.biorxiv.org/content/10.1101/730119v1 (visited on 04/27/2021).

[22] Pierre G. Coulie. “Human tumour antigens recognized by T cells: new perspectives for anti-cancer vaccines?” en. In: *Molecular Medicine Today* 3.6

(June 1997), pp. 261–268. ISSN: 1357-4310. DOI: 10.1016/S1357-4310(97)01049-6. URL: https://www.sciencedirect.com/science/article/pii/S1357431097010496 (visited on 05/04/2021).

[23] F. Crick. "Central dogma of molecular biology". eng. In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. ISSN: 0028-0836. DOI: 10.1038/227561a0.

[24] F. H. Crick. "On protein synthesis". eng. In: *Symposia of the Society for Experimental Biology* 12 (1958), pp. 138–163. ISSN: 0081-1386.

[25] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.giab008 (Feb. 2021). ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. URL: https://doi.org/10.1093/gigascience/giab008 (visited on 05/31/2021).

[26] Chi V. Dang et al. "Drugging the 'undruggable' cancer targets". eng. In: *Nature Reviews. Cancer* 17.8 (Aug. 2017), pp. 502–508. ISSN: 1474-1768. DOI: 10.1038/nrc.2017.36.

[27] Shipra Das et al. "ONCOGENIC SPLICING FACTOR SRSF1 IS A CRITICAL TRANSCRIPTIONAL TARGET OF MYC". In: *Cell reports* 1.2 (Feb. 2012), pp. 110–117. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2011.12.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334311/ (visited on 05/24/2021).

[28] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143874. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143874 (visited on 05/17/2021).

[29] Christiaan A. De Leeuw et al. "The statistical properties of gene-set analysis". In: *Nature Reviews Genetics* 17.6 (June 2016). Publisher: Nature Publishing Group, pp. 353–364. ISSN: 14710064. DOI: 10.1038/nrg.2016.29. URL: www.nature.com/nrg.

[30] Silvia De Rubeis et al. "Synaptic, transcriptional and chromatin genes disrupted in autism". en. In: *Nature* 515.7526 (Nov. 2014). Number: 7526 Publisher: Nature Publishing Group, pp. 209–215. ISSN: 1476-4687. DOI: 10.1038/nature13772. URL: https://www.nature.com/articles/nature13772 (visited on 05/25/2021).

[31] *Definition of cancer - NCI Dictionary of Cancer Terms - National Cancer Institute*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer (visited on 01/03/2021).

[32] Sven Degroeve et al. "Feature subset selection for splice site prediction". eng. In: *Bioinformatics (Oxford, England)* 18 Suppl 2 (2002), S75–83. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/18.suppl_2.s75.

[33] Joshua M. Dempster et al. *Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines*. Publication Title: bioRxiv. bioRxiv, July 2019. DOI: 10.1101/720243. URL: https://doi.org/10.1101/720243.

[34] Anamika Dhillon and Gyanendra K. Verma. "Convolutional neural network: a review of models, methodologies and applications to object detection". en.

In: *Progress in Artificial Intelligence* 9.2 (June 2020), pp. 85–112. ISSN: 2192-6360. DOI: 10.1007/s13748-019-00203-0. URL: https://doi.org/10.1007/s13748-019-00203-0 (visited on 05/25/2021).

[35] Riccardo Di Fiore et al. "RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis". eng. In: *Journal of Cellular Physiology* 228.8 (Aug. 2013), pp. 1676–1687. ISSN: 1097-4652. DOI: 10.1002/jcp.24329.

[36] John Doorbar et al. "Human papillomavirus molecular biology and disease association". In: *Reviews in Medical Virology* 25.Suppl Suppl 1 (Mar. 2015), pp. 2–23. ISSN: 1052-9276. DOI: 10.1002/rmv.1822. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5024016/ (visited on 04/22/2021).

[37] Michael Dougan, Glenn Dranoff, and Stephanie K. Dougan. "Cancer Immunotherapy: Beyond Checkpoint Blockade". In: *Annual Review of Cancer Biology* 3.1 (2019). _eprint: https://doi.org/10.1146/annurev-cancerbio-030518-055552, pp. 55–75. DOI: 10.1146/annurev-cancerbio-030518-055552. URL: https://doi.org/10.1146/annurev-cancerbio-030518-055552 (visited on 05/04/2021).

[38] R. Dulbecco. "A turning point in cancer research: sequencing the human genome". en. In: *Science* 231.4742 (Mar. 1986). Publisher: American Association for the Advancement of Science Section: Perspectives, pp. 1055–1056. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.3945817. URL: https://science.sciencemag.org/content/231/4742/1055 (visited on 04/22/2021).

[39] Steffen Durinck et al. "Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt". In: *Nature Protocols* 4.8 (2009). Publisher: Nature Publishing Group, pp. 1184–1191. ISSN: 17542189. DOI: 10.1038/nprot.2009.97. URL: https://ohsu.pure.elsevier.com/en/publications/mapping-identifiers-for-the-integration-of-genomic-datasets-with%E2%80%932.

[40] Ettaib El Marabti and Ihab Younis. *The cancer spliceome: Reprograming of alternative splicing in cancer*. Vol. 5. SEP. ISSN: 2296889X Publication Title: Frontiers in Molecular Biosciences. Frontiers Media S.A., Sept. 2018. DOI: 10.3389/fmolb.2018.00080. URL: www.frontiersin.org.

[41] Tim Elledge. "Primer on Molecular Genetics". en. In: (), p. 44.

[42] Frank Emmert-Streib et al. "An Introductory Review of Deep Learning for Prediction Models With Big Data". English. In: *Frontiers in Artificial Intelligence* 3 (2020). Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00004. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00004/full#B124 (visited on 04/27/2021).

[43] Scott J. Emrich et al. "Gene discovery and annotation using LCM-454 transcriptome sequencing". en. In: *Genome Research* 17.1 (Jan. 2007). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 69–73. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.5145806. URL: https://genome.cshlp.org/content/17/1/69 (visited on 04/24/2021).

[44] Gökcen Eraslan et al. “Deep learning: new computational modelling techniques for genomics”. en. In: *Nature Reviews Genetics* 20.7 (July 2019). Number: 7 Publisher: Nature Publishing Group, pp. 389–403. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0122-6. URL: https://www.nature.com/articles/s41576-019-0122-6 (visited on 05/24/2021).

[45] Beatrice Eymin. “Targeting the spliceosome machinery: A new therapeutic axis in cancer?” en. In: *Biochemical Pharmacology* (May 2020), p. 114039. ISSN: 0006-2952. DOI: 10.1016/j.bcp.2020.114039. URL: https://www.sciencedirect.com/science/article/pii/S0006295220302732 (visited on 06/04/2021).

[46] James D. Fackenthal and Lucy A. Godley. “Aberrant RNA splicing and its functional consequences in cancer cells”. In: *Disease Models & Mechanisms* 1.1 (2008), pp. 37–42. ISSN: 1754-8403. DOI: 10.1242/dmm.000331. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2561970/ (visited on 04/22/2021).

[47] Sofia Farkona, Eleftherios P. Diamandis, and Ivan M. Blasutig. “Cancer immunotherapy: the beginning of the end of cancer?” In: *BMC Medicine* 14 (May 2016). ISSN: 1741-7015. DOI: 10.1186/s12916-016-0623-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858828/ (visited on 05/04/2021).

[48] Rita Fior. “Cancer - when Cells Break the Rules and Hijack Their Own Planet”. In: 2019, pp. 1–20. DOI: 10.1007/978-3-030-11812-9_1.

[49] R. A. Fisher. “Statistical Methods for Research Workers”. en. In: *Breakthroughs in Statistics: Methodology and Distribution*. Ed. by Samuel Kotz and Norman L. Johnson. Springer Series in Statistics. New York, NY: Springer, 1992, pp. 66–70. ISBN: 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_6. URL: https://doi.org/10.1007/978-1-4612-4380-9_6 (visited on 05/11/2021).

[50] Luke Frankiw, David Baltimore, and Guideng Li. *Alternative mRNA splicing in cancer immunotherapy*. Vol. 19. 11. ISSN: 14741741 Publication Title: Nature Reviews Immunology. Nature Publishing Group, Nov. 2019. DOI: 10.1038/s41577-019-0195-7. URL: https://www.nature.com/articles/s41577-019-0195-7.

[51] Meital Gabay, Yulin Li, and Dean W. Felsher. “MYC Activation Is a Hallmark of Cancer Initiation and Maintenance”. In: *Cold Spring Harbor Perspectives in Medicine* 4.6 (June 2014). ISSN: 2157-1422. DOI: 10.1101/cshperspect.a014241. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4031954/ (visited on 06/03/2021).

[52] Jianjiong Gao et al. “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. eng. In: *Science Signaling* 6.269 (Apr. 2013), pl1. ISSN: 1937-9145. DOI: 10.1126/scisignal.2004088.

[53] Levi A Garraway. “Genomics-Driven Oncology: Framework for an Emerging Paradigm”. en. In: *JOURNAL OF CLINICAL ONCOLOGY* (2013), p. 9.

[54] Mahmoud Ghandi et al. “Next-generation characterization of the Cancer Cell Line Encyclopedia”. In: *Nature* 569.7757 (2019). Publisher: Springer US,

pp. 503–508. ISSN: 14764687. DOI: 10.1038/s41586-019-1186-3. URL: http://dx.doi.org/10.1038/s41586-019-1186-3.

[55] *Global Cancer Observatory*. URL: https://gco.iarc.fr/ (visited on 01/03/2021).

[56] Mehmet Gönen et al. "A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell Lines". eng. In: *Cell Systems* 5.5 (Nov. 2017), 485–497.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.09.004.

[57] Gregory J. Goodall and Vihandha O. Wickramasinghe. "RNA in cancer". en. In: *Nature Reviews Cancer* 21.1 (Jan. 2021), pp. 22–36. ISSN: 1474-175X, 1474-1768. DOI: 10.1038/s41568-020-00306-0. URL: http://www.nature.com/articles/s41568-020-00306-0 (visited on 04/22/2021).

[58] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[59] Sander Granneman et al. "The human Imp3 and Imp4 proteins form a ternary complex with hMpp10, which only interacts with the U3 snoRNA in 60–80S ribonucleoprotein complexes". In: *Nucleic Acids Research* 31.7 (Apr. 2003), pp. 1877–1887. ISSN: 0305-1048. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC152815/ (visited on 06/02/2021).

[60] William J. Greenleaf and Arend Sidow. "The future of sequencing: convergence of intelligent design and market Darwinism". In: *Genome Biology* 15.3 (Mar. 2014), p. 303. ISSN: 1474-760X. DOI: 10.1186/gb4168. URL: https://doi.org/10.1186/gb4168 (visited on 04/24/2021).

[61] Benjamin Haley and Filip Roudnicky. "Functional Genomics for Cancer Drug Target Discovery". en. In: *Cancer Cell* 38.1 (July 2020), pp. 31–43. ISSN: 15356108. DOI: 10.1016/j.ccell.2020.04.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535610820302051 (visited on 04/24/2021).

[62] Douglas Hanahan and Robert A. Weinberg. *Hallmarks of cancer: The next generation*. Vol. 144. 5. ISSN: 00928674 Publication Title: Cell. Elsevier, Mar. 2011. DOI: 10.1016/j.cell.2011.02.013. URL: http://www.cell.com/article/S0092867411001279/fulltext%20http://www.cell.com/article/S0092867411001279/abstract%20https://www.cell.com/cell/abstract/S0092-8674(11)00127-9.

[63] Seyed Hossein Hassanpour and Mohammadamin Dehghani. "Review of cancer from perspective of molecular". In: *Journal of Cancer Research and Practice* 4.4 (Dec. 2017). Publisher: Medknow, pp. 127–129. ISSN: 23113006. DOI: 10.1016/j.jcrpr.2017.07.001.

[64] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *arXiv:1512.03385 [cs]* (Dec. 2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 05/25/2021).

[65] James M. Heather and Benjamin Chain. "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1 (Jan. 2016), pp. 1–8. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/ (visited on 04/22/2021).

[66] Anna Hegele et al. "Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome". en. In: *Molecular Cell* 45.4 (Feb. 2012), pp. 567–580. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2011.12.034. URL: https:

//www.sciencedirect.com/science/article/pii/S1097276512000445 (visited on 05/28/2021).

[67] Nathaniel D. Heintzman et al. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome". en. In: *Nature Genetics* 39.3 (Mar. 2007). Number: 3 Publisher: Nature Publishing Group, pp. 311–318. ISSN: 1546-1718. DOI: 10.1038/ng1966. URL: https://www.nature.com/articles/ng1966 (visited on 05/24/2021).

[68] J. W.K. Ho et al. "Differential variability analysis of gene expression and its application to human diseases". In: *Bioinformatics* 24.13 (July 2008). Publisher: Oxford Academic, pp. i390–i398. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn142. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn142.

[69] Tiffany Y.T. Hsu et al. "The spliceosome is a therapeutic vulnerability in MYC-driven cancer". In: *Nature* 525.7569 (2015), pp. 384–388. ISSN: 14764687. DOI: 10.1038/nature14985.

[70] Yuanhua Huang and Guido Sanguinetti. "BRIE: Transcriptome-wide splicing quantification in single cells". In: *Genome Biology* 18.1 (June 2017). Publisher: BioMed Central Ltd., p. 123. ISSN: 1474760X. DOI: 10.1186/s13059-017-1248-5. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1248-5.

[71] *Illumina/SpliceAI*. original-date: 2019-01-02T20:37:36Z. Apr. 2021. URL: https://github.com/Illumina/SpliceAI (visited on 04/27/2021).

[72] Kishore Jaganathan et al. "Predicting Splicing from Primary Sequence with Deep Learning". In: *Cell* 176.3 (Jan. 2019). Publisher: Cell Press, 535–548.e24. ISSN: 10974172. DOI: 10.1016/j.cell.2018.12.015.

[73] Ian T. Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 2016). Publisher: Royal Society, p. 20150202. DOI: 10.1098/rsta.2015.0202. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (visited on 05/26/2021).

[74] Minoru Kanehisa and Susumu Goto. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Vol. 28. 1. ISSN: 03051048 Publication Title: Nucleic Acids Research. Oxford University Press, Jan. 2000. DOI: 10.1093/nar/28.1.27. URL: http://www.genome.ad.jp/kegg/.

[75] Yuri Kapustin et al. "Cryptic splice sites and split genes". In: *Nucleic Acids Research* 39.14 (Aug. 2011), pp. 5837–5844. ISSN: 0305-1048. DOI: 10.1093/nar/gkr203. URL: https://doi.org/10.1093/nar/gkr203 (visited on 05/25/2021).

[76] Matthias A. Karajannis and Rosalie E. Ferner. "NEUROFIBROMATOSIS-RELATED TUMORS: EMERGING BIOLOGY AND THERAPIES". In: *Current opinion in pediatrics* 27.1 (Feb. 2015), pp. 26–33. ISSN: 1040-8703. DOI: 10.1097/MOP.0000000000000169. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374132/ (visited on 06/03/2021).

[77] Yarden Katz et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nature Methods* 7.12 (Dec. 2010). Pub-

lisher: Nature Publishing Group, pp. 1009–1015. ISSN: 15487091. DOI: 10.1038/nmeth.1528. URL: https://www.nature.com/articles/nmeth.1528.

[78] Jens Keilwagen, Ivo Grosse, and Jan Grau. "Area under Precision-Recall Curves for Weighted and Unweighted Data". en. In: *PLOS ONE* 9.3 (Mar. 2014). Publisher: Public Library of Science, e92209. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0092209. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092209 (visited on 05/17/2021).

[79] Eddo Kim, Amir Goren, and Gil Ast. "Insights into the connection between cancer and alternative splicing". English. In: *Trends in Genetics* 24.1 (Jan. 2008). Publisher: Elsevier, pp. 7–10. ISSN: 0168-9525. DOI: 10.1016/j.tig.2007.10.001. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(07)00345-9 (visited on 04/22/2021).

[80] Jong Wook Kim et al. "Characterizing genomic alterations in cancer by complementary functional associations". eng. In: *Nature Biotechnology* 34.5 (May 2016), pp. 539–546. ISSN: 1546-1696. DOI: 10.1038/nbt.3527.

[81] Cheryl M. Koh et al. "MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis". eng. In: *Nature* 523.7558 (July 2015), pp. 96–100. ISSN: 1476-4687. DOI: 10.1038/nature14351.

[82] B. R. Korf. "Malignancy in neurofibromatosis type 1". eng. In: *The Oncologist* 5.6 (2000), pp. 477–485. ISSN: 1083-7159. DOI: 10.1634/theoncologist.5-6-477.

[83] Alberto R. Kornblihtt. "Coupling transcription and alternative splicing". eng. In: *Advances in Experimental Medicine and Biology* 623 (2007), pp. 175–189. ISSN: 0065-2598. DOI: 10.1007/978-0-387-77374-2_11.

[84] Alberto R. Kornblihtt et al. *Alternative splicing: A pivotal step between eukaryotic transcription and translation*. Vol. 14. 3. ISSN: 14710072 Publication Title: Nature Reviews Molecular Cell Biology. Nature Publishing Group, Mar. 2013. DOI: 10.1038/nrm3525. URL: www.nature.com/reviews/molcellbio.

[85] Alberto R. Kornblihtt et al. "Multiple links between transcription and splicing". en. In: *RNA* 10.10 (Jan. 2004). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1489–1498. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.7100104. URL: http://rnajournal.cshlp.org/content/10/10/1489 (visited on 04/22/2021).

[86] Guennadi Kozlov et al. "Molecular determinants of PAM2 recognition by the MLLE domain of poly(A)-binding protein". eng. In: *Journal of Molecular Biology* 397.2 (Mar. 2010), pp. 397–407. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2010.01.032.

[87] Michael Ladomery. "Aberrant alternative splicing is another hallmark of cancer". eng. In: *International Journal of Cell Biology* 2013 (2013), p. 463786. ISSN: 1687-8876. DOI: 10.1155/2013/463786.

[88] H. O. Lancaster. "THE COMBINATION OF PROBABILITIES: AN APPLICATION OF ORTHONORMAL FUNCTIONS". In: *Australian Journal of Statistics* 3.1 (Apr. 1961). Publisher: John Wiley & Sons, Ltd, pp. 20–

33. ISSN: 00049581. DOI: 10.1111/j.1467-842X.1961.tb00058.x. URL: http://doi.wiley.com/10.1111/j.1467-842X.1961.tb00058.x.

[89] Céline M. Laumont et al. "Noncoding regions are the main source of targetable tumor-specific antigens". In: *Science Translational Medicine* 10.470 (Dec. 2018). Publisher: American Association for the Advancement of Science. ISSN: 19466242. DOI: 10.1126/scitranslmed.aau5516. URL: http://stm.sciencemag.org/.

[90] Michael Lawrence et al. "Software for Computing and Annotating Genomic Ranges". en. In: *PLOS Computational Biology* 9.8 (Aug. 2013). Publisher: Public Library of Science, e1003118. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003118. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118 (visited on 06/03/2021).

[91] Walter Lawrence. "History of Surgical Oncology". en. In: *Surgery: Basic Science and Clinical Evidence*. Ed. by Jeffrey A. Norton et al. New York, NY: Springer, 2008, pp. 1889–1900. ISBN: 978-0-387-68113-9. DOI: 10.1007/978-0-387-68113-9_90. URL: https://doi.org/10.1007/978-0-387-68113-9_90 (visited on 04/21/2021).

[92] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". en. In: *Nature* 521.7553 (May 2015). Number: 7553 Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 05/25/2021).

[93] Amy S.Y. Lee et al. "eIF3d is an mRNA cap-binding protein required for specialized translation initiation". In: *Nature* 536.7614 (Aug. 2016), pp. 96–99. ISSN: 0028-0836. DOI: 10.1038/nature18954. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5003174/ (visited on 06/02/2021).

[94] Michael K. K. Leung et al. "Deep learning of the tissue-regulated splicing code". In: *Bioinformatics* 30.12 (June 2014). Publisher: Oxford University Press, pp. i121–i129. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btu277. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu277.

[95] Jake Lever, Martin Krzywinski, and Naomi Altman. "Principal component analysis". en. In: *Nature Methods* 14.7 (July 2017). Number: 7 Publisher: Nature Publishing Group, pp. 641–642. ISSN: 1548-7105. DOI: 10.1038/nmeth.4346. URL: https://www.nature.com/articles/nmeth.4346 (visited on 05/26/2021).

[96] Bo Li and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1 (Aug. 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323 (visited on 05/25/2021).

[97] Lie Li et al. "Integrative gene set enrichment analysis utilizing isoform-specific expression". In: *Genetic Epidemiology* 41.6 (Sept. 2017). Publisher: Wiley-Liss Inc., pp. 498–510. ISSN: 10982272. DOI: 10.1002/gepi.22052. URL: /pmc/articles/PMC5598160/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5598160/.

[98] Maxwell W. Libbrecht and William Stafford Noble. "Machine learning applications in genetics and genomics". en. In: *Nature Reviews Genetics* 16.6 (June 2015). Number: 6 Publisher: Nature Publishing Group, pp. 321–332. ISSN: 1471-0064. DOI: 10.1038/nrg3920. URL: https://www.nature.com/articles/nrg3920 (visited on 05/24/2021).

[99] Arthur Liberzon et al. "The Molecular Signatures Database Hallmark Gene Set Collection". In: *Cell Systems* 1.6 (Dec. 2015). Publisher: Cell Press, pp. 417–425. ISSN: 24054712. DOI: 10.1016/j.cels.2015.12.004. URL: /pmc/articles/PMC4707969/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/.

[100] Jennifer K. Litton, Harold J. Burstein, and Nicholas C. Turner. "Molecular Testing in Breast Cancer". eng. In: *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting* 39 (Jan. 2019), e1–e7. ISSN: 1548-8756. DOI: 10.1200/EDBK_237715.

[101] Ming Y. Lu et al. "AI-based pathology predicts origins for cancers of unknown primary". en. In: *Nature* (May 2021). Publisher: Nature Publishing Group, pp. 1–5. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03512-4. URL: https://www.nature.com/articles/s41586-021-03512-4 (visited on 05/24/2021).

[102] Yuanjun Lu et al. "Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy". In: *Molecular Cancer* 19 (Apr. 2020). ISSN: 1476-4598. DOI: 10.1186/s12943-020-01197-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7184703/ (visited on 05/25/2021).

[103] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. ISSN: 1533-7928. URL: http://jmlr.org/papers/v9/vandermaaten08a.html (visited on 05/26/2021).

[104] Sarah K. Madden et al. "Taking the Myc out of cancer: toward therapeutic strategies to directly inhibit c-Myc". In: *Molecular Cancer* 20.1 (Jan. 2021), p. 3. ISSN: 1476-4598. DOI: 10.1186/s12943-020-01291-6. URL: https://doi.org/10.1186/s12943-020-01291-6 (visited on 06/03/2021).

[105] Farhad Maleki et al. "Gene Set Analysis: Challenges, Opportunities, and Future Research". English. In: *Frontiers in Genetics* 11 (2020). Publisher: Frontiers. ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00654. URL: https://www.frontiersin.org/articles/10.3389/fgene.2020.00654/full (visited on 05/12/2021).

[106] Ravi Mathur et al. "Gene set analysis methods: A systematic comparison". In: *BioData Mining* 11.1 (May 2018). Publisher: BioMed Central Ltd., p. 8. ISSN: 17560381. DOI: 10.1186/s13040-018-0166-8. URL: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-018-0166-8.

[107] A M Maxam and W Gilbert. "A new method for sequencing DNA." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.2 (Feb. 1977), pp. 560–564. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392330/ (visited on 04/22/2021).

[108] Jeremy F. McRae et al. “Prevalence and architecture of de novo mutations in developmental disorders”. en. In: *Nature* 542.7642 (Feb. 2017). Number: 7642 Publisher: Nature Publishing Group, pp. 433–438. ISSN: 1476-4687. DOI: 10.1038/nature21062. URL: https://www.nature.com/articles/nature21062 (visited on 05/25/2021).

[109] Mateusz Mendel et al. “Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing”. en. In: *Cell* (Apr. 2021). ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.03.062. URL: https://www.sciencedirect.com/science/article/pii/S0092867421004359 (visited on 05/25/2021).

[110] Robin M. Meyers et al. “Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells”. In: *Nature Genetics* 49.12 (Dec. 2017). Publisher: Nature Publishing Group, pp. 1779–1784. ISSN: 15461718. DOI: 10.1038/ng.3984. URL: https://pubmed.ncbi.nlm.nih.gov/29083409/.

[111] Hong-Nan Mo and Peng Liu. “Targeting MET in cancer therapy”. In: *Chronic Diseases and Translational Medicine* 3.3 (July 2017), pp. 148–153. ISSN: 2095-882X. DOI: 10.1016/j.cdtm.2017.06.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5643781/ (visited on 06/03/2021).

[112] McKale Montgomery and Aishwarya Srinivasan. “Epigenetic Gene Regulation by Dietary Compounds in Cancer Prevention”. eng. In: *Advances in Nutrition (Bethesda, Md.)* 10.6 (Nov. 2019), pp. 1012–1028. ISSN: 2156-5376. DOI: 10.1093/advances/nmz046.

[113] Siddhartha Mukherjee. *The emperor of all maladies: a biography of cancer*. English. OCLC: 464593321. New York: Scribner, 2010. ISBN: 978-1-4391-0795-9 978-1-4391-9570-3 978-1-4391-8171-3 978-1-4104-4715-9 978-1-4391-7091-5.

[114] Dougu Nam. “Effect of the absolute statistic on gene-sampling gene-set analysis methods”. eng. In: *Statistical Methods in Medical Research* 26.3 (June 2017), pp. 1248–1260. ISSN: 1477-0334. DOI: 10.1177/0962280215574014.

[115] Bianca Nogrady. “How cancer genomics is transforming diagnosis and treatment”. en. In: *Nature* 579.7800 (Mar. 2020). Number: 7800 Publisher: Nature Publishing Group, S10–S11. DOI: 10.1038/d41586-020-00845-4. URL: https://www.nature.com/articles/d41586-020-00845-4 (visited on 04/24/2021).

[116] Gonçalo Nogueira et al. “Nonsense-mediated RNA decay and its bipolar function in cancer”. In: *Molecular Cancer* 20.1 (Apr. 2021), p. 72. ISSN: 1476-4598. DOI: 10.1186/s12943-021-01364-0. URL: https://doi.org/10.1186/s12943-021-01364-0 (visited on 06/03/2021).

[117] Uwe Ohler et al. “Computational analysis of core promoters in the Drosophila genome”. eng. In: *Genome Biology* 3.12 (2002), RESEARCH0087. ISSN: 1474-760X. DOI: 10.1186/gb-2002-3-12-research0087.

[118] Valeria Ossovskaya et al. “Upregulation of Poly (ADP-Ribose) Polymerase-1 (PARP1) in Triple-Negative Breast Cancer and Other Primary Human Tumor Types”. eng. In: *Genes & Cancer* 1.8 (Aug. 2010), pp. 812–821. ISSN: 1947-6027. DOI: 10.1177/1947601910383418.

[119] William Pao and Nicolas Girard. “New driver mutations in non-small-cell lung cancer”. en. In: *The Lancet Oncology* 12.2 (Feb. 2011), pp. 175–180.

ISSN: 1470-2045. DOI: 10.1016/S1470-2045(10)70087-5. URL: https://www.sciencedirect.com/science/article/pii/S1470204510700875 (visited on 04/26/2021).

[120] Barbara L. Parsons. *Multiclonal tumor origin: Evidence and implications*. Vol. 777. ISSN: 13882139 Publication Title: Mutation Research - Reviews in Mutation Research. Elsevier B.V., July 2018. DOI: 10.1016/j.mrrev.2018.05.001.

[121] Indrajeet Patil. "Visualizations with statistical details: The 'ggstatsplot' approach". In: *Journal of Open Source Software* 6.61 (2021), p. 3167. DOI: 10.21105/joss.03167. URL: https://doi.org/10.21105/joss.03167.

[122] Kiran Raosaheb Patil and Jens Nielsen. "Uncovering transcriptional regulation of metabolism by using metabolic network topology". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.8 (Feb. 2005), pp. 2685–2689. ISSN: 0027-8424. DOI: 10.1073/pnas.0406811102.

[123] Karl Pearson. "LIII. *On lines and planes of closest fit to systems of points in space*". en. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1901), pp. 559–572. ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440109462720. URL: https://www.tandfonline.com/doi/full/10.1080/14786440109462720 (visited on 05/18/2021).

[124] Mihaela Pertea, Xiaoying Lin, and Steven L. Salzberg. "GeneSplicer: a new computational method for splice site prediction". In: *Nucleic Acids Research* 29.5 (Mar. 2001), pp. 1185–1190. ISSN: 0305-1048. DOI: 10.1093/nar/29.5.1185. URL: https://doi.org/10.1093/nar/29.5.1185 (visited on 05/25/2021).

[125] A. Petitjean et al. "TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes". en. In: *Oncogene* 26.15 (Apr. 2007). Number: 15 Publisher: Nature Publishing Group, pp. 2157–2165. ISSN: 1476-5594. DOI: 10.1038/sj.onc.1210302. URL: https://www.nature.com/articles/1210302 (visited on 06/03/2021).

[126] Vincent Piras, Masaru Tomita, and Kumar Selvarajoo. "Is central dogma a global property of cellular information flow?" English. In: *Frontiers in Physiology* 3 (2012). Publisher: Frontiers. ISSN: 1664-042X. DOI: 10.3389/fphys.2012.00439. URL: https://www.frontiersin.org/articles/10.3389/fphys.2012.00439/full#B6 (visited on 04/22/2021).

[127] Maximilian W Popp and Lynne E Maquat. "Nonsense-mediated mRNA Decay and Cancer". en. In: *Current Opinion in Genetics & Development*. Cancer genomics 48 (Feb. 2018), pp. 44–50. ISSN: 0959-437X. DOI: 10.1016/j.gde.2017.10.007. URL: https://www.sciencedirect.com/science/article/pii/S0959437X17301041 (visited on 06/03/2021).

[128] Michael A. Postow, Margaret K. Callahan, and Jedd D. Wolchok. "Immune Checkpoint Blockade in Cancer Therapy". In: *Journal of Clinical Oncology* 33.17 (June 2015), pp. 1974–1982. ISSN: 0732-183X. DOI: 10.1200/JCO.2014.59.4358. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4980573/ (visited on 05/04/2021).

[129] Alex H. Ramos et al. “Oncotator: Cancer Variant Annotation Tool”. en. In: *Human Mutation* 36.4 (Apr. 2015), E2423–E2429. ISSN: 10597794. DOI: 10.1002/humu.22771. URL: http://doi.wiley.com/10.1002/humu.22771 (visited on 05/25/2021).

[130] Matthew G. Rees et al. “Correlating chemical sensitivity and basal gene expression reveals mechanism of action”. eng. In: *Nature Chemical Biology* 12.2 (Feb. 2016), pp. 109–116. ISSN: 1552-4469. DOI: 10.1038/nchembio.1986.

[131] M. G. Reese et al. “Improved splice site detection in Genie”. eng. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 4.3 (1997), pp. 311–323. ISSN: 1066-5277. DOI: 10.1089/cmb.1997.4.311.

[132] Christy L. Rhine et al. “Hereditary cancer genes are highly susceptible to splicing mutations”. en. In: *PLOS Genetics* 14.3 (Mar. 2018). Ed. by Rolf I. Skotheim, e1007231. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007231. URL: https://dx.plos.org/10.1371/journal.pgen.1007231 (visited on 06/03/2021).

[133] Xavier Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12.1 (Mar. 2011), p. 77. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-77. URL: https://doi.org/10.1186/1471-2105-12-77 (visited on 05/17/2021).

[134] Peyton Rous. “A TRANSMISSIBLE AVIAN NEOPLASM. (SARCOMA OF THE COMMON FOWL.)” In: *The Journal of Experimental Medicine* 12.5 (Sept. 1910), pp. 696–705. ISSN: 0022-1007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2124810/ (visited on 04/22/2021).

[135] Eiji Sakashita et al. “Human RNPS1 and its associated factors: a versatile alternative pre-mRNA splicing regulator in vivo”. eng. In: *Molecular and Cellular Biology* 24.3 (Feb. 2004), pp. 1174–1187. ISSN: 0270-7306. DOI: 10.1128/MCB.24.3.1174-1187.2004.

[136] Stephan J. Sanders et al. “De novo mutations revealed by whole-exome sequencing are strongly associated with autism”. en. In: *Nature* 485.7397 (May 2012). Number: 7397 Publisher: Nature Publishing Group, pp. 237–241. ISSN: 1476-4687. DOI: 10.1038/nature10945. URL: https://www.nature.com/articles/nature10945 (visited on 05/25/2021).

[137] F. Sanger, G. G. Brownlee, and B. G. Barrell. “A two-dimensional fractionation procedure for radioactive nucleotides”. en. In: *Journal of Molecular Biology* 13.2 (Sept. 1965), 373–IN4. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(65)80104-8. URL: https://www.sciencedirect.com/science/article/pii/S0022283665801048 (visited on 04/22/2021).

[138] Dominic Schmidt et al. “ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions”. eng. In: *Methods (San Diego, Calif.)* 48.3 (July 2009), pp. 240–248. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2009.03.001.

[139] Rocco Sciarrillo et al. *The role of alternative splicing in cancer: From oncogenesis to drug resistance*. Vol. 53. ISSN: 15322084 Publication Title: Drug Resistance Updates. Churchill Livingstone, Dec. 2020. DOI: 10.1016/j.drup.2020.100728.

[140] Brinton Seashore-Ludlow et al. “Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset”. eng. In: *Cancer Discovery* 5.11 (Nov. 2015), pp. 1210–1223. ISSN: 2159-8290. DOI: 10.1158/2159-8290.CD-15-0235.

[141] Eran Segal et al. “A genomic code for nucleosome positioning”. en. In: *Nature* 442.7104 (Aug. 2006). Number: 7104 Publisher: Nature Publishing Group, pp. 772–778. ISSN: 1476-4687. DOI: 10.1038/nature04979. URL: https://www.nature.com/articles/nature04979 (visited on 05/24/2021).

[142] Michael Seiler et al. “Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types”. In: *Cell reports* 23.1 (Apr. 2018), 282–296.e4. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2018.01.088. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5933844/ (visited on 05/24/2021).

[143] Phillip A. Sharp. “The discovery of split genes and RNA splicing”. English. In: *Trends in Biochemical Sciences* 30.6 (June 2005). Publisher: Elsevier, pp. 279–281. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2005.04.002. URL: https://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004(05)00092-7 (visited on 04/22/2021).

[144] Karen Sheppard et al. “Targeting PI3 kinase/AKT/mTOR signaling in cancer”. eng. In: *Critical Reviews in Oncogenesis* 17.1 (2012), pp. 69–95. ISSN: 0893-9675. DOI: 10.1615/critrevoncog.v17.i1.60.

[145] Johannes Smolander, Matthias Dehmer, and Frank Emmert-Streib. “Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders”. en. In: *FEBS Open Bio* 9.7 (2019). _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/2211-5463.12652, pp. 1232–1248. ISSN: 2211-5463. DOI: https://doi.org/10.1002/2211-5463.12652. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/2211-5463.12652 (visited on 04/27/2021).

[146] Johannes Smolander et al. “Comparing biological information contained in mRNA and non-coding RNAs for classification of lung cancer patients”. In: *BMC Cancer* 19.1 (Dec. 2019), p. 1176. ISSN: 1471-2407. DOI: 10.1186/s12885-019-6338-1. URL: https://doi.org/10.1186/s12885-019-6338-1 (visited on 04/27/2021).

[147] Ana M. Soto and Carlos Sonnenschein. “The somatic mutation theory of cancer: growing problems with the paradigm?” eng. In: *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 26.10 (Oct. 2004), pp. 1097–1107. ISSN: 0265-9247. DOI: 10.1002/bies.20087.

[148] Fran Supek, Ben Lehner, and Rik G. H. Lindeboom. “To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases”. English. In: *Trends in Genetics* 0.0 (Dec. 2020). Publisher: Elsevier. ISSN: 0168-9525. DOI: 10.1016/j.tig.2020.11.002. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(20)30307-3 (visited on 06/03/2021).

[149] *Surgery for Cancer - National Cancer Institute*. en. cgvArticle. Archive Location: nciglobal,ncienterprise. Apr. 2015. URL: https://www.cancer.gov/about-cancer/treatment/types/surgery (visited on 04/27/2021).

[150] A. Sveen et al. “Aberrant RNA splicing in cancer; Expression changes and driver mutations of splicing factor genes”. In: *Oncogene* 35.19 (2016). Publisher: Nature Publishing Group, pp. 2413–2427. ISSN: 14765594. DOI: 10.1038/onc.2015.318. URL: http://dx.doi.org/10.1038/onc.2015.318.

[151] *The Human Genome Project*. en. URL: https://www.genome.gov/human-genome-project (visited on 04/23/2021).

[152] Elizabeth A. Tovar and Carrie R. Graveel. “MET in human cancer: germline and somatic mutations”. In: *Annals of Translational Medicine* 5.10 (May 2017). ISSN: 2305-5839. DOI: 10.21037/atm.2017.03.64. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5451609/ (visited on 06/03/2021).

[153] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. In: *Nature Biotechnology* 28.5 (May 2010). Publisher: Nature Publishing Group, pp. 511–515. ISSN: 10870156. DOI: 10.1038/nbt.1621. URL: https://www.nature.com/articles/nbt.1621.

[154] Aviad Tsherniak et al. “Defining a Cancer Dependency Map”. eng. In: *Cell* 170.3 (July 2017), 564–576.e16. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.06.010.

[155] Tychele N. Turner et al. “Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA”. eng. In: *American Journal of Human Genetics* 98.1 (Jan. 2016), pp. 58–74. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2015.11.023.

[156] Leif Väremo, Jens Nielsen, and Intawat Nookaew. “Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods”. In: *Nucleic Acids Research* 41.8 (Apr. 2013). Publisher: Nucleic Acids Res, pp. 4378–4391. ISSN: 03051048. DOI: 10.1093/nar/gkt111. URL: https://pubmed.ncbi.nlm.nih.gov/23444143/.

[157] David L. Vaux. “In defense of the somatic mutation theory of cancer”. en. In: *BioEssays* 33.5 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201100022, pp. 341–343. ISSN: 1521-1878. DOI: https://doi.org/10.1002/bies.201100022. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201100022 (visited on 04/23/2021).

[158] Julian P. Venables. “Aberrant and Alternative Splicing in Cancer”. en. In: *Cancer Research* 64.21 (Nov. 2004). Publisher: American Association for Cancer Research Section: Review, pp. 7647–7654. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-04-1910. URL: https://cancerres.aacrjournals.org/content/64/21/7647 (visited on 04/22/2021).

[159] Dass S. Vinay et al. *Immune evasion in cancer: Mechanistic basis and therapeutic strategies*. Vol. 35. ISSN: 10963650 Publication Title: Seminars in Cancer Biology. Academic Press, Dec. 2015. DOI: 10.1016/j.semcancer.2015.03.004.

[160] Emily A. Vucic et al. “Translating cancer ‘omics’ to improved outcomes”. In: *Genome Research* 22.2 (Feb. 2012), pp. 188–195. ISSN: 1088-9051. DOI:

10.1101/gr.124354.111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3266027/ (visited on 04/24/2021).

[161] Markus C. Wahl, Cindy L. Will, and Reinhard Lührmann. *The Spliceosome: Design Principles of a Dynamic RNP Machine*. Vol. 136. 4. ISSN: 00928674 Publication Title: Cell. Cell Press, Feb. 2009. DOI: 10.1016/j.cell.2009.02.009.

[162] Thomas A. Waldmann. "Immunotherapy: past, present and future". en. In: *Nature Medicine* 9.3 (Mar. 2003). Number: 3 Publisher: Nature Publishing Group, pp. 269–277. ISSN: 1546-170X. DOI: 10.1038/nm0303-269. URL: https://www.nature.com/articles/nm0303-269 (visited on 05/04/2021).

[163] Chen Wang et al. "Alternative approaches to target Myc for cancer treatment". en. In: *Signal Transduction and Targeted Therapy* 6.1 (Mar. 2021). Number: 1 Publisher: Nature Publishing Group, pp. 1–14. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00500-y. URL: https://www.nature.com/articles/s41392-021-00500-y (visited on 06/03/2021).

[164] Eric Wang and Iannis Aifantis. *RNA Splicing and Cancer*. Vol. 6. 8. ISSN: 24058033 Publication Title: Trends in Cancer. Cell Press, Aug. 2020. DOI: 10.1016/j.trecan.2020.04.011. URL: http://www.cell.com/article/S2405803320301412/fulltext.

[165] Eric T. Wang et al. "Alternative isoform regulation in human tissue transcriptomes". en. In: *Nature* 456.7221 (Nov. 2008). Number: 7221 Publisher: Nature Publishing Group, pp. 470–476. ISSN: 1476-4687. DOI: 10.1038/nature07509. URL: https://www.nature.com/articles/nature07509 (visited on 04/22/2021).

[166] Guey Shin Wang and Thomas A. Cooper. *Splicing in disease: Disruption of the splicing code and the decoding machinery*. Vol. 8. 10. ISSN: 14710056 Publication Title: Nature Reviews Genetics. Nature Publishing Group, Oct. 2007. DOI: 10.1038/nrg2164. URL: www.nature.com/reviews/genetics.

[167] Kai Wang et al. *Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes*. Tech. rep. 3. Publication Title: Int. J. Computational Biology and Drug Design Volume: 7. 2014, pp. 11–13.

[168] Kai Wang et al. "EntropyExplorer: An R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression". In: *BMC Research Notes* 8.1 (Dec. 2015). Publisher: BioMed Central Ltd., p. 832. ISSN: 17560500. DOI: 10.1186/s13104-015-1786-4. URL: https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-015-1786-4.

[169] Luyao Wang et al. "PARP1 in Carcinomas and PARP1 Inhibitors as Antineoplastic Drugs". In: *International Journal of Molecular Sciences* 18.10 (Oct. 2017). ISSN: 1422-0067. DOI: 10.3390/ijms18102111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5666793/ (visited on 06/03/2021).

[170] Zefeng Wang and Christopher B. Burge. "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code". eng. In: *RNA (New York, N.Y.)* 14.5 (May 2008), pp. 802–813. ISSN: 1469-9001. DOI: 10.1261/rna.876308.

[171] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews. Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/ (visited on 04/24/2021).

[172] James D. Watson. *Manuscript: "The Human Genome Project: Past, Present, and Future"*. 1990. URL: http://libgallery.cshl.edu/items/show/53660 (visited on 04/23/2021).

[173] Julia Weber et al. "In vivo functional screening for systems-level integrative cancer genomics". en. In: *Nature Reviews Cancer* 20.10 (Oct. 2020), pp. 573–593. ISSN: 1474-175X, 1474-1768. DOI: 10.1038/s41568-020-0275-9. URL: http://www.nature.com/articles/s41568-020-0275-9 (visited on 04/24/2021).

[174] James West et al. "Differential network entropy reveals cancer system hallmarks". In: *Scientific Reports* 2.1 (Nov. 2012). Publisher: Nature Publishing Group, pp. 1–8. ISSN: 20452322. DOI: 10.1038/srep00802. URL: www.nature.com/scientificreports.

[175] Brian T. Wilhelm and Josette-Renée Landry. "RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing". en. In: *Methods*. Global approaches to study gene regulation 48.3 (July 2009), pp. 249–257. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2009.03.016. URL: https://www.sciencedirect.com/science/article/pii/S1046202309000632 (visited on 04/24/2021).

[176] Hui Y. Xiong et al. "The human splicing code reveals new insights into the genetic determinants of disease". In: *Science* 347.6218 (Jan. 2015). Publisher: American Association for the Advancement of Science. ISSN: 10959203. DOI: 10.1126/science.1254806. URL: http://science.sciencemag.org/.

[177] Gene Yeo and Christopher B. Burge. "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals". eng. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11.2-3 (2004), pp. 377–394. ISSN: 1066-5277. DOI: 10.1089/1066527041410418.

[178] Lynn Yi et al. "Gene-level differential analysis at transcript-level resolution". In: *Genome Biology* 19.1 (Apr. 2018). Publisher: BioMed Central Ltd., p. 53. ISSN: 1474760X. DOI: 10.1186/s13059-018-1419-z. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1419-z.

[179] Channing Yu et al. "High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines". eng. In: *Nature Biotechnology* 34.4 (Apr. 2016), pp. 419–423. ISSN: 1546-1696. DOI: 10.1038/nbt.3460.

[180] Ahmet Zehir et al. "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients". en. In: *Nature Medicine* 23.6 (June 2017). Number: 6 Publisher: Nature Publishing Group, pp. 703–713. ISSN: 1546-170X. DOI: 10.1038/nm.4333. URL: https://www.nature.com/articles/nm.4333 (visited on 05/25/2021).

[181] Yuanjiao Zhang et al. "Alternative splicing and cancer: a systematic review". en. In: *Signal Transduction and Targeted Therapy* 6.1 (Feb. 2021). Number:

1 Publisher: Nature Publishing Group, pp. 1–14. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00486-7. URL: https://www.nature.com/articles/s41392-021-00486-7 (visited on 06/03/2021).

[182] Zijun Zhang et al. "Deep-learning augmented RNA-seq analysis of transcript splicing". In: *Nature Methods* 16.4 (Apr. 2019). Publisher: Nature Publishing Group, pp. 307–310. ISSN: 15487105. DOI: 10.1038/s41592-019-0351-9. URL: https://doi.org/10.1038/s41592-019-0351-9.

[183] Ran Zhao et al. "Implications of Genetic and Epigenetic Alterations of CDKN2A (p16INK4a) in Cancer". In: *EBioMedicine* 8 (May 2016), pp. 30–39. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2016.04.017. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4919535/ (visited on 06/03/2021).

[184] Jian Zhou and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". en. In: *Nature Methods* 12.10 (Oct. 2015). Number: 10 Publisher: Nature Publishing Group, pp. 931–934. ISSN: 1548-7105. DOI: 10.1038/nmeth.3547. URL: https://www.nature.com/articles/nmeth.3547 (visited on 05/24/2021).

[185] James Zou et al. "A primer on deep learning in genomics". en. In: *Nature Genetics* 51.1 (Jan. 2019). Number: 1 Publisher: Nature Publishing Group, pp. 12–18. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0295-5. URL: https://www.nature.com/articles/s41588-018-0295-5 (visited on 05/24/2021).

A

Appendix 1 - Exploratory analysis

A.1 Splicing factors

List of splicing factors

Symbol	Description	Category	Mutation events
SPEN	spen family transcriptional repressor	other	251
RNF213	ring finger protein 213	other	250
SRRM2	serine/arginine repetitive matrix 2	core	202
INTS1	integrator complex subunit 1	other	169
ZC3H13	zinc finger CCCH-type containing 13	other	136
KIAA1429		other	128
SF3B1	splicing factor 3b subunit 1	core	123
NUMA1	nuclear mitotic apparatus protein 1	other	114
RBBP6	RB binding protein 6, ubiquitin ligase	no_core	114
GPATCH8	G-patch domain containing 8	other	111
SNRNP200	small nuclear ribonucleoprotein U5 subunit 200	core	109
ZC3H4	zinc finger CCCH-type containing 4	other	108
GEMIN5	gem nuclear organelle associated protein 5	other	102
TCERG1	transcription elongation regulator 1	core	100
PRPF8	pre-mRNA processing factor 8	core	98
ZC3H18	zinc finger CCCH-type containing 18	no_core	93
AQR	aquarius intron-binding spliceosomal factor	core	91
CWC22	CWC22 spliceosome associated protein homolog	core	89
ELAVL2	ELAV like RNA binding protein 2	other	89
DHX57	DExH-box helicase 57	no_core	88
TTC14	tetratricopeptide repeat domain 14	no_core	85
RBM47	RNA binding motif protein 47	other	84
SEC31B	SEC31 homolog B, COPII coat complex component	no_core	84
THOC2	THO complex 2	no_core	84
ACIN1	apoptotic chromatin condensation inducer 1	no_core	83
DDX3X	DEAD-box helicase 3 X-linked	no_core	83
DHX38	DEAH-box helicase 38	core	82
EIF3A	eukaryotic translation initiation factor 3 subunit A	other	82
CLASRP	CLK4 associating serine/arginine rich protein	other	81
CDK12	cyclin dependent kinase 12	other	79
RBM27	RNA binding motif protein 27	other	79
DHX34	DExH-box helicase 34	other	78
CACTIN	cactin, spliceosome C complex subunit	core	77
ZMYM3	zinc finger MYM-type containing 3	other	77
DHX16	DEAH-box helicase 16	core	76
TRIM24	tripartite motif containing 24	other	75
CCAR1	cell division cycle and apoptosis regulator 1	core	74
INTS5	integrator complex subunit 5	other	73
ZFR	zinc finger RNA binding protein	other	73
DHX36	DEAH-box helicase 36	other	72
RBM15B	RNA binding motif protein 15B	other	72
ILF3	interleukin enhancer binding factor 3	no_core	68
RBM26	RNA binding motif protein 26	other	68
SF1	splicing factor 1	core	68
PAXBP1	PAX3 and PAX7 binding protein 1	no_core	67
RNF20	ring finger protein 20	other	67
SRRT	serrate, RNA effector molecule	no_core	67
DHX9	DExH-box helicase 9	no_core	66
PPIG	peptidylprolyl isomerase G	core	66

Continued on next page

Table A.1 – continued from previous page

Symbol	Description	Category	Mutation events
ZNF326	zinc finger protein 326	other	66
DHX8	DEAH-box helicase 8	core	65
NOVA1	NOVA alternative splicing regulator 1	other	64
RBM15	RNA binding motif protein 15	other	64
CDK11A	cyclin dependent kinase 11A	core	63
DNAJC6	DnaJ heat shock protein family (Hsp40) member C6	no_core	63
PRPF40A	pre-mRNA processing factor 40 homolog A	core	63
TAF15	TATA-box binding protein associated factor 15	other	63
PRPF4B	pre-mRNA processing factor 4B	core	62
RBM25	RNA binding motif protein 25	core	62
EFTUD2	elongation factor Tu GTP binding domain containing 2	core	61
PNN	pinin, desmosome associated protein	no_core	61
RNF40	ring finger protein 40	other	61
AGGF1	angiogenic factor with G-patch and FHA domains 1	no_core	60
HSPA8	heat shock protein family A (Hsp70) member 8	core	60
INTS7	integrator complex subunit 7	other	60
PPWD1	peptidylprolyl isomerase domain and WD repeat containing 1	core	60
PRPF40B	pre-mRNA processing factor 40 homolog B	other	60
SYNCRIP	synaptotagmin binding cytoplasmic RNA interacting protein	no_core	60
DDX27	DEAD-box helicase 27	other	59
HNRNPUL2	heterogeneous nuclear ribonucleoprotein U like 2	other	59
RBMXL2	RBMX like 2	other	59
THRAP3	thyroid hormone receptor associated protein 3	core	59
ZC3HAV1	zinc finger CCCH-type containing, antiviral 1	other	59
MATR3	matrin 3	no_core	58
RBM10	RNA binding motif protein 10	core	58
SDE2	SDE2 telomere maintenance homolog	core	58
GPATCH1	G-patch domain containing 1	core	57
HNRNPM	heterogeneous nuclear ribonucleoprotein M	no_core	57
HNRNPUL1	heterogeneous nuclear ribonucleoprotein U like 1	no_core	57
INTS6	integrator complex subunit 6	other	57
RBM23	RNA binding motif protein 23	no_core	57
DDX26B		other	56
SART1	spliceosome associated factor 1, recruiter of U4/U6.U5 tri-snRNP	core	55
XAB2	XPA binding protein 2	core	55
INTS4	integrator complex subunit 4	other	54
TFIP11	tuftelin interacting protein 11	no_core	54
SRRM1	serine and arginine repetitive matrix 1	no_core	53
DDX42	DEAD-box helicase 42	core	52
KHSRP	KH-type splicing regulatory protein	other	52
MYEF2	myelin expression factor 2	other	52
RALYL	RALY RNA binding protein like	other	52
SKIV2L2		other	52
U2SURP	U2 snRNP associated SURP domain containing	core	52
ZC3H11A	zinc finger CCCH-type containing 11A	other	52
SF3B2	splicing factor 3b subunit 2	core	51
CDC40	cell division cycle 40	core	50
CDC5L	cell division cycle 5 like	core	50
DDX50	DExD-box helicase 50	other	50
DHX30	DExH-box helicase 30	other	50
PABPC1	poly(A) binding protein cytoplasmic 1	no_core	50
C17orf85		no_core	49
FUBP1	far upstream element binding protein 1	other	49
JUP	junction plakoglobin	no_core	49
TOPORS	TOP1 binding arginine/serine rich protein, E3 ubiquitin ligase	other	49
ZCCHC8	zinc finger CCHC-type containing 8	other	49
CELF4	CUGBP Elav-like family member 4	other	48
DDX21	DExD-box helicase 21	other	48
DDX46	DEAD-box helicase 46	core	48
FMR1	FMRP translational regulator 1	other	48
TNPO1	transportin 1	other	48
BCAS1	brain enriched myelin associated protein 1	other	47
DDX18	DEAD-box helicase 18	other	47
HNRNPU	heterogeneous nuclear ribonucleoprotein U	no_core	47
PRPF4	pre-mRNA processing factor 4	core	47
TOP1MT	DNA topoisomerase I mitochondrial	other	47
CLK2	CDC like kinase 2	other	46
DDX17	DEAD-box helicase 17	no_core	45
HNRNPL	heterogeneous nuclear ribonucleoprotein L	other	45
LSM1	LSM1 homolog, mRNA degradation associated	other	45

Continued on next page

Table A.1 – continued from previous page

Symbol	Description	Category	Mutation events
RAVER1	ribonucleoprotein, PTB binding 1	other	45
RBM5	RNA binding motif protein 5	core	45
SF3B3	splicing factor 3b subunit 3	core	45
SRPK1	SRSF protein kinase 1	core	45
SRPK2	SRSF protein kinase 2	core	45
DDX41	DEAD-box helicase 41	core	44
SFPQ	splicing factor proline and glutamine rich	other	44
SNRPN	small nuclear ribonucleoprotein polypeptide N	other	44
HTATSF1	HIV-1 Tat specific factor 1	core	43
MOV10	Mov10 RISC complex RNA helicase	other	43
SNRNP70	small nuclear ribonucleoprotein U1 subunit 70	core	43
DDX20	DEAD-box helicase 20	other	42
DDX23	DEAD-box helicase 23	core	42
DHX35	DEAH-box helicase 35	core	42
FUS	FUS RNA binding protein	core	42
HNRNPCL1	heterogeneous nuclear ribonucleoprotein C like 1	other	42
RAVER2	ribonucleoprotein, PTB binding 2	other	42
RBM39	RNA binding motif protein 39	no_core	42
ZNF131	zinc finger protein 131	other	42
DDX1	DEAD-box helicase 1	other	41
DHX15	DEAH-box helicase 15	core	41
SUGP1	SURP and G-patch domain containing 1	core	41
CRNKL1	crooked neck pre-mRNA splicing factor 1	core	40
DDX5	DEAD-box helicase 5	core	40
DHX40	DEAH-box helicase 40	other	40
NCBP1	nuclear cap binding protein subunit 1	core	40
NSRP1	nuclear speckle splicing regulatory protein 1	no_core	40
PRCC	proline rich mitotic checkpoint control factor	core	40
RBM14	RNA binding motif protein 14	other	40
DDX39B	DExD-box helicase 39B	no_core	39
DGCR14		no_core	39
INTS3	integrator complex subunit 3	other	39
LENG1	leukocyte receptor cluster member 1	other	39
QKI	QKI, KH domain containing RNA binding	no_core	39
SF3A1	splicing factor 3a subunit 1	core	39
SRPK3	SRSF protein kinase 3	other	39
CDK10	cyclin dependent kinase 10	no_core	38
CTNNBL1	catenin beta like 1	core	38
FUBP3	far upstream element binding protein 3	other	38
GRSF1	G-rich RNA sequence binding factor 1	other	38
HSPA5	heat shock protein family A (Hsp70) member 5	other	38
PRPF3	pre-mRNA processing factor 3	core	38
PTBP1	polypyrimidine tract binding protein 1	no_core	38
CELF2	CUGBP Elav-like family member 2	other	37
CLK1	CDC like kinase 1	other	37
GPATCH3	G-patch domain containing 3	other	37
KHDRBS3	KH RNA binding domain containing, signal transduction associated 3	no_core	37
SNRNP48	small nuclear ribonucleoprotein U11/U12 subunit 48	other	37
SRSF4	serine and arginine rich splicing factor 4	no_core	37
SRSF6	serine and arginine rich splicing factor 6	no_core	37
CELF3	CUGBP Elav-like family member 3	other	36
CELF5	CUGBP Elav-like family member 5	other	36
HNRNPDL	heterogeneous nuclear ribonucleoprotein D like	other	36
KIN	Kin17 DNA and RNA binding protein	core	36
PRPF19	pre-mRNA processing factor 19	core	36
PRPF38B	pre-mRNA processing factor 38B	other	36
SF3A2	splicing factor 3a subunit 2	core	36
SREK1	splicing regulatory glutamic acid and lysine rich protein 1	other	36
WBP11	WW domain binding protein 11	core	36
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	no_core	35
HNRNPH1	heterogeneous nuclear ribonucleoprotein H1	no_core	35
HNRNPR	heterogeneous nuclear ribonucleoprotein R	no_core	35
IGF2BP3	insulin like growth factor 2 mRNA binding protein 3	other	35
MBNL1	muscleblind like splicing regulator 1	other	35
MBNL2	muscleblind like splicing regulator 2	other	35
NONO	non-POU domain containing octamer binding	other	35
BUD13	BUD13 homolog	core	34
RBM42	RNA binding motif protein 42	no_core	34
RNPC3	RNA binding region (RNP1, RRM) containing 3	other	34
TIAL1	TIA1 cytotoxic granule associated RNA binding protein like 1	other	34

Continued on next page

Table A.1 – continued from previous page

Symbol	Description	Category	Mutation events
ZNF207	zinc finger protein 207	no_core	34
CWC25	CWC25 spliceosome associated protein homolog	other	33
DDX19B	DEAD-box helicase 19B	other	33
ELAVL4	ELAV like RNA binding protein 4	other	33
HNRNPA3	heterogeneous nuclear ribonucleoprotein A3	no_core	33
HNRNPC	heterogeneous nuclear ribonucleoprotein C	no_core	33
HNRNPK	heterogeneous nuclear ribonucleoprotein K	no_core	33
KHDRBS1	KH RNA binding domain containing, signal transduction associated 1	no_core	33
MFSD11	major facilitator superfamily domain containing 11	other	33
NOVA2	NOVA alternative splicing regulator 2	other	33
PPM1G	protein phosphatase, Mg2+/Mn2+ dependent 1G	other	33
PRPF39	pre-mRNA processing factor 39	other	33
PSIP1	PC4 and SFRS1 interacting protein 1	other	33
WDR77	WD repeat domain 77	no_core	33
ZNF830	zinc finger protein 830	core	33
CLK4	CDC like kinase 4	other	32
CWC27	CWC27 spliceosome associated cyclophilin	core	32
ELAVL1	ELAV like RNA binding protein 1	no_core	32
NELFE	negative elongation factor complex member E	other	32
PRMT5	protein arginine methyltransferase 5	no_core	32
SLU7	SLU7 homolog, splicing factor	core	32
THOC5	THO complex 5	no_core	32
YBX3	Y-box binding protein 3	no_core	32
ELAVL3	ELAV like RNA binding protein 3	other	31
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	no_core	31
PCBP4	poly(rC) binding protein 4	other	31
PPIL2	peptidylprolyl isomerase like 2	core	31
PTBP3	polypyrimidine tract binding protein 3	other	31
RBFOX2	RNA binding fox-1 homolog 2	no_core	31
RBM45	RNA binding motif protein 45	other	31
THOC1	THO complex 1	no_core	31
ZRSR1		other	31
CLK3	CDC like kinase 3	other	30
CPSF6	cleavage and polyadenylation specific factor 6	other	30
U2AF1	U2 small nuclear RNA auxiliary factor 1	core	30
HNRNPD	heterogeneous nuclear ribonucleoprotein D	no_core	29
HNRNPH2	heterogeneous nuclear ribonucleoprotein H2	no_core	29
PCBP1	poly(rC) binding protein 1	no_core	29
PLRG1	pleiotropic regulator 1	core	29
PUF60	poly(U) binding splicing factor 60	core	29
RNF113A	ring finger protein 113A	core	29
SRSF11	serine and arginine rich splicing factor 11	other	29
TRA2B	transformer 2 beta homolog	no_core	29
CD2BP2	CD2 cytoplasmic tail binding protein 2	core	28
HNRNPLL	heterogeneous nuclear ribonucleoprotein L like	other	28
IK	IK cytokine	core	28
LUC7L	LUC7 like	no_core	28
PPIL4	peptidylprolyl isomerase like 4	no_core	28
RBM22	RNA binding motif protein 22	core	28
RNF34	ring finger protein 34	other	28
SNRPA	small nuclear ribonucleoprotein polypeptide A	core	28
DDX6	DEAD-box helicase 6	other	27
FRA10AC1	FRA10A associated CGG repeat 1	no_core	27
PPP1CA	protein phosphatase 1 catalytic subunit alpha	no_core	27
PRPF31	pre-mRNA processing factor 31	core	27
PTBP2	polypyrimidine tract binding protein 2	no_core	27
SNW1	SNW domain containing 1	core	27
U2AF2	U2 small nuclear RNA auxiliary factor 2	core	27
DDX39A	DExD-box helicase 39A	other	26
FAM50B	family with sequence similarity 50 member B	core	26
ZRSR2	zinc finger CCCH-type, RNA binding motif and serine/arginine rich 2	other	26
CELF1	CUGBP Elav-like family member 1	no_core	25
EEF1A1	eukaryotic translation elongation factor 1 alpha 1	other	25
LUC7L3	LUC7 like 3 pre-mRNA splicing factor	other	25
MBNL3	muscleblind like splicing regulator 3	other	25
NKAP	NFKB activating protein	no_core	25
NOSIP	nitric oxide synthase interacting protein	core	25
THOC3	THO complex 3	no_core	25
TIA1	TIA1 cytotoxic granule associated RNA binding protein	other	25
TOE1	target of EGR1, exonuclease	no_core	25

Continued on next page

Table A.1 – continued from previous page

Symbol	Description	Category	Mutation events
CCDC94		other	24
DDX19A	DEAD-box helicase 19A	other	24
FAM50A	family with sequence similarity 50 member A	core	24
MSI1	musashi RNA binding protein 1	other	24
MSI2	musashi RNA binding protein 2	other	24
PPIE	peptidylprolyl isomerase E	core	24
PPP1R8	protein phosphatase 1 regulatory subunit 8	no_core	24
SRSF7	serine and arginine rich splicing factor 7	no_core	24
WBP4	WW domain binding protein 4	other	24
WDR83	WD repeat domain 83	core	24
WTAP	WT1 associated protein	other	24
CHERP	calcium homeostasis endoplasmic reticulum protein	core	23
GNB2L1		no_core	23
MFAP1	microfibril associated protein 1	core	23
RALY	RALY heterogeneous nuclear ribonucleoprotein	no_core	23
SF3A3	splicing factor 3a subunit 3	core	23
YBX1	Y-box binding protein 1	no_core	23
ZNF346	zinc finger protein 346	other	23
FAM58A		other	22
NRIP2	nuclear receptor interacting protein 2	no_core	22
RBM4	RNA binding motif protein 4	no_core	22
RBMX	RNA binding motif protein X-linked	no_core	22
SNIP1	Smad nuclear interacting protein 1	core	22
CELF6	CUGBP Elav-like family member 6	other	21
EIF4A3	eukaryotic translation initiation factor 4A3	core	21
GPKOW	G-patch domain and KOW motifs	core	21
HNRNPF	heterogeneous nuclear ribonucleoprotein F	no_core	21
PCBP3	poly(rC) binding protein 3	other	21
PDCD7	programmed cell death 7	other	21
RBM17	RNA binding motif protein 17	core	21
RBMS1	RNA binding motif single stranded interacting protein 1	other	21
RBMXL1	RBMX like 1	other	21
SNRNP35	small nuclear ribonucleoprotein U11/U12 subunit 35	other	21
SRSF5	serine and arginine rich splicing factor 5	no_core	21
THOC6	THO complex 6	other	21
U2AF1L4	U2 small nuclear RNA auxiliary factor 1 like 4	other	21
USP39	ubiquitin specific peptidase 39	core	21
CCDC130	coiled-coil domain containing 130	no_core	20
CSN3	casein kappa	other	20
SSB	small RNA binding exonuclease protection factor La	other	20
EIF2S2	eukaryotic translation initiation factor 2 subunit beta	other	19
ISY1	ISY1 splicing factor homolog	core	19
PCBP2	poly(rC) binding protein 2	no_core	19
PSEN1	presenilin 1	other	19
RBM4B	RNA binding motif protein 4B	other	19
RBMX2	RNA binding motif protein X-linked 2	core	19
SYF2	SYF2 pre-mRNA splicing factor	core	19
CIRBP	cold inducible RNA binding protein	no_core	18
HNRNPA0	heterogeneous nuclear ribonucleoprotein A0	no_core	18
PQBP1	polyglutamine binding protein 1	core	18
SF3B4	splicing factor 3b subunit 4	core	18
SRSF3	serine and arginine rich splicing factor 3	no_core	18
ZCCHC10	zinc finger CCHC-type containing 10	no_core	18
ALYREF	Aly/REF export factor	no_core	17
BCAS2	BCAS2 pre-mRNA processing factor	core	17
BUB3	BUB3 mitotic checkpoint protein	core	17
DNAJC8	DnaJ heat shock protein family (Hsp40) member C8	core	17
HSPA1B	heat shock protein family A (Hsp70) member 1B	other	17
PRPF18	pre-mRNA processing factor 18	core	17
SRSF9	serine and arginine rich splicing factor 9	no_core	17
C9orf78	chromosome 9 open reading frame 78	core	16
HNRNPH3	heterogeneous nuclear ribonucleoprotein H3	no_core	16
HSPA1A	heat shock protein family A (Hsp70) member 1A	no_core	16
HSPB1	heat shock protein family B (small) member 1	core	16
PRPF38A	pre-mRNA processing factor 38A	core	16
RNPS1	RNA binding protein with serine rich domain 1	no_core	16
SNRNP40	small nuclear ribonucleoprotein U5 subunit 40	core	16
SRSF1	serine and arginine rich splicing factor 1	no_core	16
SRSF2	serine and arginine rich splicing factor 2	no_core	16
ARGLU1	arginine and glutamate rich 1	no_core	15

Continued on next page

Table A.1 – continued from previous page

Symbol	Description	Category	Mutation events
FRG1	FSHD region gene 1	core	15
SMU1	SMU1 DNA replication regulator and spliceosomal factor	core	15
SNRNP27	small nuclear ribonucleoprotein U4/U6.U5 subunit 27	core	15
BAG2	BAG cochaperone 2	no_core	14
ILF2	interleukin enhancer binding factor 2	no_core	14
LSM10	LSM10, U7 small nuclear RNA associated	other	14
LSM4	LSM4 homolog, U6 small nuclear RNA and mRNA degradation associated	core	14
SNRPC	small nuclear ribonucleoprotein polypeptide C	core	14
THOC7	THO complex 7	no_core	14
TRA2A	transformer 2 alpha homolog	no_core	14
ZCRB1	zinc finger CCHC-type and RNA binding motif containing 1	other	14
CXorf56		core	13
SRSF12	serine and arginine rich splicing factor 12	other	13
GEMIN2	gem nuclear organelle associated protein 2	other	12
RBM3	RNA binding motif protein 3	other	12
RBM7	RNA binding motif protein 7	no_core	12
SAP30BP	SAP30 binding protein	no_core	12
ZMAT5	zinc finger matrin-type 5	other	12
BUD31	BUD31 homolog	core	11
RBM8A	RNA binding motif protein 8A	core	11
SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	core	11
ZMAT2	zinc finger matrin-type 2	core	11
C19orf43		core	10
CLNS1A	chloride nucleotide-sensitive channel 1A	no_core	10
SMNDC1	survival motor neuron domain containing 1	core	10
C1QBP	complement C1q binding protein	other	9
DDX3Y	DEAD-box helicase 3 Y-linked	other	9
SAP18	Sin3A associated protein 18	core	9
SNRNP25	small nuclear ribonucleoprotein U11/U12 subunit 25	other	9
LSM7	LSM7 homolog, U6 small nuclear RNA and mRNA degradation associated	core	8
NUDT21	nudix hydrolase 21	other	8
SNRPA1	small nuclear ribonucleoprotein polypeptide A'	core	8
SNRPD2	small nuclear ribonucleoprotein D2 polypeptide	core	8
SNRPG	small nuclear ribonucleoprotein polypeptide G	core	8
CCDC12	coiled-coil domain containing 12	core	7
CWC15	CWC15 spliceosome associated protein homolog	core	7
HNRNPAB	heterogeneous nuclear ribonucleoprotein A/B	no_core	7
LSMD1		other	7
NCBP2	nuclear cap binding protein subunit 2	core	7
PPIL1	peptidylprolyl isomerase like 1	core	7
SMN1	survival of motor neuron 1, telomeric	no_core	7
SNRPD3	small nuclear ribonucleoprotein D3 polypeptide	core	7
LSM2	LSM2 homolog, U6 small nuclear RNA and mRNA degradation associated	core	6
SNRPB2	small nuclear ribonucleoprotein polypeptide B2	core	6
SNRPD1	small nuclear ribonucleoprotein D1 polypeptide	core	6
PPIL3	peptidylprolyl isomerase like 3	core	5
SNURF	SNRPN upstream reading frame	other	5
LSM3	LSM3 homolog, U6 small nuclear RNA and mRNA degradation associated	core	4
LSM5	LSM5 homolog, U6 small nuclear RNA and mRNA degradation associated	core	4
MAGOH	mago homolog, exon junction complex subunit	core	4
PPIH	peptidylprolyl isomerase H	core	4
SF3B5	splicing factor 3b subunit 5	core	4
FAM32A	family with sequence similarity 32 member A	core	3
SNRPF	small nuclear ribonucleoprotein polypeptide F	core	3
LSM6	LSM6 homolog, U6 small nuclear RNA and mRNA degradation associated	core	2
LUC7L2	LUC7 like 2, pre-mRNA splicing factor	other	2
TXNL4A	thioredoxin like 4A	core	2
PHF5A	PHD finger protein 5A	core	1
SNRPE	small nuclear ribonucleoprotein polypeptide E	core	1
UBL5	ubiquitin like 5	core	1

A.2 Mutational landscape of splicing factors

Here we include several plots that help to understand the contents of the somatic mutations dataset from DepMap.

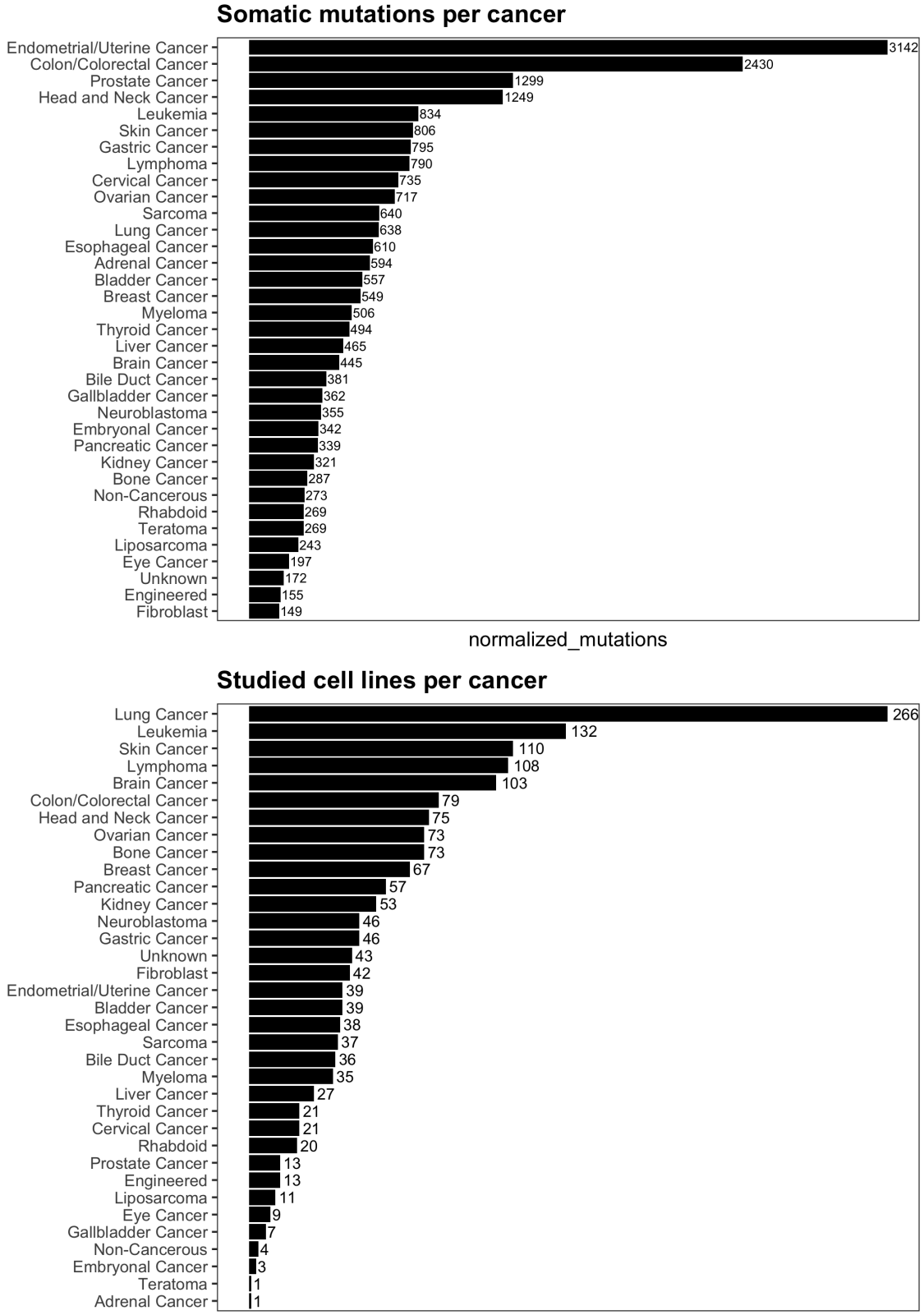

Figure A.1: Top graph shows the amount of normalized somatic mutations - including mutations in the splicing factors - reported per cancer cell lines. Bottom shows the number of analyzed cell lines per cancer.

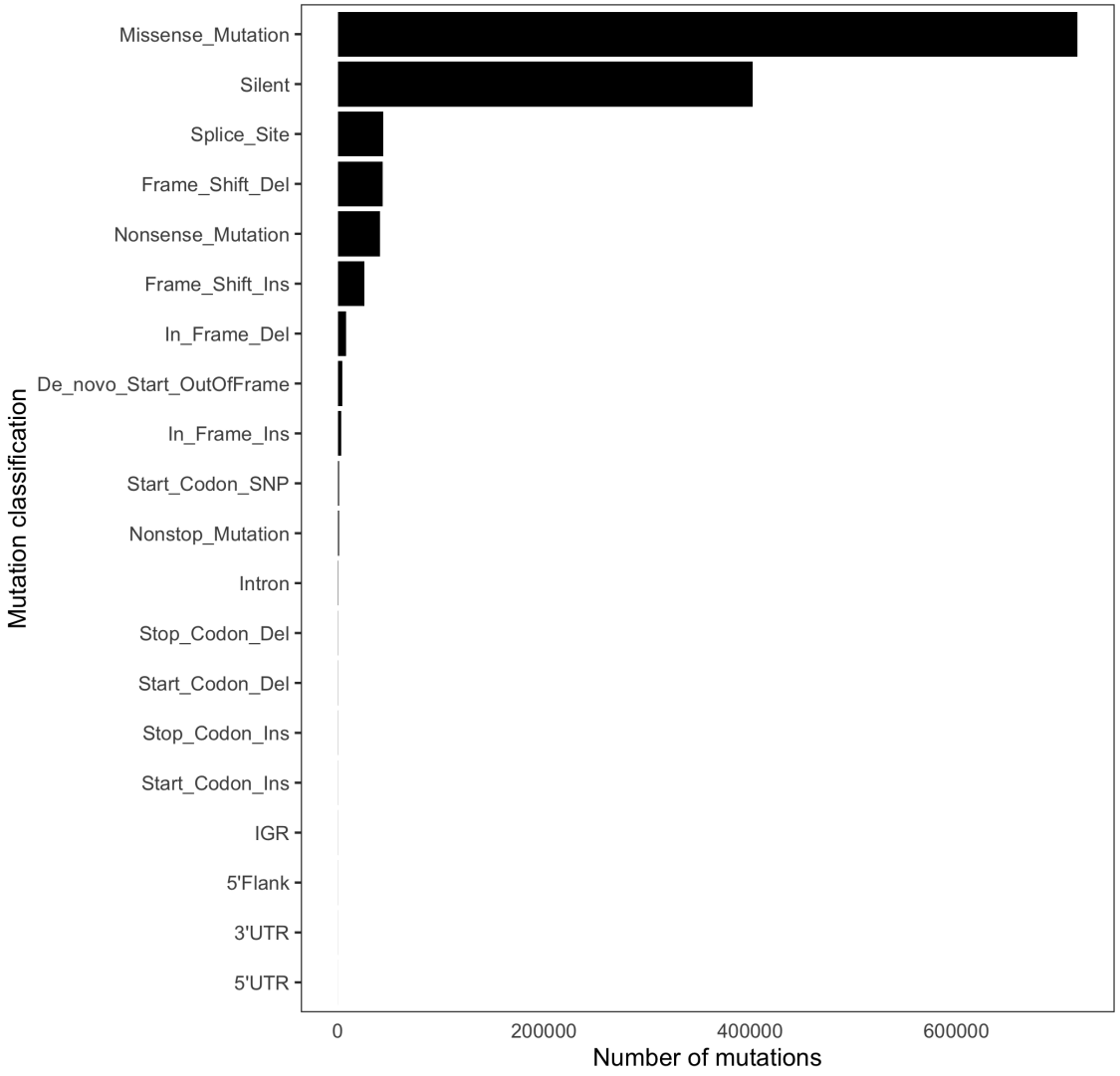

Figure A.2: Count of the somatic mutations reported by DepMap in the CCLE project classified according to mutation type.

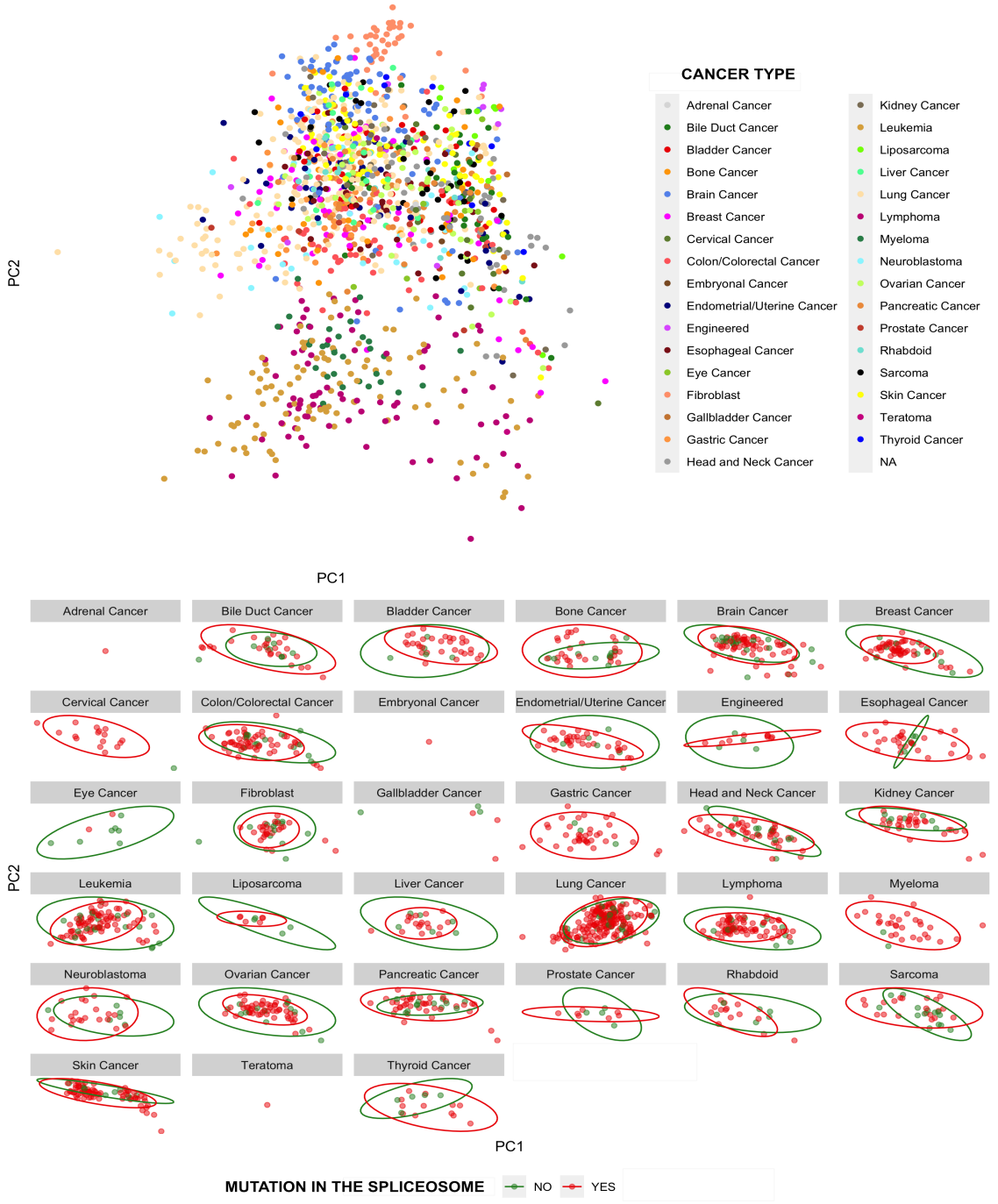

Figure A.3: PCA performed labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the non-core spliceosome machinery

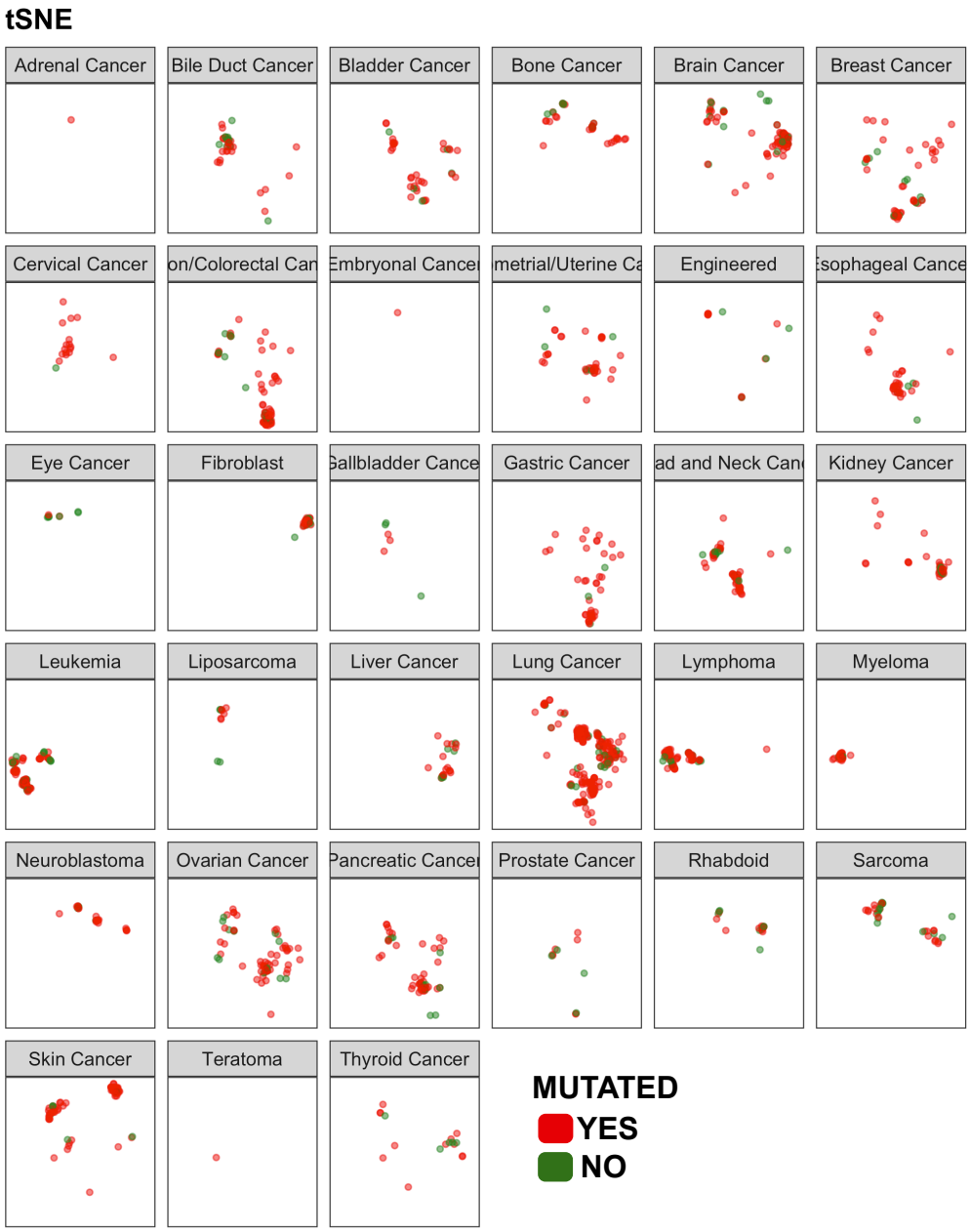

Figure A.4: t-SNE performed labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the no-core spliceosome machinery

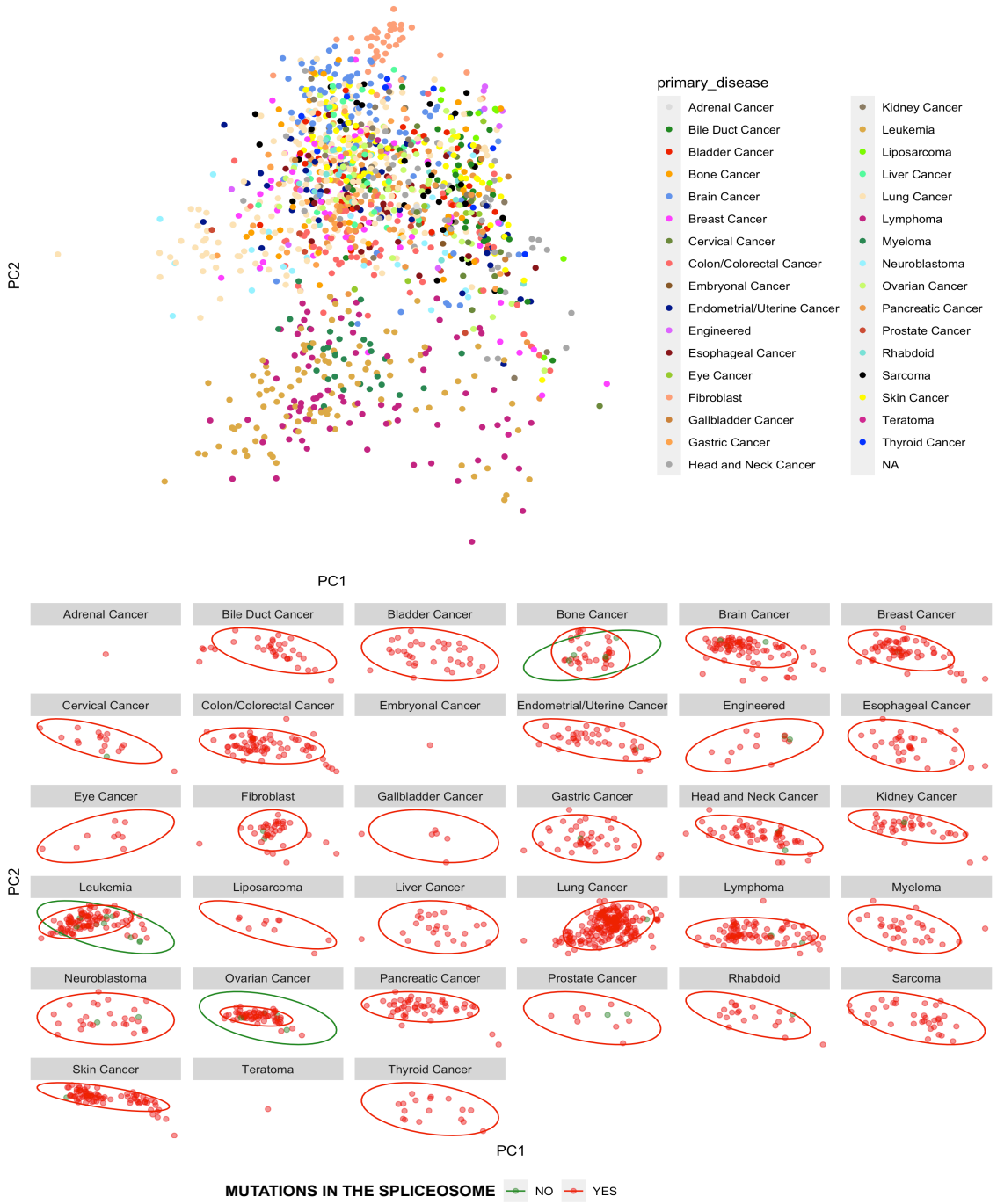

Figure A.5: PCA performed labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the "other" spliceosome machinery

Figure A.6: t-SNE performed labeling the cell lines depending on presence or absence of mutations on the genes encoding for proteins that are part of the "other" spliceosome machinery

A.3 Heatmaps

To produce the binary heatmaps (see below and figure 4.2) we used the R package *tidyheatmap*, that provides an interface for the package *pheatmap*. To draw the heatmaps we used the clustering method "complete" and the clustering distance "binary" for clustering the splicing factors (rows).

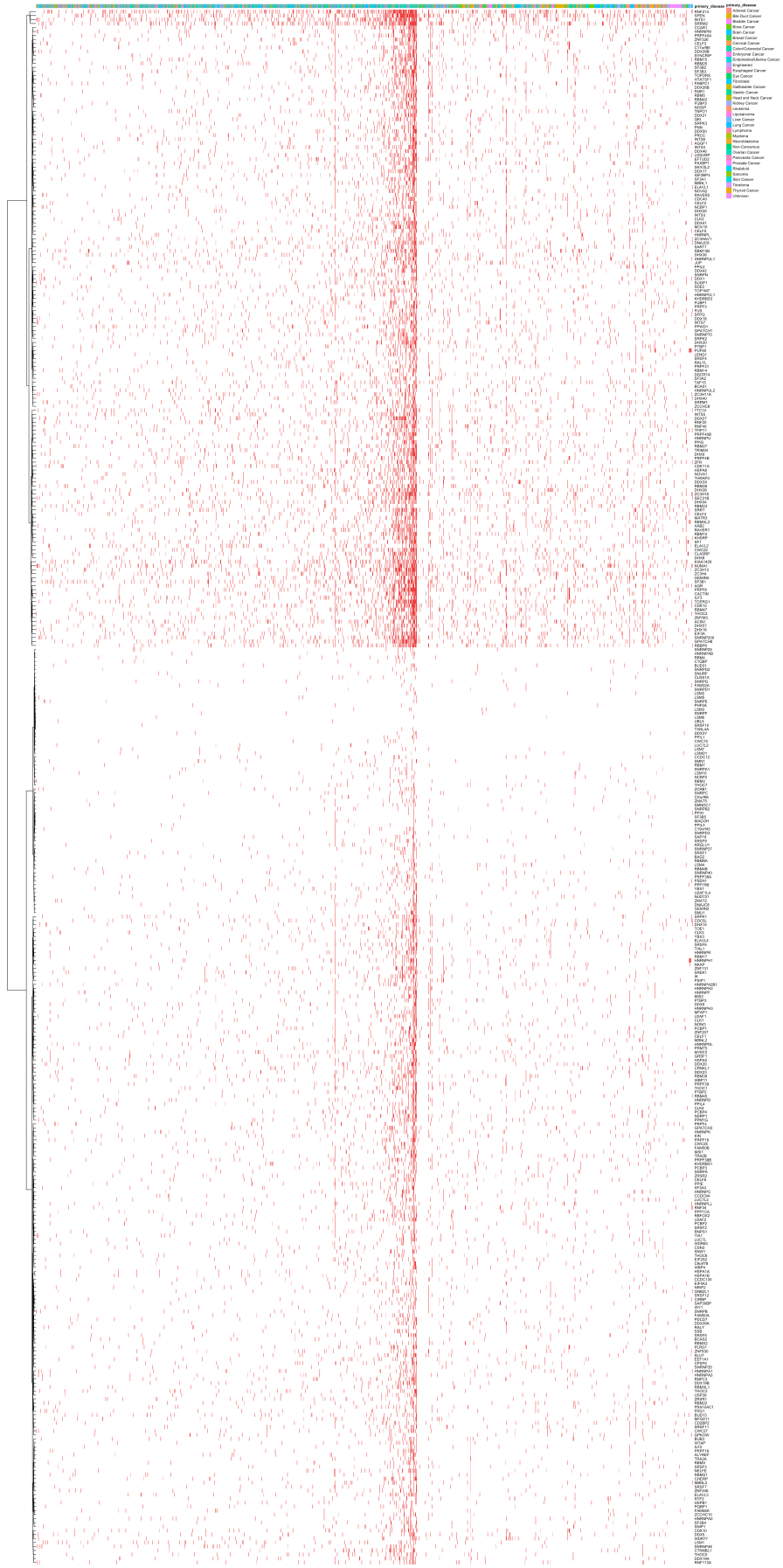

Figure A.7: Heatmap showing the all mutations occurring on the splicing factors across the studied cell lines.

B

Appendix 2 - Transcriptional differences related to spliceosome deficiencies in cancer

B.1 MYC targets gene sets

List of genes in the MYC target gene sets

SYMBOL	Description
ABCE1	ATP binding cassette subfamily E member 1
ACP1	acid phosphatase 1
AIMP2	aminoacyl tRNA synthetase complex interacting multifunctional protein 2
AP3S1	adaptor related protein complex 3 subunit sigma 1
APEX1	apurinic/apyrimidinic endodeoxyribonuclease 1
BUB3	BUB3 mitotic checkpoint protein
BYSL	bystin like
C1QBP	complement C1q binding protein
CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase
CANX	calnexin
CBX3	chromobox 3
CCNA2	cyclin A2
CCT2	chaperonin containing TCP1 subunit 2
CCT3	chaperonin containing TCP1 subunit 3
CCT4	chaperonin containing TCP1 subunit 4
CCT5	chaperonin containing TCP1 subunit 5
CCT7	chaperonin containing TCP1 subunit 7
CDC20	cell division cycle 20
CDC45	cell division cycle 45
CDK2	cyclin dependent kinase 2
CDK4	cyclin dependent kinase 4
CLNS1A	chloride nucleotide-sensitive channel 1A
CNBP	CCHC-type zinc finger nucleic acid binding protein
COPS5	COP9 signalosome subunit 5
COX5A	cytochrome c oxidase subunit 5A
CSTF2	cleavage stimulation factor subunit 2
CTPS1	CTP synthase 1
CUL1	cullin 1
CYC1	cytochrome c1
DCTPP1	dCTP pyrophosphatase 1
DDX18	DEAD-box helicase 18
DDX21	DExD-box helicase 21
DEK	DEK proto-oncogene
DHX15	DEAH-box helicase 15
DUSP2	dual specificity phosphatase 2
DUT	deoxyuridine triphosphatase
EEF1B2	eukaryotic translation elongation factor 1 beta 2
EIF1AX	eukaryotic translation initiation factor 1A X-linked
EIF2S1	eukaryotic translation initiation factor 2 subunit alpha
EIF2S2	eukaryotic translation initiation factor 2 subunit beta
EIF3B	eukaryotic translation initiation factor 3 subunit B
EIF3D	eukaryotic translation initiation factor 3 subunit D

Continued on next page

Table B.1 – continued from previous page

SYMBOL	Description
EIF3J	eukaryotic translation initiation factor 3 subunit J
EIF4A1	eukaryotic translation initiation factor 4A1
EIF4E	eukaryotic translation initiation factor 4E
EIF4G2	eukaryotic translation initiation factor 4 gamma 2
EIF4H	eukaryotic translation initiation factor 4H
EPRS1	glutamyl-prolyl-tRNA synthetase 1
ERH	ERH mRNA splicing and mitosis factor
ETF1	eukaryotic translation termination factor 1
EXOSC5	exosome component 5
EXOSC7	exosome component 7
FAM120A	family with sequence similarity 120A
FARSA	phenylalanyl-tRNA synthetase subunit alpha
FBL	fibrillarin
G3BP1	G3BP stress granule assembly factor 1
GLO1	glyoxalase I
GNL3	G protein nucleolar 3
GOT2	glutamic-oxaloacetic transaminase 2
GRWD1	glutamate rich WD repeat containing 1
GSPT1	G1 to S phase transition 1
H2AZ1	H2A.Z variant histone 1
HDAC2	histone deacetylase 2
HDDC2	HD domain containing 2
HDGF	heparin binding growth factor
HK2	hexokinase 2
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1
HNRNPA3	heterogeneous nuclear ribonucleoprotein A3
HNRNPC	heterogeneous nuclear ribonucleoprotein C
HNRNPD	heterogeneous nuclear ribonucleoprotein D
HNRNPR	heterogeneous nuclear ribonucleoprotein R
HNRNPU	heterogeneous nuclear ribonucleoprotein U
HPRT1	hypoxanthine phosphoribosyltransferase 1
HSP90AB1	heat shock protein 90 alpha family class B member 1
HSPD1	heat shock protein family D (Hsp60) member 1
HSPE1	heat shock protein family E (Hsp10) member 1
IARS1	isoleucyl-tRNA synthetase 1
IFRD1	interferon related developmental regulator 1
ILF2	interleukin enhancer binding factor 2
IMP4	IMP U3 small nucleolar ribonucleoprotein 4
IMPDH2	inosine monophosphate dehydrogenase 2
IPO4	importin 4
KARS1	lysyl-tRNA synthetase 1
KPNA2	karyopherin subunit alpha 2
KPNB1	karyopherin subunit beta 1
LAS1L	LAS1 like ribosome biogenesis factor
LDHA	lactate dehydrogenase A
LSM2	LSM2 homolog, U6 small nuclear RNA and mRNA degradation associated
LSM7	LSM7 homolog, U6 small nuclear RNA and mRNA degradation associated
MAD2L1	mitotic arrest deficient 2 like 1
MAP3K6	mitogen-activated protein kinase kinase kinase 6
MCM2	minichromosome maintenance complex component 2
MCM4	minichromosome maintenance complex component 4
MCM5	minichromosome maintenance complex component 5
MCM6	minichromosome maintenance complex component 6
MCM7	minichromosome maintenance complex component 7
MPHOSPH10	M-phase phosphoprotein 10
MRPL23	mitochondrial ribosomal protein L23
MRPL9	mitochondrial ribosomal protein L9
MRPS18B	mitochondrial ribosomal protein S18B
MRTO4	MRT4 homolog, ribosome maturation factor
MYBBP1A	MYB binding protein 1a
MYC	MYC proto-oncogene, bHLH transcription factor
NAP1L1	nucleosome assembly protein 1 like 1
NCBP1	nuclear cap binding protein subunit 1
NCBP2	nuclear cap binding protein subunit 2
NDUFAB1	NADH:ubiquinone oxidoreductase subunit AB1
NDUFAF4	NADH:ubiquinone oxidoreductase complex assembly factor 4
NHP2	NHP2 ribonucleoprotein
NIP7	nucleolar pre-rRNA processing protein NIP7
NME1	NME/NM23 nucleoside diphosphate kinase 1

Continued on next page

Table B.1 – continued from previous page

SYMBOL	Description
NOC4L	nucleolar complex associated 4 homolog
NOLC1	nucleolar and coiled-body phosphoprotein 1
NOP16	NOP16 nucleolar protein
NOP2	NOP2 nucleolar protein
NOP56	NOP56 ribonucleoprotein
NPM1	nucleophosmin 1
ODC1	ornithine decarboxylase 1
ORC2	origin recognition complex subunit 2
PA2G4	proliferation-associated 2G4
PABPC1	poly(A) binding protein cytoplasmic 1
PABPC4	poly(A) binding protein cytoplasmic 4
PCBP1	poly(rC) binding protein 1
PCNA	proliferating cell nuclear antigen
PES1	pescadillo ribosomal biogenesis factor 1
PGK1	phosphoglycerate kinase 1
PHB	prohibitin
PHB2	prohibitin 2
PLK1	polo like kinase 1
PLK4	polo like kinase 4
POLD2	DNA polymerase delta 2, accessory subunit
POLE3	DNA polymerase epsilon 3, accessory subunit
PPAN	peter pan homolog
PPIA	peptidylprolyl isomerase A
PPM1G	protein phosphatase, Mg2+/Mn2+ dependent 1G
PPRC1	PPARG related coactivator 1
PRDX3	peroxiredoxin 3
PRDX4	peroxiredoxin 4
PRMT3	protein arginine methyltransferase 3
PRPF31	pre-mRNA processing factor 31
PRPS2	phosphoribosyl pyrophosphate synthetase 2
PSMA1	proteasome 20S subunit alpha 1
PSMA2	proteasome 20S subunit alpha 2
PSMA4	proteasome 20S subunit alpha 4
PSMA6	proteasome 20S subunit alpha 6
PSMA7	proteasome 20S subunit alpha 7
PSMB2	proteasome 20S subunit beta 2
PSMB3	proteasome 20S subunit beta 3
PSMC4	proteasome 26S subunit, ATPase 4
PSMC6	proteasome 26S subunit, ATPase 6
PSMD1	proteasome 26S subunit, non-ATPase 1
PSMD14	proteasome 26S subunit, non-ATPase 14
PSMD3	proteasome 26S subunit, non-ATPase 3
PSMD7	proteasome 26S subunit, non-ATPase 7
PSMD8	proteasome 26S subunit, non-ATPase 8
PTGES3	prostaglandin E synthase 3
PUS1	pseudouridine synthase 1
PWP1	PWP1 homolog, endonuclein
RABEPK	Rab9 effector protein with kelch motifs
RACK1	receptor for activated C kinase 1
RAD23B	RAD23 homolog B, nucleotide excision repair protein
RAN	RAN, member RAS oncogene family
RANBP1	RAN binding protein 1
RCL1	RNA terminal phosphate cyclase like 1
RFC4	replication factor C subunit 4
RNPS1	RNA binding protein with serine rich domain 1
RPL14	ribosomal protein L14
RPL18	ribosomal protein L18
RPL22	ribosomal protein L22
RPL34	ribosomal protein L34
RPL6	ribosomal protein L6
RPLP0	ribosomal protein lateral stalk subunit P0
RPS10	ribosomal protein S10
RPS2	ribosomal protein S2
RPS3	ribosomal protein S3
RPS5	ribosomal protein S5
RPS6	ribosomal protein S6
RRM1	ribonucleotide reductase catalytic subunit M1
RRP12	ribosomal RNA processing 12 homolog
RRP9	ribosomal RNA processing 9, U3 small nucleolar RNA binding protein
RSL1D1	ribosomal L1 domain containing 1

Continued on next page

Table B.1 – continued from previous page

SYMBOL	Description
RUVBL2	RuvB like AAA ATPase 2
SERBP1	SERPINE1 mRNA binding protein 1
SET	SET nuclear proto-oncogene
SF3A1	splicing factor 3a subunit 1
SF3B3	splicing factor 3b subunit 3
SLC19A1	solute carrier family 19 member 1
SLC25A3	solute carrier family 25 member 3
SLC29A2	solute carrier family 29 member 2
SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily c member 1
SNRPA	small nuclear ribonucleoprotein polypeptide A
SNRPA1	small nuclear ribonucleoprotein polypeptide A'
SNRPB2	small nuclear ribonucleoprotein polypeptide B2
SNRPD1	small nuclear ribonucleoprotein D1 polypeptide
SNRPD2	small nuclear ribonucleoprotein D2 polypeptide
SNRPD3	small nuclear ribonucleoprotein D3 polypeptide
SNRPG	small nuclear ribonucleoprotein polypeptide G
SORD	sorbitol dehydrogenase
SRM	spermidine synthase
SRPK1	SRSF protein kinase 1
SRSF1	serine and arginine rich splicing factor 1
SRSF2	serine and arginine rich splicing factor 2
SRSF3	serine and arginine rich splicing factor 3
SRSF7	serine and arginine rich splicing factor 7
SSB	small RNA binding exonuclease protection factor La
SSBP1	single stranded DNA binding protein 1
STARD7	StAR related lipid transfer domain containing 7
SUPV3L1	Suv3 like RNA helicase
SYNCRIP	synaptotagmin binding cytoplasmic RNA interacting protein
TARDBP	TAR DNA binding protein
TBRG4	transforming growth factor beta regulator 4
TCOF1	treacle ribosome biogenesis factor 1
TCP1	t-complex 1
TFB2M	transcription factor B2, mitochondrial
TFDP1	transcription factor Dp-1
TMEM97	transmembrane protein 97
TOMM70	translocase of outer mitochondrial membrane 70
TRA2B	transformer 2 beta homolog
TRIM28	tripartite motif containing 28
TUFM	Tu translation elongation factor, mitochondrial
TXNL4A	thioredoxin like 4A
TYMS	thymidylate synthetase
U2AF1	U2 small nuclear RNA auxiliary factor 1
UBA2	ubiquitin like modifier activating enzyme 2
UBE2E1	ubiquitin conjugating enzyme E2 E1
UBE2L3	ubiquitin conjugating enzyme E2 L3
UNG	uracil DNA glycosylase
USP1	ubiquitin specific peptidase 1
UTP20	UTP20 small subunit processome component
VBP1	VHL binding protein 1
VDAC1	voltage dependent anion channel 1
VDAC3	voltage dependent anion channel 3
WDR43	WD repeat domain 43
WDR74	WD repeat domain 74
XPO1	exportin 1
XPOT	exportin for tRNA
XRCC6	X-ray repair cross complementing 6
YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon
YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta

C

Appendix 3 - Predicting individual mutation-associated splicing events using SpliceAI

C.1 Preparation to run SpliceAI

As explained in section 3.6.1, the CSV file downloaded from DepMap was parsed into a VCF file. Then, the file was sorted and indexed using *bgzip* and *tabix*, both included in the *samtools* suit [25] and installed locally in a conda environment.

```
#Parsing to a VCF file
awk -F ',' 'BEGIN {printf("##fileformat=VCFv4.2
\n##reference=GRCh37/hg19\n#CHROM\tPOS\tID\tREF\tALT\tQUAL\
tFILTER\tINFO\n");}
NR > 1 {printf("%s\t%s\t%s\t%s\t%s\t.\t.\n",
$4,$5,$36,$10,$11);}'
../path/to/my/file/filename.csv > /path/to/output.vcf

# Sorting the file
## Print headers
grep "^#" CCLE.vcf > CCLE_out.vcf
## Sort by chromosome and position
grep -v "^#" CCLE.vcf| sort -k1,1V -k2,2g >> CCLE_out.vcf

#Indexing
bgzip CCLE_out.vcf
tabix CCLE_out.vcf

#To run SpliceAI on the cluster
spliceai -I mutations_CCLE.vcf -O output.vcf
-R hg19.fa -A grch37

#Where the mutations CCLE file is the somatic mutations
#dataset generated by CCLE, the output file contains the
#probabilities calculated by SpliceAI and its position and
#the hg19 fasta file corresponds to the
#Genome Reference Consortium Human Build 37 (GRCh37).
```

C.2 SpliceAI

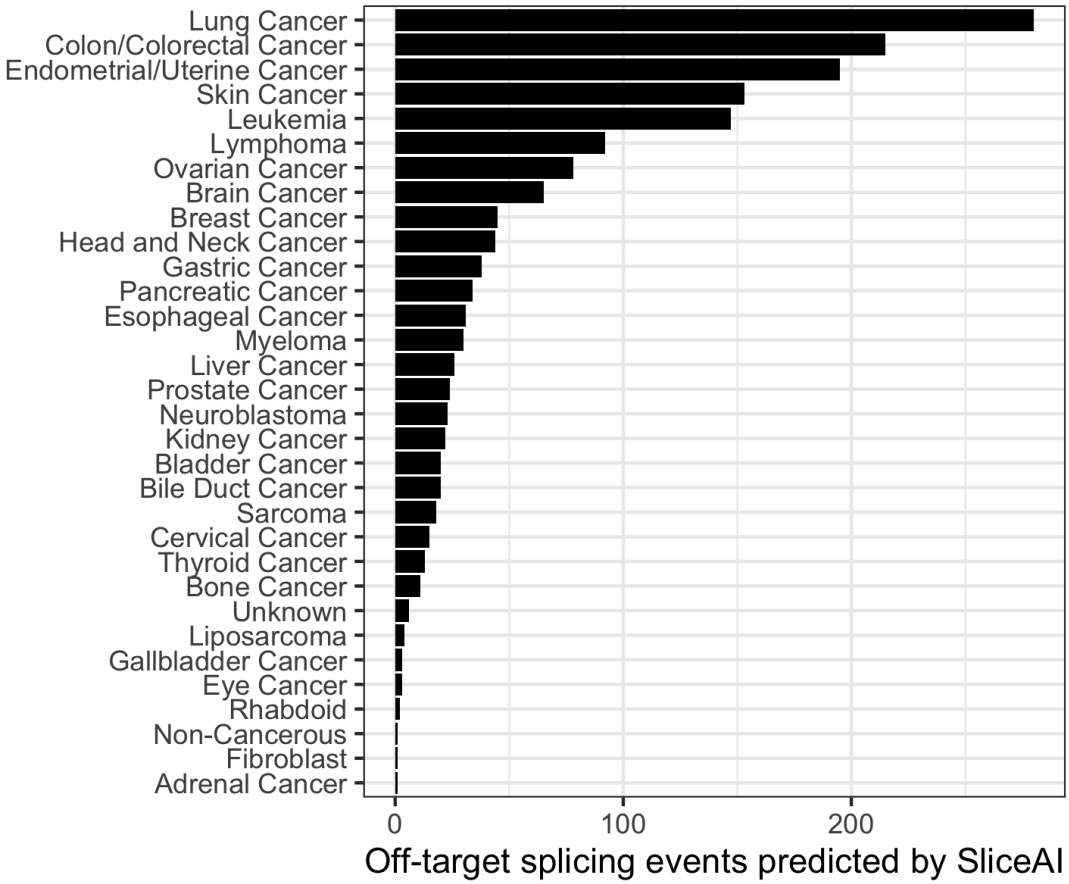

Figure C.1: Number of off-target mutation events predicted by SpliceAI per cancer

List of off-target splicing events predicted by SpliceAI

DepMap ID	SYMBOL	DISEASE
ACH-000512	OR2L13	Myeloma
ACH-001328	OR2W3	Skin Cancer
ACH-000941	MTERF4	Endometrial/Uterine Cancer
ACH-000946	SNED1	Endometrial/Uterine Cancer
ACH-000579	CROCC2	Skin Cancer
ACH-000314	AL646016.1	Lung Cancer
ACH-000851	AL646016.1	Lung Cancer
ACH-000942	UBE2F-SCLY	Leukemia
ACH-000608	RYR2	Ovarian Cancer
ACH-000901	RYR2	Lung Cancer
ACH-001518	ERO1B	Endometrial/Uterine Cancer
ACH-002257	ERO1B	Brain Cancer
ACH-000316	ERO1B	Liver Cancer
ACH-000928	ERO1B	Endometrial/Uterine Cancer
ACH-000955	RP5-862P8.2	Colon/Colorectal Cancer
ACH-001321	DAW1	Thyroid Cancer
ACH-000517	FARSB	Pancreatic Cancer
ACH-000793	RP11-33O4.2	Gastric Cancer
ACH-000787	RP11-33O4.2	Lung Cancer
ACH-000972	RP11-33O4.2	Endometrial/Uterine Cancer
ACH-000730	RP11-33O4.2	Skin Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000548	RP11-33O4.2	Head and Neck Cancer
ACH-001664	RP11-33O4.2	Gastric Cancer
ACH-001517	RP11-33O4.2	Endometrial/Uterine Cancer
ACH-000941	RP11-33O4.2	Endometrial/Uterine Cancer
ACH-000991	FCMR	Colon/Colorectal Cancer
ACH-001517	RAB29	Endometrial/Uterine Cancer
ACH-000985	RAB29	Colon/Colorectal Cancer
ACH-000274	NBEAL1	Skin Cancer
ACH-000157	RNPEP	Lymphoma
ACH-000957	HSPE1-MOB4	Colon/Colorectal Cancer
ACH-000969	HSPE1-MOB4	Colon/Colorectal Cancer
ACH-000692	RUBCN	Head and Neck Cancer
ACH-000568	RP4-608O15.3	Breast Cancer
ACH-000987	RP4-608O15.3	Skin Cancer
ACH-000215	RP4-608O15.3	Brain Cancer
ACH-001137	RP4-608O15.3	Lung Cancer
ACH-000219	RP4-608O15.3	Skin Cancer
ACH-002001	RP4-608O15.3	Skin Cancer
ACH-002003	RP4-608O15.3	Skin Cancer
ACH-001081	TM4SF19-TCTEX1D2	Colon/Colorectal Cancer
ACH-000991	TM4SF19-TCTEX1D2	Colon/Colorectal Cancer
ACH-002253	RP11-447L10.1	Lymphoma
ACH-002253	TCTEX1D2	Lymphoma
ACH-002253	TM4SF19-TCTEX1D2	Lymphoma
ACH-000380	NEMP2	Myeloma
ACH-001339	INPP1	Skin Cancer
ACH-001796	P3H2	Liposarcoma
ACH-001498	P3H2	Lymphoma
ACH-000169	CENPU	Sarcoma
ACH-001790	CENPU	Sarcoma
ACH-000985	CENPU	Colon/Colorectal Cancer
ACH-001151	NCKAP1	Ovarian Cancer
ACH-000274	USP13	Skin Cancer
ACH-000960	TTN	Leukemia
ACH-001339	PIK3CA	Skin Cancer
ACH-000901	ASB5	Lung Cancer
ACH-000888	AC009336.19	Lung Cancer
ACH-000888	HOXD4	Lung Cancer
ACH-000608	CDHR2	Ovarian Cancer
ACH-000157	SFXN1	Lymphoma
ACH-000662	CDCA7	Lung Cancer
ACH-000562	AC013461.1	Lung Cancer
ACH-000198	AC013461.1	Leukemia
ACH-000077	AC013461.1	Lymphoma
ACH-000590	AC013461.1	Lung Cancer
ACH-000996	AC013461.1	Endometrial/Uterine Cancer
ACH-000011	AC013461.1	Bladder Cancer
ACH-000026	AC013461.1	Bladder Cancer
ACH-001137	HMP19	Lung Cancer
ACH-000608	NEK1	Ovarian Cancer
ACH-000699	BBS5	Breast Cancer
ACH-000274	ANXA10	Skin Cancer
ACH-000662	TTC21B	Lung Cancer
ACH-000468	QKI	Pancreatic Cancer
ACH-000025	HMMR	Brain Cancer
ACH-000900	CCDC190	Lung Cancer
ACH-000458	PLG	Skin Cancer
ACH-000999	RP11-544M22.13	Colon/Colorectal Cancer
ACH-000123	SLC22A3	Ovarian Cancer
ACH-002163	LY75	Breast Cancer
ACH-002166	LY75	Skin Cancer
ACH-000927	LY75	Breast Cancer
ACH-001344	LY75	Neuroblastoma
ACH-000996	LY75	Endometrial/Uterine Cancer
ACH-002100	LY75	Skin Cancer
ACH-002193	LY75	Lung Cancer
ACH-000986	LY75	Colon/Colorectal Cancer
ACH-000782	LY75	Leukemia
ACH-001735	LY75	Leukemia
ACH-000993	LY75-CD302	Endometrial/Uterine Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000784	LY75	Esophageal Cancer
ACH-000719	LY75	Ovarian Cancer
ACH-000977	CD302	Prostate Cancer
ACH-000999	RP11-574F21.3	Colon/Colorectal Cancer
ACH-000921	RP11-574F21.3	Lung Cancer
ACH-000638	RP11-574F21.3	Lung Cancer
ACH-000997	RP11-432B6.3	Colon/Colorectal Cancer
ACH-000901	SMC4	Lung Cancer
ACH-000946	RP11-432B6.3	Endometrial/Uterine Cancer
ACH-000787	RP11-432B6.3	Lung Cancer
ACH-001517	RP11-432B6.3	Endometrial/Uterine Cancer
ACH-000954	RP11-432B6.3	Endometrial/Uterine Cancer
ACH-000941	RP11-432B6.3	Endometrial/Uterine Cancer
ACH-001333	SCHIP1	Cervical Cancer
ACH-000984	ACKR1	Endometrial/Uterine Cancer
ACH-000718	PQLC2L	Lung Cancer
ACH-000992	AC026407.1	Head and Neck Cancer
ACH-000804	GLMP	Neuroblastoma
ACH-000608	THBS3	Ovarian Cancer
ACH-000757	DCST2	Lung Cancer
ACH-000978	CNKSR3	Endometrial/Uterine Cancer
ACH-000317	CNKSR3	Kidney Cancer
ACH-000963	CMC4	Colon/Colorectal Cancer
ACH-000965	GATB	Endometrial/Uterine Cancer
ACH-001328	NEB	Skin Cancer
ACH-000355	TMEM14EP	Lung Cancer
ACH-000907	ARMT1	Kidney Cancer
ACH-000404	ERICH6	Skin Cancer
ACH-001127	ERICH6	Leukemia
ACH-000858	ERICH6	Lung Cancer
ACH-000972	ERICH6	Endometrial/Uterine Cancer
ACH-000504	ERICH6	Brain Cancer
ACH-000969	CIART	Colon/Colorectal Cancer
ACH-000563	PRMT9	Lung Cancer
ACH-000749	PRMT9	Lung Cancer
ACH-000946	AF011889.5	Endometrial/Uterine Cancer
ACH-000963	AF011889.5	Colon/Colorectal Cancer
ACH-000759	AF011889.5	Breast Cancer
ACH-000988	AF011889.5	Endometrial/Uterine Cancer
ACH-002130	RP11-449H3.3	Lung Cancer
ACH-000987	RP11-449H3.3	Skin Cancer
ACH-000977	RP11-449H3.3	Prostate Cancer
ACH-001328	HHIP	Skin Cancer
ACH-000989	NBPF10	Colon/Colorectal Cancer
ACH-000911	HGH1	Gastric Cancer
ACH-000681	WDR97	Lung Cancer
ACH-001627	WDR97	Prostate Cancer
ACH-000558	WDR97	Brain Cancer
ACH-000236	WDR97	Colon/Colorectal Cancer
ACH-001991	RP3-468K18.7	Ovarian Cancer
ACH-001328	CYP11B2	Skin Cancer
ACH-000781	ADGRB1	Lung Cancer
ACH-000628	ADGRB1	Lung Cancer
ACH-002309	ADGRG6	Sarcoma
ACH-001518	ADGRG6	Endometrial/Uterine Cancer
ACH-000901	TRPV5	Lung Cancer
ACH-000751	ADGRG6	Leukemia
ACH-000987	MGAM2	Skin Cancer
ACH-002119	MGAM2	Bone Cancer
ACH-001524	MGAM2	Skin Cancer
ACH-002004	MGAM2	Skin Cancer
ACH-000993	TBC1D9	Endometrial/Uterine Cancer
ACH-000993	CACNA1B	Endometrial/Uterine Cancer
ACH-000757	CACNA1B	Lung Cancer
ACH-000940	PCDHA1	Endometrial/Uterine Cancer
ACH-000940	PCDHA2	Endometrial/Uterine Cancer
ACH-000940	PCDHA3	Endometrial/Uterine Cancer
ACH-000940	PCDHA4	Endometrial/Uterine Cancer
ACH-000940	PCDHA5	Endometrial/Uterine Cancer
ACH-000940	PCDHA6	Endometrial/Uterine Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000940	PCDHA7	Endometrial/Uterine Cancer
ACH-000940	PCDHA8	Endometrial/Uterine Cancer
ACH-000940	PCDHA9	Endometrial/Uterine Cancer
ACH-000940	PCDHA10	Endometrial/Uterine Cancer
ACH-000940	PCDHA11	Endometrial/Uterine Cancer
ACH-000940	PCDHA12	Endometrial/Uterine Cancer
ACH-001559	PCDHA1	Lung Cancer
ACH-001559	PCDHA2	Lung Cancer
ACH-001559	PCDHA3	Lung Cancer
ACH-001559	PCDHA4	Lung Cancer
ACH-001559	PCDHA5	Lung Cancer
ACH-001559	PCDHA6	Lung Cancer
ACH-001559	PCDHA7	Lung Cancer
ACH-001559	PCDHA8	Lung Cancer
ACH-001559	PCDHA9	Lung Cancer
ACH-001559	PCDHA10	Lung Cancer
ACH-001559	PCDHA11	Lung Cancer
ACH-000838	PCDHA1	Myeloma
ACH-002038	PCDHA1	Ovarian Cancer
ACH-000838	PCDHA2	Myeloma
ACH-000838	PCDHA3	Myeloma
ACH-000838	PCDHA4	Myeloma
ACH-000838	PCDHA5	Myeloma
ACH-000838	PCDHA6	Myeloma
ACH-000838	PCDHA7	Myeloma
ACH-000838	PCDHA8	Myeloma
ACH-000838	PCDHA9	Myeloma
ACH-000838	PCDHA10	Myeloma
ACH-002038	PCDHA2	Ovarian Cancer
ACH-002038	PCDHA3	Ovarian Cancer
ACH-002038	PCDHA4	Ovarian Cancer
ACH-002038	PCDHA5	Ovarian Cancer
ACH-002038	PCDHA6	Ovarian Cancer
ACH-002038	PCDHA7	Ovarian Cancer
ACH-002038	PCDHA8	Ovarian Cancer
ACH-002038	PCDHA9	Ovarian Cancer
ACH-002038	PCDHA10	Ovarian Cancer
ACH-000652	PCDHA1	Pancreatic Cancer
ACH-000652	PCDHA2	Pancreatic Cancer
ACH-000652	PCDHA3	Pancreatic Cancer
ACH-000652	PCDHA4	Pancreatic Cancer
ACH-000652	PCDHA5	Pancreatic Cancer
ACH-000652	PCDHA6	Pancreatic Cancer
ACH-000652	PCDHA7	Pancreatic Cancer
ACH-000652	PCDHA8	Pancreatic Cancer
ACH-000984	PCDHA1	Endometrial/Uterine Cancer
ACH-000984	PCDHA2	Endometrial/Uterine Cancer
ACH-000984	PCDHA3	Endometrial/Uterine Cancer
ACH-000984	PCDHA4	Endometrial/Uterine Cancer
ACH-000981	PCDHA1	Leukemia
ACH-000981	PCDHA2	Leukemia
ACH-000981	PCDHA3	Leukemia
ACH-000981	PCDHA4	Leukemia
ACH-000960	ENTPD2	Leukemia
ACH-000398	ENTPD2	Lymphoma
ACH-000727	RP11-229P13.27	Lung Cancer
ACH-002245	ANKHD1-EIF4EBP3	Lymphoma
ACH-000860	ANKHD1-EIF4EBP3	Lung Cancer
ACH-001523	FBXW5	Skin Cancer
ACH-000171	KDM7A	Kidney Cancer
ACH-000123	SLC4A9	Ovarian Cancer
ACH-001433	RP11-216L13.17	Sarcoma
ACH-001849	RP11-216L13.17	Bile Duct Cancer
ACH-001433	CCDC183	Sarcoma
ACH-001849	CCDC183	Bile Duct Cancer
ACH-000594	RP11-216L13.16	Lung Cancer
ACH-001377	LUC7L2	Pancreatic Cancer
ACH-000925	C7orf55	Lung Cancer
ACH-000925	C7orf55-LUC7L2	Lung Cancer
ACH-001061	ARFGEF3	Colon/Colorectal Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000997	ARFGEF3	Colon/Colorectal Cancer
ACH-000989	ARFGEF3	Colon/Colorectal Cancer
ACH-000876	ARFGEF3	Breast Cancer
ACH-000901	DGKI	Lung Cancer
ACH-002370	ADGRG4	Unknown
ACH-001529	ADGRG4	Endometrial/Uterine Cancer
ACH-000998	ADGRG4	Colon/Colorectal Cancer
ACH-002397	ADGRG4	Thyroid Cancer
ACH-000671	CFAP77	Liver Cancer
ACH-001345	CFAP77	Colon/Colorectal Cancer
ACH-000274	AL161645.2	Skin Cancer
ACH-002389	MTG1	Neuroblastoma
ACH-000676	MTG1	Brain Cancer
ACH-000781	MTG1	Lung Cancer
ACH-000914	PRAP1	Lymphoma
ACH-000950	CFAP46	Colon/Colorectal Cancer
ACH-000970	CFAP46	Colon/Colorectal Cancer
ACH-000941	CFAP46	Endometrial/Uterine Cancer
ACH-002127	CFAP46	Lung Cancer
ACH-000937	CFAP46	Leukemia
ACH-000880	CFAP46	Gastric Cancer
ACH-001061	CFAP46	Colon/Colorectal Cancer
ACH-000997	CFAP46	Colon/Colorectal Cancer
ACH-000986	CFAP46	Colon/Colorectal Cancer
ACH-001835	CFAP46	Bile Duct Cancer
ACH-000998	CFAP46	Colon/Colorectal Cancer
ACH-000952	CFAP46	Prostate Cancer
ACH-000929	CFAP46	Lung Cancer
ACH-000757	CFAP46	Lung Cancer
ACH-001328	CFAP46	Skin Cancer
ACH-000641	CFAP46	Leukemia
ACH-002005	CFAP46	Skin Cancer
ACH-001081	CT55	Colon/Colorectal Cancer
ACH-000814	CT55	Skin Cancer
ACH-002340	CT55	Neuroblastoma
ACH-000992	JADE2	Head and Neck Cancer
ACH-001623	JADE2	Brain Cancer
ACH-000662	ANHX	Lung Cancer
ACH-000894	CTD-2140B24.4	Lung Cancer
ACH-001328	C3orf36	Skin Cancer
ACH-001991	CTD-2410N18.5	Ovarian Cancer
ACH-000990	LYPD1	Endometrial/Uterine Cancer
ACH-002302	RP11-240B13.2	Leukemia
ACH-000941	RP11-240B13.2	Endometrial/Uterine Cancer
ACH-000941	HHLA1	Endometrial/Uterine Cancer
ACH-000993	OC90	Endometrial/Uterine Cancer
ACH-000993	RP11-240B13.2	Endometrial/Uterine Cancer
ACH-000274	OC90	Skin Cancer
ACH-001456	RP11-240B13.2	Colon/Colorectal Cancer
ACH-001041	RP11-240B13.2	Lung Cancer
ACH-001496	RP11-240B13.2	Esophageal Cancer
ACH-000274	RP11-240B13.2	Skin Cancer
ACH-000672	RP11-240B13.2	Lung Cancer
ACH-001536	RP11-240B13.2	Bile Duct Cancer
ACH-001791	RP11-240B13.2	Liposarcoma
ACH-001559	RP11-240B13.2	Lung Cancer
ACH-001568	RP11-240B13.2	Skin Cancer
ACH-001632	RP11-240B13.2	Ovarian Cancer
ACH-001856	RP11-240B13.2	Bile Duct Cancer
ACH-001751	RP11-240B13.2	Sarcoma
ACH-001645	RP11-240B13.2	Skin Cancer
ACH-000556	RP11-240B13.2	Cervical Cancer
ACH-001654	RP11-240B13.2	Esophageal Cancer
ACH-001414	RP11-240B13.2	Bladder Cancer
ACH-002002	RP11-240B13.2	Skin Cancer
ACH-000637	RP11-240B13.2	Esophageal Cancer
ACH-000820	RP11-240B13.2	Colon/Colorectal Cancer
ACH-000416	RP11-240B13.2	Lung Cancer
ACH-000921	RP11-240B13.2	Lung Cancer
ACH-000662	ADCY8	Lung Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000026	ADGRD1	Bladder Cancer
ACH-002275	ADGRD1	Leukemia
ACH-001414	CTB-127M13.1	Bladder Cancer
ACH-000813	CTB-127M13.1	Lung Cancer
ACH-000900	CTB-127M13.1	Lung Cancer
ACH-000940	CTB-127M13.1	Endometrial/Uterine Cancer
ACH-000984	CTB-127M13.1	Endometrial/Uterine Cancer
ACH-000973	STK26	Bladder Cancer
ACH-002238	CTC-432M15.3	Leukemia
ACH-001650	CTC-432M15.3	Cervical Cancer
ACH-002026	CTC-432M15.3	Endometrial/Uterine Cancer
ACH-000974	CTC-432M15.3	Endometrial/Uterine Cancer
ACH-000164	CTC-432M15.3	Pancreatic Cancer
ACH-000959	CTC-432M15.3	Colon/Colorectal Cancer
ACH-000274	COL6A5	Skin Cancer
ACH-000928	CPA1	Endometrial/Uterine Cancer
ACH-000992	RALGPS1	Head and Neck Cancer
ACH-000037	RALGPS1	Sarcoma
ACH-000995	JADE1	Leukemia
ACH-002258	JADE1	Leukemia
ACH-000999	ABHD18	Colon/Colorectal Cancer
ACH-000250	ISY1-RAB43	Kidney Cancer
ACH-000736	EFCC1	Gastric Cancer
ACH-000901	SMARCA1	Lung Cancer
ACH-000941	RP3-403A15.5	Endometrial/Uterine Cancer
ACH-000841	EDRF1	Lung Cancer
ACH-001550	ADGRD2	Skin Cancer
ACH-001555	ADGRD2	Lung Cancer
ACH-000248	ADGRD2	Breast Cancer
ACH-000017	ADGRD2	Breast Cancer
ACH-001610	ADGRD2	Brain Cancer
ACH-001517	ADGRD2	Endometrial/Uterine Cancer
ACH-000954	ADGRD2	Endometrial/Uterine Cancer
ACH-000941	ADGRD2	Endometrial/Uterine Cancer
ACH-000988	ADGRD2	Endometrial/Uterine Cancer
ACH-001518	METTL10	Endometrial/Uterine Cancer
ACH-002253	CFAP100	Lymphoma
ACH-000865	CFAP100	Esophageal Cancer
ACH-000398	DNAH10	Lymphoma
ACH-000157	CCDC14	Lymphoma
ACH-000901	MYLK	Lung Cancer
ACH-000954	RP11-512M8.5	Endometrial/Uterine Cancer
ACH-000959	RP11-512M8.5	Colon/Colorectal Cancer
ACH-000681	NIFK	Lung Cancer
ACH-002337	NIFK	Lung Cancer
ACH-000345	PLPP4	Neuroblastoma
ACH-000025	TBC1D32	Brain Cancer
ACH-001118	COX6A1	Brain Cancer
ACH-000152	COX6A1	Brain Cancer
ACH-001419	GCN1	Breast Cancer
ACH-000984	GCN1	Endometrial/Uterine Cancer
ACH-000334	GCN1	Lymphoma
ACH-001664	GCN1	Gastric Cancer
ACH-000982	CFAP221	Colon/Colorectal Cancer
ACH-001345	CFAP221	Colon/Colorectal Cancer
ACH-000398	HGD	Lymphoma
ACH-001303	CFAP221	Neuroblastoma
ACH-001145	CFAP221	Ovarian Cancer
ACH-000901	PLA1A	Lung Cancer
ACH-000829	MFRP	Myeloma
ACH-000123	VPS11	Ovarian Cancer
ACH-000960	SEPT6	Leukemia
ACH-001365	DCBLD1	Lung Cancer
ACH-000990	LSM8	Endometrial/Uterine Cancer
ACH-000999	LSM8	Colon/Colorectal Cancer
ACH-000901	DOCK11	Lung Cancer
ACH-000999	RP1-179P9.3	Colon/Colorectal Cancer
ACH-000568	FXYD6-FXYD2	Breast Cancer
ACH-001991	RP1-179P9.3	Ovarian Cancer
ACH-000684	RP1-179P9.3	Kidney Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000437	RP1-179P9.3	Brain Cancer
ACH-001127	RP1-179P9.3	Leukemia
ACH-001339	ROS1	Skin Cancer
ACH-001339	RP1-179P9.3	Skin Cancer
ACH-000662	RP1-179P9.3	Lung Cancer
ACH-000990	RP1-179P9.3	Endometrial/Uterine Cancer
ACH-000038	ZPR1	Lymphoma
ACH-000295	ZPR1	Leukemia
ACH-000396	ZPR1	Bladder Cancer
ACH-000732	ZPR1	Head and Neck Cancer
ACH-000655	ZPR1	Brain Cancer
ACH-000901	CASQ2	Lung Cancer
ACH-002301	CCDC186	Lymphoma
ACH-000988	CCDC186	Endometrial/Uterine Cancer
ACH-000394	CCDC186	Lung Cancer
ACH-001229	CTD-2287O16.3	Head and Neck Cancer
ACH-001500	CTD-2287O16.3	Esophageal Cancer
ACH-000806	CTD-2287O16.3	Lymphoma
ACH-000963	CTD-2287O16.3	Colon/Colorectal Cancer
ACH-000846	CTD-2287O16.3	Head and Neck Cancer
ACH-000607	LVRN	Rhabdoid
ACH-002247	LVRN	Non-Cancerous
ACH-000787	DNAJC25-GNG10	Lung Cancer
ACH-000563	LRCH2	Lung Cancer
ACH-001539	LRCH2	Lymphoma
ACH-000947	LRCH2	Ovarian Cancer
ACH-001151	RBM19	Ovarian Cancer
ACH-000804	RP11-212D19.4	Neuroblastoma
ACH-000936	RP11-212D19.4	Ovarian Cancer
ACH-001333	CCDC191	Cervical Cancer
ACH-001300	CCDC191	Neuroblastoma
ACH-000025	USP28	Brain Cancer
ACH-000662	USP28	Lung Cancer
ACH-000503	ZGRF1	Head and Neck Cancer
ACH-000937	ZGRF1	Leukemia
ACH-000998	ZGRF1	Colon/Colorectal Cancer
ACH-000998	USF3	Colon/Colorectal Cancer
ACH-000158	USF3	Lymphoma
ACH-000969	CFAP44	Colon/Colorectal Cancer
ACH-001023	CFAP44	Thyroid Cancer
ACH-000990	CFAP44	Endometrial/Uterine Cancer
ACH-000784	PALM2	Esophageal Cancer
ACH-000750	RP11-162P23.2	Skin Cancer
ACH-000956	RP11-162P23.2	Prostate Cancer
ACH-000641	RP11-162P23.2	Leukemia
ACH-001061	RP11-162P23.2	Colon/Colorectal Cancer
ACH-000997	RP11-162P23.2	Colon/Colorectal Cancer
ACH-000952	TMIGD3	Prostate Cancer
ACH-000999	RP11-108O10.8	Colon/Colorectal Cancer
ACH-002222	RP11-108O10.8	Leukemia
ACH-002256	RP11-475E11.9	Leukemia
ACH-000993	SELPLG	Endometrial/Uterine Cancer
ACH-000757	KIAA1524	Lung Cancer
ACH-000946	COL4A5	Endometrial/Uterine Cancer
ACH-000990	CFAP58	Endometrial/Uterine Cancer
ACH-000717	CFAP58	Esophageal Cancer
ACH-002121	CFAP43	Skin Cancer
ACH-000993	CFAP43	Endometrial/Uterine Cancer
ACH-000977	CFAP43	Prostate Cancer
ACH-000978	CFAP43	Endometrial/Uterine Cancer
ACH-000901	CXorf57	Lung Cancer
ACH-000274	PACS2	Skin Cancer
ACH-000989	NEURL1	Colon/Colorectal Cancer
ACH-000274	PDCD11	Skin Cancer
ACH-000946	TXNRD1	Endometrial/Uterine Cancer
ACH-002294	BORCS7-ASMT	Head and Neck Cancer
ACH-002294	AS3MT	Head and Neck Cancer
ACH-000974	BORCS7	Endometrial/Uterine Cancer
ACH-000974	BORCS7-ASMT	Endometrial/Uterine Cancer
ACH-001599	GRIN3A	Lung Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000949	RP11-73M18.2	Gastric Cancer
ACH-000852	RP11-73M18.2	Lung Cancer
ACH-000757	STAB2	Lung Cancer
ACH-002508	PDGFD	Skin Cancer
ACH-000916	BIVM-ERCC5	Lung Cancer
ACH-000971	BIVM-ERCC5	Colon/Colorectal Cancer
ACH-000522	BIVM	Bladder Cancer
ACH-000517	TPP2	Pancreatic Cancer
ACH-000157	DYNC2H1	Lymphoma
ACH-000999	SLF2	Colon/Colorectal Cancer
ACH-001550	SLF2	Skin Cancer
ACH-000483	RP11-514P8.6	Liver Cancer
ACH-000483	RP11-577H5.5	Liver Cancer
ACH-000621	RP11-411B6.6	Breast Cancer
ACH-000941	RP11-411B6.6	Endometrial/Uterine Cancer
ACH-000225	RP11-514P8.6	Gastric Cancer
ACH-000757	CPN1	Lung Cancer
ACH-000662	COL15A1	Lung Cancer
ACH-001151	EMCN	Ovarian Cancer
ACH-001677	LINS1	Lymphoma
ACH-001328	GAS2L3	Skin Cancer
ACH-000901	SLC25A47	Lung Cancer
ACH-001075	RPL36A-HNRNPH2	Lung Cancer
ACH-000965	ADGRG7	Endometrial/Uterine Cancer
ACH-000403	ADGRG7	Colon/Colorectal Cancer
ACH-002221	ADGRG7	Lymphoma
ACH-000901	CMSS1	Lung Cancer
ACH-000556	C2orf15	Cervical Cancer
ACH-000556	RP11-111H13.1	Cervical Cancer
ACH-000988	PLPPR4	Endometrial/Uterine Cancer
ACH-000467	PLPPR4	Colon/Colorectal Cancer
ACH-002244	RP11-548K23.11	Head and Neck Cancer
ACH-000986	CYP3A7-CYP3A51P	Colon/Colorectal Cancer
ACH-001625	ERICH5	Head and Neck Cancer
ACH-002025	ERICH5	Colon/Colorectal Cancer
ACH-000999	PTCD1	Colon/Colorectal Cancer
ACH-000880	PTCD1	Gastric Cancer
ACH-001127	PTCD1	Leukemia
ACH-000249	ARHGAP19-SLIT1	Colon/Colorectal Cancer
ACH-000996	ARHGAP19-SLIT1	Endometrial/Uterine Cancer
ACH-000946	FAM169B	Endometrial/Uterine Cancer
ACH-000911	ARHGAP19-SLIT1	Gastric Cancer
ACH-000341	ARHGAP19-SLIT1	Neuroblastoma
ACH-000157	TRRAP	Lymphoma
ACH-001328	KLHL32	Skin Cancer
ACH-002171	MTERF3	Lung Cancer
ACH-000985	MTERF3	Colon/Colorectal Cancer
ACH-000913	CFAP54	Endometrial/Uterine Cancer
ACH-000538	CFAP54	Gastric Cancer
ACH-001081	CFAP54	Colon/Colorectal Cancer
ACH-002509	CFAP54	Skin Cancer
ACH-000967	CFAP54	Colon/Colorectal Cancer
ACH-002098	CFAP54	Skin Cancer
ACH-000434	CFAP54	Lung Cancer
ACH-000444	CFAP54	Lung Cancer
ACH-000991	CFAP54	Colon/Colorectal Cancer
ACH-002272	CFAP54	Lymphoma
ACH-000616	CFAP54	Gastric Cancer
ACH-000998	ACSM6	Colon/Colorectal Cancer
ACH-001363	ACSM6	Lung Cancer
ACH-000988	ACSM6	Endometrial/Uterine Cancer
ACH-000825	CFAP54	Lung Cancer
ACH-001151	CFAP54	Ovarian Cancer
ACH-001339	CYP2C9	Skin Cancer
ACH-000695	CDK17	Lung Cancer
ACH-000852	RP11-400G3.5	Lung Cancer
ACH-002282	RP11-400G3.5	Neuroblastoma
ACH-000993	PCSK1	Endometrial/Uterine Cancer
ACH-002510	TMEM56-RWDD3	Skin Cancer
ACH-002338	TMEM56-RWDD3	Skin Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-001364	CENPP	Lung Cancer
ACH-000977	CEP83	Prostate Cancer
ACH-000053	CEP83	Lymphoma
ACH-000770	CEP83	Leukemia
ACH-000912	IZUMO1R	Lung Cancer
ACH-001137	SLF1	Lung Cancer
ACH-001081	CEP295	Colon/Colorectal Cancer
ACH-000717	CEP295	Esophageal Cancer
ACH-000914	CEP295	Lymphoma
ACH-000993	CCDC67	Endometrial/Uterine Cancer
ACH-000980	VPS50	Lung Cancer
ACH-000714	VPS50	Myeloma
ACH-000963	VPS50	Colon/Colorectal Cancer
ACH-000982	VPS50	Colon/Colorectal Cancer
ACH-001345	VPS50	Colon/Colorectal Cancer
ACH-000901	LRRC69	Lung Cancer
ACH-001339	PEX1	Skin Cancer
ACH-000997	PPP4R3A	Colon/Colorectal Cancer
ACH-000967	PPP4R3A	Colon/Colorectal Cancer
ACH-001199	PPP4R3A	Colon/Colorectal Cancer
ACH-001719	MTERF1	Ovarian Cancer
ACH-000562	NGRN	Lung Cancer
ACH-000562	RP11-697E2.12	Lung Cancer
ACH-000989	NGRN	Colon/Colorectal Cancer
ACH-000989	RP11-697E2.12	Colon/Colorectal Cancer
ACH-002097	NGRN	Skin Cancer
ACH-002097	RP11-697E2.12	Skin Cancer
ACH-000848	RP11-697E2.6	Liver Cancer
ACH-000946	LIPM	Endometrial/Uterine Cancer
ACH-001610	ADGRV1	Brain Cancer
ACH-000950	ADGRV1	Colon/Colorectal Cancer
ACH-001119	ADGRV1	Lymphoma
ACH-000829	ADGRV1	Myeloma
ACH-000620	ADGRV1	Liver Cancer
ACH-000920	ADGRV1	Leukemia
ACH-000479	ADGRV1	Brain Cancer
ACH-002259	ADGRV1	Brain Cancer
ACH-001203	ADGRV1	Lymphoma
ACH-001509	ADGRV1	Head and Neck Cancer
ACH-000019	ADGRV1	Breast Cancer
ACH-000930	ADGRV1	Breast Cancer
ACH-000943	ADGRV1	Colon/Colorectal Cancer
ACH-002163	ADGRV1	Breast Cancer
ACH-000481	ADGRV1	Lung Cancer
ACH-001664	ADGRV1	Gastric Cancer
ACH-000991	ADGRV1	Colon/Colorectal Cancer
ACH-000924	ADGRV1	Lung Cancer
ACH-000891	ADGRV1	Lung Cancer
ACH-002040	ADGRV1	Skin Cancer
ACH-000989	ADGRV1	Colon/Colorectal Cancer
ACH-000123	CDK10	Ovarian Cancer
ACH-000999	WAPL	Colon/Colorectal Cancer
ACH-001610	WAPL	Brain Cancer
ACH-001751	WAPL	Sarcoma
ACH-000963	RP3-382I10.7	Colon/Colorectal Cancer
ACH-000999	CFAP206	Colon/Colorectal Cancer
ACH-000999	RP3-382I10.7	Colon/Colorectal Cancer
ACH-000843	CFAP206	Lung Cancer
ACH-000843	RP3-382I10.7	Lung Cancer
ACH-000458	AFF1	Skin Cancer
ACH-000684	RP5-1052I5.2	Kidney Cancer
ACH-000599	RP11-178L8.4	Pancreatic Cancer
ACH-000901	CD8B	Lung Cancer
ACH-001610	RNF103-CHMP3	Brain Cancer
ACH-000261	CA1	Lung Cancer
ACH-000960	DACH2	Leukemia
ACH-001521	CEP162	Skin Cancer
ACH-000431	CEP162	Lung Cancer
ACH-001306	CEP162	Thyroid Cancer
ACH-000855	SPATA31D3	Esophageal Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000481	GPAT3	Lung Cancer
ACH-000999	GPAT3	Colon/Colorectal Cancer
ACH-001339	TTLL7	Skin Cancer
ACH-000261	APOOL	Lung Cancer
ACH-000143	RP11-505K9.4	Lung Cancer
ACH-000025	DLG2	Brain Cancer
ACH-001081	RP11-152F13.10	Colon/Colorectal Cancer
ACH-000572	RP11-152F13.10	Skin Cancer
ACH-001203	RP11-152F13.10	Lymphoma
ACH-000650	ADGRL2	Skin Cancer
ACH-000550	ADGRL2	Skin Cancer
ACH-000470	ADGRL2	Colon/Colorectal Cancer
ACH-000733	ADGRL2	Lung Cancer
ACH-000971	ADGRL2	Colon/Colorectal Cancer
ACH-000437	ADGRL2	Brain Cancer
ACH-002164	ADGRL2	Pancreatic Cancer
ACH-000886	ADGRL2	Lung Cancer
ACH-000898	PPFIA2	Gastric Cancer
ACH-000901	CACNA2D1	Lung Cancer
ACH-002111	BCO1	Bone Cancer
ACH-000961	BCO1	Endometrial/Uterine Cancer
ACH-000947	CEMIP	Ovarian Cancer
ACH-000061	CEMIP	Lymphoma
ACH-001850	CEMIP	Gallbladder Cancer
ACH-000846	RP11-26J3.4	Head and Neck Cancer
ACH-000398	FASN	Lymphoma
ACH-001063	DHFR	Ovarian Cancer
ACH-000624	FAAP100	Breast Cancer
ACH-000886	ADGRL4	Lung Cancer
ACH-000383	ADGRL4	Esophageal Cancer
ACH-001529	BAHCC1	Endometrial/Uterine Cancer
ACH-000820	BAHCC1	Colon/Colorectal Cancer
ACH-001127	BAHCC1	Leukemia
ACH-000922	BAHCC1	Leukemia
ACH-002089	BAHCC1	Lung Cancer
ACH-000956	BAHCC1	Prostate Cancer
ACH-000963	ADGRL4	Colon/Colorectal Cancer
ACH-002017	ADGRL4	Eye Cancer
ACH-000744	ADGRL4	Lung Cancer
ACH-000028	ADGRL4	Breast Cancer
ACH-000019	ADGRL4	Breast Cancer
ACH-000806	CEP131	Lymphoma
ACH-000784	CEP131	Esophageal Cancer
ACH-001862	CEP131	Bile Duct Cancer
ACH-000943	CEP131	Colon/Colorectal Cancer
ACH-000973	CEP131	Bladder Cancer
ACH-002296	C14orf178	Head and Neck Cancer
ACH-000960	POMT2	Leukemia
ACH-000901	ROBO2	Lung Cancer
ACH-000799	CARNMT1	Skin Cancer
ACH-000695	TSPAN3	Lung Cancer
ACH-000963	LRRC74A	Colon/Colorectal Cancer
ACH-001566	LRRC74A	Skin Cancer
ACH-001356	LRRC74A	Thyroid Cancer
ACH-000261	ANGEL1	Lung Cancer
ACH-000888	FAM47E-STBD1	Lung Cancer
ACH-001328	WDR41	Skin Cancer
ACH-000932	CNTNAP4	Gastric Cancer
ACH-002098	TMEM266	Skin Cancer
ACH-001523	ADK	Skin Cancer
ACH-001321	KARS	Thyroid Cancer
ACH-000973	RP11-77K12.7	Bladder Cancer
ACH-000938	RP11-77K12.7	Leukemia
ACH-000999	RP11-574K11.31	Colon/Colorectal Cancer
ACH-000981	RP11-574K11.31	Leukemia
ACH-000014	ZSWIM8	Skin Cancer
ACH-000995	RP11-77K12.1	Leukemia
ACH-000993	RP11-77K12.1	Endometrial/Uterine Cancer
ACH-000334	RP11-77K12.1	Lymphoma
ACH-000662	HIP1	Lung Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000960	HIP1	Leukemia
ACH-000974	ERICH3	Endometrial/Uterine Cancer
ACH-000948	ERICH3	Gastric Cancer
ACH-000991	CFAP70	Colon/Colorectal Cancer
ACH-000692	ERICH3	Head and Neck Cancer
ACH-001402	CFAP70	Cervical Cancer
ACH-001334	CFAP70	Cervical Cancer
ACH-000957	CFAP70	Colon/Colorectal Cancer
ACH-002133	ERICH3	Lung Cancer
ACH-000860	CFAP70	Lung Cancer
ACH-001024	CFAP70	Skin Cancer
ACH-000871	TNNI3K	Lung Cancer
ACH-000999	TNNI3K	Colon/Colorectal Cancer
ACH-000982	TNNI3K	Colon/Colorectal Cancer
ACH-001345	TNNI3K	Colon/Colorectal Cancer
ACH-000569	TNNI3K	Lung Cancer
ACH-000399	TNNI3K	Lung Cancer
ACH-002336	TNNI3K	Unknown
ACH-000988	TNNI3K	Endometrial/Uterine Cancer
ACH-000157	LOXL3	Lymphoma
ACH-000963	RP11-463D19.2	Colon/Colorectal Cancer
ACH-000981	RP11-463D19.2	Leukemia
ACH-001845	TNNI3K	Bile Duct Cancer
ACH-000014	MLKL	Skin Cancer
ACH-001414	TNNI3K	Bladder Cancer
ACH-001751	BBOF1	Sarcoma
ACH-000662	GLG1	Lung Cancer
ACH-000186	SLC4A5	Lung Cancer
ACH-000996	SLC4A5	Endometrial/Uterine Cancer
ACH-000879	RP11-287D1.3	Endometrial/Uterine Cancer
ACH-000980	RP11-287D1.3	Lung Cancer
ACH-000988	RP11-287D1.3	Endometrial/Uterine Cancer
ACH-000537	NPIPB15	Liver Cancer
ACH-001390	RP5-1021I20.4	Breast Cancer
ACH-000901	ABCB7	Lung Cancer
ACH-000995	TEN1-CDK3	Leukemia
ACH-001061	TMEM94	Colon/Colorectal Cancer
ACH-000997	TMEM94	Colon/Colorectal Cancer
ACH-000986	TMEM94	Colon/Colorectal Cancer
ACH-002256	TMEM94	Leukemia
ACH-000992	TMEM94	Head and Neck Cancer
ACH-000928	TMEM94	Endometrial/Uterine Cancer
ACH-000696	TMEM94	Ovarian Cancer
ACH-000548	TMEM94	Head and Neck Cancer
ACH-000988	TMEM94	Endometrial/Uterine Cancer
ACH-002222	PPP4R2	Leukemia
ACH-000697	RP11-106M3.2	Lymphoma
ACH-002146	RP11-106M3.2	Ovarian Cancer
ACH-000192	RP11-106M3.2	Endometrial/Uterine Cancer
ACH-000695	RP11-106M3.5	Lung Cancer
ACH-000278	RP11-106M3.5	Ovarian Cancer
ACH-000974	RP11-106M3.5	Endometrial/Uterine Cancer
ACH-000695	CELF6	Lung Cancer
ACH-000695	RP11-106M3.2	Lung Cancer
ACH-000278	RP11-106M3.2	Ovarian Cancer
ACH-000974	RP11-106M3.2	Endometrial/Uterine Cancer
ACH-001328	TTYH2	Skin Cancer
ACH-000914	RP11-293I14.2	Lymphoma
ACH-000662	DYSF	Lung Cancer
ACH-002174	TEX261	Lung Cancer
ACH-000966	TEX261	Ovarian Cancer
ACH-000559	SYNJ2BP-COX16	Lung Cancer
ACH-001339	BDP1	Skin Cancer
ACH-000901	KLHL1	Lung Cancer
ACH-000986	UGT2A2	Colon/Colorectal Cancer
ACH-000375	UGT2A2	Kidney Cancer
ACH-000897	UGT2A2	Thyroid Cancer
ACH-000956	RP11-529K1.3	Prostate Cancer
ACH-000955	RP11-529K1.3	Colon/Colorectal Cancer
ACH-000907	ADGRB3	Kidney Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000998	ADGRB3	Colon/Colorectal Cancer
ACH-000901	ADGRB3	Lung Cancer
ACH-000508	ADGRB3	Lung Cancer
ACH-002309	ADGRB3	Sarcoma
ACH-002387	ADGRB3	Skin Cancer
ACH-001536	ADGRB3	Bile Duct Cancer
ACH-000860	ADGRB3	Lung Cancer
ACH-000050	ADGRB3	Myeloma
ACH-001001	ADGRB3	Bone Cancer
ACH-000729	ADGRB3	Lung Cancer
ACH-000988	RP11-343C2.9	Endometrial/Uterine Cancer
ACH-000988	RP11-343C2.7	Endometrial/Uterine Cancer
ACH-000979	RP11-343C2.11	Prostate Cancer
ACH-002225	RP11-315D16.2	Brain Cancer
ACH-000279	RP11-315D16.2	Bone Cancer
ACH-000555	RP11-474G23.1	Kidney Cancer
ACH-000157	RP11-474G23.1	Lymphoma
ACH-000838	RP11-474G23.1	Myeloma
ACH-000535	RP11-474G23.1	Pancreatic Cancer
ACH-000580	RP11-474G23.1	Skin Cancer
ACH-000120	RP11-474G23.1	Neuroblastoma
ACH-000805	RP11-474G23.1	Skin Cancer
ACH-000786	RP11-474G23.1	Lymphoma
ACH-000056	RP11-474G23.1	Lymphoma
ACH-000198	RP11-474G23.1	Leukemia
ACH-000487	RP11-474G23.1	Leukemia
ACH-000027	RP11-474G23.1	Brain Cancer
ACH-000111	RP11-474G23.1	Breast Cancer
ACH-000725	RP11-474G23.1	Breast Cancer
ACH-000946	RP11-474G23.1	Endometrial/Uterine Cancer
ACH-000004	RP11-474G23.1	Leukemia
ACH-000069	RP11-474G23.1	Lymphoma
ACH-000799	RP11-474G23.1	Skin Cancer
ACH-000118	RP11-474G23.1	Pancreatic Cancer
ACH-000650	RP11-474G23.1	Skin Cancer
ACH-000476	RP11-474G23.1	Liver Cancer
ACH-000166	RP11-474G23.1	Leukemia
ACH-000501	RP11-474G23.1	Colon/Colorectal Cancer
ACH-000032	RP11-474G23.1	Leukemia
ACH-000156	RP11-474G23.1	Leukemia
ACH-000391	RP11-474G23.1	Bone Cancer
ACH-000077	RP11-474G23.1	Lymphoma
ACH-000462	RP11-474G23.1	Leukemia
ACH-000130	RP11-474G23.1	Leukemia
ACH-000129	RP11-474G23.1	Lung Cancer
ACH-000845	RP11-474G23.1	Lung Cancer
ACH-000766	RP11-474G23.1	Lung Cancer
ACH-000781	RP11-474G23.1	Lung Cancer
ACH-000481	RP11-474G23.1	Lung Cancer
ACH-000628	RP11-474G23.1	Lung Cancer
ACH-000491	RP11-474G23.1	Colon/Colorectal Cancer
ACH-000297	RP11-474G23.1	Lung Cancer
ACH-000050	RP11-474G23.1	Myeloma
ACH-000336	RP11-474G23.1	Leukemia
ACH-001182	RP11-474G23.1	Unknown
ACH-000441	RP11-474G23.1	Skin Cancer
ACH-000437	RP11-474G23.1	Brain Cancer
ACH-000146	RP11-474G23.1	Leukemia
ACH-002166	RP11-474G23.1	Skin Cancer
ACH-000820	RP11-474G23.1	Colon/Colorectal Cancer
ACH-000398	RANBP10	Lymphoma
ACH-000785	C8orf44-SGK3	Lung Cancer
ACH-001203	C8orf44-SGK3	Lymphoma
ACH-000417	AP003419.11	Pancreatic Cancer
ACH-000999	CKLF	Colon/Colorectal Cancer
ACH-000754	RP11-745O10.4	Lymphoma
ACH-000963	RBM14-RBM4	Colon/Colorectal Cancer
ACH-000694	RBM14-RBM4	Esophageal Cancer
ACH-000999	BBS1	Colon/Colorectal Cancer
ACH-000960	EPHA5	Leukemia

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-001328	BPTF	Skin Cancer
ACH-000623	AC068533.7	Brain Cancer
ACH-000999	AC068533.7	Colon/Colorectal Cancer
ACH-000830	SIPA1	Lung Cancer
ACH-000996	CHURC1-FNTB	Endometrial/Uterine Cancer
ACH-001551	SIPA1	Skin Cancer
ACH-001402	SIPA1	Cervical Cancer
ACH-000863	SIPA1	Brain Cancer
ACH-000994	SIPA1	Endometrial/Uterine Cancer
ACH-000981	SIPA1	Leukemia
ACH-002166	SIPA1	Skin Cancer
ACH-001345	SIPA1	Colon/Colorectal Cancer
ACH-000992	SIPA1	Head and Neck Cancer
ACH-001852	SIPA1	Bile Duct Cancer
ACH-000071	AC069368.3	Fibroblast
ACH-000398	PPWD1	Lymphoma
ACH-002256	ARL2-SNX15	Leukemia
ACH-002256	SNX15	Leukemia
ACH-002256	CTD-2116N17.1	Leukemia
ACH-001151	KCNH5	Ovarian Cancer
ACH-000891	ADGRL3	Lung Cancer
ACH-001401	ADGRL3	Adrenal Cancer
ACH-002283	ADGRL3	Neuroblastoma
ACH-000283	ADGRL3	Brain Cancer
ACH-000893	LKAAEAR1	Lung Cancer
ACH-001517	ZNF512B	Endometrial/Uterine Cancer
ACH-000954	ZNF512B	Endometrial/Uterine Cancer
ACH-000941	ZNF512B	Endometrial/Uterine Cancer
ACH-001638	ZNF512B	Lymphoma
ACH-000350	HNRNPUL2-BSCL2	Colon/Colorectal Cancer
ACH-000972	HNRNPUL2-BSCL2	Endometrial/Uterine Cancer
ACH-001685	HNRNPUL2-BSCL2	Lymphoma
ACH-001610	LIME1	Brain Cancer
ACH-001233	ZGPAT	Lung Cancer
ACH-000328	ZGPAT	Brain Cancer
ACH-002004	EEF1G	Skin Cancer
ACH-002004	RP11-864I4.1	Skin Cancer
ACH-000941	RTEL1-TNFRSF6B	Endometrial/Uterine Cancer
ACH-000808	RTEL1-TNFRSF6B	Bile Duct Cancer
ACH-002166	RTEL1-TNFRSF6B	Skin Cancer
ACH-000885	RTEL1-TNFRSF6B	Ovarian Cancer
ACH-000989	RTEL1-TNFRSF6B	Colon/Colorectal Cancer
ACH-002297	RTEL1-TNFRSF6B	Head and Neck Cancer
ACH-001413	RTEL1-TNFRSF6B	Bladder Cancer
ACH-001529	RTEL1-TNFRSF6B	Endometrial/Uterine Cancer
ACH-000353	RTEL1-TNFRSF6B	Esophageal Cancer
ACH-000236	RTEL1-TNFRSF6B	Colon/Colorectal Cancer
ACH-000914	RTEL1-TNFRSF6B	Lymphoma
ACH-000981	RTEL1-TNFRSF6B	Leukemia
ACH-000681	RP11-310K10.1	Lung Cancer
ACH-000123	COL20A1	Ovarian Cancer
ACH-000988	RP11-51F16.8	Endometrial/Uterine Cancer
ACH-000945	RP11-51F16.8	Lung Cancer
ACH-000468	FADS3	Pancreatic Cancer
ACH-000930	CTD-2501B8.1	Breast Cancer
ACH-002026	CTD-2501B8.1	Endometrial/Uterine Cancer
ACH-000974	CTD-2501B8.1	Endometrial/Uterine Cancer
ACH-000982	CTD-2501B8.1	Colon/Colorectal Cancer
ACH-000207	CTD-2501B8.1	Head and Neck Cancer
ACH-000782	ICE2	Leukemia
ACH-001735	ICE2	Leukemia
ACH-000990	ICE2	Endometrial/Uterine Cancer
ACH-001134	ICE2	Lymphoma
ACH-000274	SS18L1	Skin Cancer
ACH-000840	PCNXL4	Lung Cancer
ACH-000025	STX3	Brain Cancer
ACH-000997	CTD-3138B18.4	Colon/Colorectal Cancer
ACH-000997	AC010642.1	Colon/Colorectal Cancer
ACH-000999	ZFP91-CNTF	Colon/Colorectal Cancer
ACH-000269	RP11-80H18.3	Brain Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000979	CFAP20	Prostate Cancer
ACH-000929	AC003005.4	Lung Cancer
ACH-002175	AC003005.4	Lung Cancer
ACH-002238	GCOM1	Leukemia
ACH-001650	GCOM1	Cervical Cancer
ACH-002215	AC003002.6	Lymphoma
ACH-000461	DCTN2	Bile Duct Cancer
ACH-000999	AC003002.6	Colon/Colorectal Cancer
ACH-000999	AC004076.7	Colon/Colorectal Cancer
ACH-000009	AC003002.4	Colon/Colorectal Cancer
ACH-000763	AC003002.4	Myeloma
ACH-000763	ZNF547	Myeloma
ACH-000997	TRAPPC2P1	Colon/Colorectal Cancer
ACH-000997	AC003002.4	Colon/Colorectal Cancer
ACH-000409	DRC7	Ovarian Cancer
ACH-000988	DRC7	Endometrial/Uterine Cancer
ACH-000942	ADGRG1	Leukemia
ACH-000951	ADGRG1	Lung Cancer
ACH-000963	ADGRG1	Colon/Colorectal Cancer
ACH-000992	RP11-123K3.4	Head and Neck Cancer
ACH-000887	TMX2-CTNND1	Brain Cancer
ACH-000757	STAT6	Lung Cancer
ACH-000992	TMX2-CTNND1	Head and Neck Cancer
ACH-000960	MYO1A	Leukemia
ACH-001494	ZIM2	Bile Duct Cancer
ACH-001030	CTD-2510F5.6	Bone Cancer
ACH-000992	CTD-2510F5.6	Head and Neck Cancer
ACH-000786	CTD-2510F5.6	Lymphoma
ACH-000995	CTD-2510F5.6	Leukemia
ACH-000825	CTD-2510F5.6	Lung Cancer
ACH-001151	CTD-2510F5.6	Ovarian Cancer
ACH-001151	SMG8	Ovarian Cancer
ACH-001490	CTD-2510F5.6	Lung Cancer
ACH-000990	STX16-NPEPL1	Endometrial/Uterine Cancer
ACH-001151	PRR11	Ovarian Cancer
ACH-000258	PLPP3	Breast Cancer
ACH-000941	PLPP3	Endometrial/Uterine Cancer
ACH-000662	STAT2	Lung Cancer
ACH-000620	RP11-977G19.10	Liver Cancer
ACH-001339	ANKRD52	Skin Cancer
ACH-001738	RP11-603J24.9	Leukemia
ACH-000744	SARNP	Lung Cancer
ACH-000977	RP11-762I7.5	Prostate Cancer
ACH-000992	RP11-644F5.10	Head and Neck Cancer
ACH-001848	BLOC1S1	Bile Duct Cancer
ACH-000025	KTN1	Brain Cancer
ACH-002509	CTD-2105E13.6	Skin Cancer
ACH-000929	PPP4R3B	Lung Cancer
ACH-001061	PPP4R3B	Colon/Colorectal Cancer
ACH-000997	PPP4R3B	Colon/Colorectal Cancer
ACH-002217	PPP4R3B	Neuroblastoma
ACH-000937	PPP4R3B	Leukemia
ACH-000930	PPP4R3B	Breast Cancer
ACH-000974	PPP4R3B	Endometrial/Uterine Cancer
ACH-000157	CTD-2587H24.4	Lymphoma
ACH-000274	CTD-2587H24.4	Skin Cancer
ACH-000157	DNAAF3	Lymphoma
ACH-001329	DNAAF3	Brain Cancer
ACH-001278	DNAAF3	Ovarian Cancer
ACH-001454	DNAAF3	Colon/Colorectal Cancer
ACH-001456	DNAAF3	Colon/Colorectal Cancer
ACH-000206	DNAAF3	Lymphoma
ACH-001433	DNAAF3	Sarcoma
ACH-001041	DNAAF3	Lung Cancer
ACH-001053	DNAAF3	Brain Cancer
ACH-001054	DNAAF3	Brain Cancer
ACH-001496	DNAAF3	Esophageal Cancer
ACH-002349	DNAAF3	Unknown
ACH-000941	DNAAF3	Endometrial/Uterine Cancer
ACH-002026	DNAAF3	Endometrial/Uterine Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000274	DNAAF3	Skin Cancer
ACH-001737	DNAAF3	Leukemia
ACH-002044	DNAAF3	Head and Neck Cancer
ACH-001834	DNAAF3	Bile Duct Cancer
ACH-001836	DNAAF3	Bile Duct Cancer
ACH-001530	DNAAF3	Endometrial/Uterine Cancer
ACH-001539	DNAAF3	Lymphoma
ACH-000128	DNAAF3	Brain Cancer
ACH-001550	DNAAF3	Skin Cancer
ACH-001554	DNAAF3	Eye Cancer
ACH-001559	DNAAF3	Lung Cancer
ACH-001569	DNAAF3	Skin Cancer
ACH-001570	DNAAF3	Skin Cancer
ACH-000514	DNAAF3	Lung Cancer
ACH-000434	DNAAF3	Lung Cancer
ACH-000841	DNAAF3	Lung Cancer
ACH-000929	DNAAF3	Lung Cancer
ACH-000251	DNAAF3	Lung Cancer
ACH-001366	DNAAF3	Neuroblastoma
ACH-001368	DNAAF3	Esophageal Cancer
ACH-001719	DNAAF3	Ovarian Cancer
ACH-000247	DNAAF3	Gastric Cancer
ACH-001624	DNAAF3	Brain Cancer
ACH-000779	DNAAF3	Lung Cancer
ACH-001711	DNAAF3	Brain Cancer
ACH-001173	DNAAF3	Brain Cancer
ACH-001386	DNAAF3	Lung Cancer
ACH-000246	DNAAF3	Kidney Cancer
ACH-000172	DNAAF3	Rhabdoid
ACH-001685	DNAAF3	Lymphoma
ACH-002509	DNAAF3	Skin Cancer
ACH-002022	TNNI3	Colon/Colorectal Cancer
ACH-000295	TNNI3	Leukemia
ACH-000901	NCR1	Lung Cancer
ACH-000912	LEXM	Lung Cancer
ACH-000535	LEXM	Pancreatic Cancer
ACH-000954	LEXM	Endometrial/Uterine Cancer
ACH-000314	MROH7-TTC4	Lung Cancer
ACH-000851	MROH7-TTC4	Lung Cancer
ACH-000991	RP11-231C18.3	Colon/Colorectal Cancer
ACH-000999	RP11-231C18.3	Colon/Colorectal Cancer
ACH-000835	RP11-231C18.3	Sarcoma
ACH-000929	MROH7-TTC4	Lung Cancer
ACH-001355	MROH7-TTC4	Neuroblastoma
ACH-000551	RP11-231C18.3	Leukemia
ACH-000996	RP11-231C18.3	Endometrial/Uterine Cancer
ACH-000845	RP11-231C18.3	Lung Cancer
ACH-001819	RP11-231C18.3	Breast Cancer
ACH-000157	PDE1B	Lymphoma
ACH-000848	PLPP1	Liver Cancer
ACH-000025	CNOT3	Brain Cancer
ACH-000683	RP11-446E24.4	Colon/Colorectal Cancer
ACH-000989	RP11-231C18.3	Colon/Colorectal Cancer
ACH-000763	RP11-231C18.3	Myeloma
ACH-000871	RP11-793H13.10	Lung Cancer
ACH-000282	ASB3	Lung Cancer
ACH-000343	RP11-793H13.10	Lung Cancer
ACH-000625	RP11-793H13.10	Liver Cancer
ACH-000575	RP11-793H13.10	Lung Cancer
ACH-000962	RP11-793H13.10	Ovarian Cancer
ACH-000863	TNS2	Brain Cancer
ACH-000940	TNS2	Endometrial/Uterine Cancer
ACH-000998	TNS2	Colon/Colorectal Cancer
ACH-000997	TNS2	Colon/Colorectal Cancer
ACH-000941	TNS2	Endometrial/Uterine Cancer
ACH-001991	TNS2	Ovarian Cancer
ACH-001664	ERO1A	Gastric Cancer
ACH-000800	ERO1A	Lung Cancer
ACH-001517	ERO1A	Endometrial/Uterine Cancer
ACH-000954	ERO1A	Endometrial/Uterine Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000941	ERO1A	Endometrial/Uterine Cancer
ACH-000995	RP5-966M1.6	Leukemia
ACH-000123	MYO5C	Ovarian Cancer
ACH-000307	RP11-330H6.5	Pancreatic Cancer
ACH-000307	TWF2	Pancreatic Cancer
ACH-002222	AC018755.18	Leukemia
ACH-000996	AC018755.18	Endometrial/Uterine Cancer
ACH-000314	AC018755.18	Lung Cancer
ACH-000961	ABHD14A-ACY1	Endometrial/Uterine Cancer
ACH-000925	ABHD14A-ACY1	Lung Cancer
ACH-002077	CTD-2616J11.11	Lung Cancer
ACH-000025	GLDN	Brain Cancer
ACH-000768	CTC-518B2.8	Breast Cancer
ACH-000952	CTD-2545M3.6	Prostate Cancer
ACH-000407	FIGNL1	Endometrial/Uterine Cancer
ACH-000459	FIGNL1	Kidney Cancer
ACH-000157	SMARCD1	Lymphoma
ACH-000997	CTC-326K19.6	Colon/Colorectal Cancer
ACH-001495	AL627171.1	Endometrial/Uterine Cancer
ACH-000928	ATP8B4	Endometrial/Uterine Cancer
ACH-000278	PRR12	Ovarian Cancer
ACH-000999	CTD-3148I10.9	Colon/Colorectal Cancer
ACH-001362	CTD-3148I10.9	Lung Cancer
ACH-000589	CTD-3148I10.9	Lung Cancer
ACH-000972	CTD-3148I10.9	Endometrial/Uterine Cancer
ACH-001321	GYS1	Thyroid Cancer
ACH-001339	COPS2	Skin Cancer
ACH-000928	CACNA1F	Endometrial/Uterine Cancer
ACH-000985	CYTH2	Colon/Colorectal Cancer
ACH-001364	RP11-310N16.1	Lung Cancer
ACH-000683	RP11-310N16.1	Colon/Colorectal Cancer
ACH-000274	RP11-310N16.1	Skin Cancer
ACH-000274	LHCGR	Skin Cancer
ACH-000937	STON1-GTF2A1L	Leukemia
ACH-000972	GTF2A1L	Endometrial/Uterine Cancer
ACH-000384	STON1-GTF2A1L	Bladder Cancer
ACH-000243	RP11-729L2.2	Pancreatic Cancer
ACH-000992	RP11-729L2.2	Head and Neck Cancer
ACH-000458	PFKM	Skin Cancer
ACH-000274	LONP2	Skin Cancer
ACH-000877	CFAP53	Lymphoma
ACH-001064	CFAP53	Lymphoma
ACH-000987	ADGRF4	Skin Cancer
ACH-000458	CORIN	Skin Cancer
ACH-000807	RP11-761B3.1	Brain Cancer
ACH-000506	RP11-761B3.1	Lung Cancer
ACH-000998	RP11-886H22.1	Colon/Colorectal Cancer
ACH-000025	PACSIN3	Brain Cancer
ACH-000924	RPL17-C18orf32	Lung Cancer
ACH-001374	ADGRF1	Ovarian Cancer
ACH-000746	ADGRF1	Gastric Cancer
ACH-002098	ABC7-42404400C24.1	Skin Cancer
ACH-002302	JADE3	Leukemia
ACH-001484	PRSS50	Leukemia
ACH-001485	PRSS50	Leukemia
ACH-000274	ALS2CL	Skin Cancer
ACH-001151	CPB2	Ovarian Cancer
ACH-001339	GABRA2	Skin Cancer
ACH-000958	BHMG1	Colon/Colorectal Cancer
ACH-001852	ERICH6B	Bile Duct Cancer
ACH-000288	ERICH6B	Breast Cancer
ACH-000839	ERICH6B	Bladder Cancer
ACH-000183	ERICH6B	Myeloma
ACH-001518	RP11-96O20.4	Endometrial/Uterine Cancer
ACH-001061	RP11-96O20.4	Colon/Colorectal Cancer
ACH-000997	RP11-96O20.4	Colon/Colorectal Cancer
ACH-000379	CH507-9B2.1	Lung Cancer
ACH-000963	CH507-9B2.1	Colon/Colorectal Cancer
ACH-000912	CH507-9B2.4	Lung Cancer
ACH-000912	CH507-9B2.3	Lung Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-001517	CH507-9B2.5	Endometrial/Uterine Cancer
ACH-000941	CH507-9B2.5	Endometrial/Uterine Cancer
ACH-000988	CH507-9B2.5	Endometrial/Uterine Cancer
ACH-002152	RP11-290H9.2	Skin Cancer
ACH-000959	RP11-290H9.2	Colon/Colorectal Cancer
ACH-000416	CH507-42P11.8	Lung Cancer
ACH-000817	U2AF1L5	Myeloma
ACH-001577	U2AF1L5	Leukemia
ACH-000993	CBSL	Endometrial/Uterine Cancer
ACH-000978	SPATS1	Endometrial/Uterine Cancer
ACH-001422	SPATS1	Prostate Cancer
ACH-000274	LOXHD1	Skin Cancer
ACH-000903	L34079.2	Thyroid Cancer
ACH-000989	L34079.2	Colon/Colorectal Cancer
ACH-000942	RP11-296A16.1	Leukemia
ACH-000974	RP11-296A16.1	Endometrial/Uterine Cancer
ACH-001127	RP11-296A16.1	Leukemia
ACH-002344	RP11-296A16.1	Unknown
ACH-002090	RP11-296A16.1	Kidney Cancer
ACH-001339	PHLDB3	Skin Cancer
ACH-000993	THADA	Endometrial/Uterine Cancer
ACH-000995	CFAP57	Leukemia
ACH-000873	CFAP57	Esophageal Cancer
ACH-000157	ZNF318	Lymphoma
ACH-001162	P3H1	Myeloma
ACH-000998	P3H1	Colon/Colorectal Cancer
ACH-000104	P3H1	Leukemia
ACH-001203	P3H1	Lymphoma
ACH-000784	NIM1K	Esophageal Cancer
ACH-000351	NIM1K	Gastric Cancer
ACH-000159	NIM1K	Kidney Cancer
ACH-000966	LINC01620	Ovarian Cancer
ACH-000855	LINC01620	Esophageal Cancer
ACH-001417	RP11-111K18.1	Lung Cancer
ACH-000609	KRBOX1	Brain Cancer
ACH-000979	KRBOX1	Prostate Cancer
ACH-000889	RP4-613B23.5	Myeloma
ACH-000295	AC006486.9	Leukemia
ACH-002240	AC006486.9	Lymphoma
ACH-002271	AC006486.9	Sarcoma
ACH-000416	AC006486.9	Lung Cancer
ACH-000795	AC006486.9	Leukemia
ACH-000981	AC006486.9	Leukemia
ACH-001321	TRERF1	Thyroid Cancer
ACH-000795	AC011513.3	Leukemia
ACH-000605	AC011513.3	Esophageal Cancer
ACH-002237	RP1-138B7.6	Bile Duct Cancer
ACH-001858	RP1-138B7.6	Bile Duct Cancer
ACH-002309	RP1-138B7.6	Sarcoma
ACH-000197	RP1-138B7.6	Leukemia
ACH-001655	JMJD7	Endometrial/Uterine Cancer
ACH-000583	GPAT4	Lymphoma
ACH-001333	RAB4B-EGLN2	Cervical Cancer
ACH-000695	BRCA1	Lung Cancer
ACH-000478	PTGES3L-AARSD1	Liver Cancer
ACH-000157	SHKBP1	Lymphoma
ACH-000993	C7	Endometrial/Uterine Cancer
ACH-000946	C7	Endometrial/Uterine Cancer
ACH-000794	SUGCT	Head and Neck Cancer
ACH-000025	STAT5B	Brain Cancer
ACH-001345	SUGCT	Colon/Colorectal Cancer
ACH-001190	SUGCT	Skin Cancer
ACH-001137	SUGCT	Lung Cancer
ACH-000950	CTAGE5	Colon/Colorectal Cancer
ACH-000948	CTAGE5	Gastric Cancer
ACH-000981	CTAGE5	Leukemia
ACH-001190	RP11-407N17.3	Skin Cancer
ACH-000662	RP11-407N17.3	Lung Cancer
ACH-000662	MIA2	Lung Cancer
ACH-000973	CTC-360G5.8	Bladder Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-001634	SARS2	Leukemia
ACH-000978	AC104534.3	Endometrial/Uterine Cancer
ACH-000356	AC104534.3	Gastric Cancer
ACH-001539	AC104534.3	Lymphoma
ACH-000274	WDR19	Skin Cancer
ACH-001239	CATSPERG	Skin Cancer
ACH-000695	ATL2	Lung Cancer
ACH-000998	PLPP5	Colon/Colorectal Cancer
ACH-000948	PLPP5	Gastric Cancer
ACH-000960	PLPP5	Leukemia
ACH-000473	PLPP5	Bladder Cancer
ACH-001339	SIM2	Skin Cancer
ACH-000475	RP5-972B16.2	Liver Cancer
ACH-000800	RP5-972B16.2	Lung Cancer
ACH-000997	RP5-972B16.2	Colon/Colorectal Cancer
ACH-000937	RP5-972B16.2	Leukemia
ACH-000993	HYPM	Endometrial/Uterine Cancer
ACH-000998	RP11-613M10.9	Colon/Colorectal Cancer
ACH-000634	RP11-613M10.9	Brain Cancer
ACH-000912	ADGRA2	Lung Cancer
ACH-000985	ADGRA2	Colon/Colorectal Cancer
ACH-000567	ADGRA2	Lymphoma
ACH-002511	ADGRA2	Skin Cancer
ACH-000992	ADGRA2	Head and Neck Cancer
ACH-000998	RP5-972B16.2	Colon/Colorectal Cancer
ACH-002278	RP5-972B16.2	Neuroblastoma
ACH-001363	RP5-972B16.2	Lung Cancer
ACH-001786	RP5-972B16.2	Colon/Colorectal Cancer
ACH-000875	RP5-972B16.2	Lung Cancer
ACH-000979	RP5-972B16.2	Prostate Cancer
ACH-001339	ANKRD30A	Skin Cancer
ACH-001861	CEBPZOS	Gallbladder Cancer
ACH-000971	NWD2	Colon/Colorectal Cancer
ACH-001061	RP5-972B16.2	Colon/Colorectal Cancer
ACH-000157	ARHGAP40	Lymphoma
ACH-000990	CCDC169-SOHLH2	Endometrial/Uterine Cancer
ACH-000999	CCDC169-SOHLH2	Colon/Colorectal Cancer
ACH-000925	TBC1D3K	Lung Cancer
ACH-002238	PROSER3	Leukemia
ACH-001650	PROSER3	Cervical Cancer
ACH-000570	PROSER3	Brain Cancer
ACH-000395	AC002398.9	Lung Cancer
ACH-000991	AD000671.6	Colon/Colorectal Cancer
ACH-000521	CFAP47	Lung Cancer
ACH-000624	CFAP47	Breast Cancer
ACH-000924	CFAP47	Lung Cancer
ACH-000550	CREB3	Skin Cancer
ACH-002256	RP11-561B11.2	Leukemia
ACH-000028	RP11-244H3.4	Breast Cancer
ACH-000019	RP11-244H3.4	Breast Cancer
ACH-000993	LHX1	Endometrial/Uterine Cancer
ACH-000990	PHF24	Endometrial/Uterine Cancer
ACH-000157	PARD3	Lymphoma
ACH-000963	AP000295.9	Colon/Colorectal Cancer
ACH-000621	RP11-195F19.29	Breast Cancer
ACH-000971	AP000295.9	Colon/Colorectal Cancer
ACH-002020	RPS10-NUDT3	Cervical Cancer
ACH-000657	CCL15-CCL14	Ovarian Cancer
ACH-000941	CCL14	Endometrial/Uterine Cancer
ACH-000941	CCL15-CCL14	Endometrial/Uterine Cancer
ACH-000852	CCL15-CCL14	Lung Cancer
ACH-001517	RPS10-NUDT3	Endometrial/Uterine Cancer
ACH-000990	HEATR9	Endometrial/Uterine Cancer
ACH-000865	HEATR9	Esophageal Cancer
ACH-000705	HEATR9	Lung Cancer
ACH-001113	HEATR9	Lung Cancer
ACH-000159	HEATR9	Kidney Cancer
ACH-000652	HEATR9	Pancreatic Cancer
ACH-000954	HEATR9	Endometrial/Uterine Cancer
ACH-001053	HEATR9	Brain Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-001054	HEATR9	Brain Cancer
ACH-000980	C1QTNF3-AMACR	Lung Cancer
ACH-000980	C1QTNF3	Lung Cancer
ACH-000810	GDF5	Skin Cancer
ACH-001539	C1QTNF3-AMACR	Lymphoma
ACH-000471	C21orf59	Liver Cancer
ACH-002166	C21orf59	Skin Cancer
ACH-000993	RYR3	Endometrial/Uterine Cancer
ACH-000758	AZIN2	Gastric Cancer
ACH-000813	AZIN2	Lung Cancer
ACH-000996	AZIN2	Endometrial/Uterine Cancer
ACH-000983	AZIN2	Leukemia
ACH-000157	NRP1	Lymphoma
ACH-002136	FAAP24	Lung Cancer
ACH-001544	RAD51L3-RFFL	Esophageal Cancer
ACH-000662	BAG1	Lung Cancer
ACH-001328	YARS	Skin Cancer
ACH-000982	RP1-27O5.3	Colon/Colorectal Cancer
ACH-001850	ZBTB8B	Gallbladder Cancer
ACH-001610	XXbac-BPG181M17.5	Brain Cancer
ACH-000718	XXbac-BPG181M17.5	Lung Cancer
ACH-000124	XXbac-BPG181M17.5	Lymphoma
ACH-002256	XXbac-BPG246D15.9	Leukemia
ACH-000649	XXbac-BPG246D15.9	Kidney Cancer
ACH-000584	HLA-DOB	Ovarian Cancer
ACH-001328	CCL11	Skin Cancer
ACH-000928	MTMR12	Endometrial/Uterine Cancer
ACH-000984	ADGRB2	Endometrial/Uterine Cancer
ACH-000480	ADGRB2	Liver Cancer
ACH-000938	ADGRB2	Leukemia
ACH-000948	ADGRB2	Gastric Cancer
ACH-000947	PPT2-EGFL8	Ovarian Cancer
ACH-000952	PPT2-EGFL8	Prostate Cancer
ACH-000998	PPT2-EGFL8	Colon/Colorectal Cancer
ACH-000999	PPT2-EGFL8	Colon/Colorectal Cancer
ACH-000617	PPT2-EGFL8	Ovarian Cancer
ACH-000517	STK19	Pancreatic Cancer
ACH-000920	XXbac-BPG116M5.17	Leukemia
ACH-000967	XXbac-BPG116M5.17	Colon/Colorectal Cancer
ACH-001199	XXbac-BPG116M5.17	Colon/Colorectal Cancer
ACH-000945	XXbac-BPG116M5.17	Lung Cancer
ACH-000318	C2	Esophageal Cancer
ACH-000318	XXbac-BPG116M5.17	Esophageal Cancer
ACH-000986	XXbac-BPG116M5.17	Colon/Colorectal Cancer
ACH-001061	C2	Colon/Colorectal Cancer
ACH-001061	XXbac-BPG116M5.17	Colon/Colorectal Cancer
ACH-000468	METTL20	Pancreatic Cancer
ACH-000930	MSH5-SAPCD1	Breast Cancer
ACH-000639	MSH5-SAPCD1	Lung Cancer
ACH-000965	MSH5-SAPCD1	Endometrial/Uterine Cancer
ACH-001956	MSH5-SAPCD1	Bone Cancer
ACH-000986	MSH5-SAPCD1	Colon/Colorectal Cancer
ACH-000978	MSH5-SAPCD1	Endometrial/Uterine Cancer
ACH-000745	MSH5-SAPCD1	Myeloma
ACH-000997	MSH5-SAPCD1	Colon/Colorectal Cancer
ACH-001127	XXbac-BPG32J3.20	Leukemia
ACH-002256	XXbac-BPG32J3.22	Leukemia
ACH-002256	LY6G5B	Leukemia
ACH-002510	CSNK2B	Skin Cancer
ACH-002338	CSNK2B	Skin Cancer
ACH-002510	XXbac-BPG32J3.22	Skin Cancer
ACH-002338	XXbac-BPG32J3.22	Skin Cancer
ACH-000992	XXbac-BPG32J3.22	Head and Neck Cancer
ACH-000662	BPIFB2	Lung Cancer
ACH-000901	CCDC129	Lung Cancer
ACH-000444	ATP6V1G2	Lung Cancer
ACH-000992	ATP6V1G2-DDX39B	Head and Neck Cancer
ACH-001127	ATP6V1G2-DDX39B	Leukemia
ACH-000981	ATP6V1G2-DDX39B	Leukemia
ACH-002256	ATP6V1G2-DDX39B	Leukemia

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-002195	ATP6V1G2-DDX39B	Kidney Cancer
ACH-000995	ATP6V1G2-DDX39B	Leukemia
ACH-000342	AC135048.1	Colon/Colorectal Cancer
ACH-000739	RP5-877J2.1	Liver Cancer
ACH-000869	RP5-877J2.1	Lung Cancer
ACH-001137	INMT-FAM188B	Lung Cancer
ACH-000590	INMT-FAM188B	Lung Cancer
ACH-000538	CCDC189	Gastric Cancer
ACH-000639	RP11-2C24.9	Lung Cancer
ACH-002059	RP11-2C24.9	Leukemia
ACH-001610	RP11-2C24.9	Brain Cancer
ACH-000620	RP11-2C24.9	Liver Cancer
ACH-000979	RP11-2C24.9	Prostate Cancer
ACH-000995	AC002310.13	Leukemia
ACH-001127	AC002310.11	Leukemia
ACH-000979	AC002310.11	Prostate Cancer
ACH-000608	CCT8	Ovarian Cancer
ACH-000560	GS1-114I9.3	Gastric Cancer
ACH-000974	TRIM39	Endometrial/Uterine Cancer
ACH-001339	DOC2A	Skin Cancer
ACH-000971	SARAF	Colon/Colorectal Cancer
ACH-000901	SEZ6L2	Lung Cancer
ACH-000515	AC009133.22	Lung Cancer
ACH-001328	OVCH1	Skin Cancer
ACH-000468	HMBOX1	Pancreatic Cancer
ACH-000998	RP11-435I10.4	Colon/Colorectal Cancer
ACH-000957	RP11-435I10.4	Colon/Colorectal Cancer
ACH-000999	RP11-435I10.4	Colon/Colorectal Cancer
ACH-000167	AC110084.1	Gastric Cancer
ACH-001642	AC110084.1	Lymphoma
ACH-001321	STMN4	Thyroid Cancer
ACH-000557	ADGRF3	Leukemia
ACH-000956	ADGRF3	Prostate Cancer
ACH-000584	RP11-96L14.7	Ovarian Cancer
ACH-001339	BTN2A1	Skin Cancer
ACH-000852	NUP58	Lung Cancer
ACH-000166	NUP58	Leukemia
ACH-002245	RSRP1	Lymphoma
ACH-000296	RSRP1	Colon/Colorectal Cancer
ACH-001339	EFR3B	Skin Cancer
ACH-000996	RP11-717K11.2	Endometrial/Uterine Cancer
ACH-000930	RP11-717K11.2	Breast Cancer
ACH-001096	RP11-701H24.9	Sarcoma
ACH-000900	RP11-701H24.9	Lung Cancer
ACH-000899	RP11-701H24.9	Skin Cancer
ACH-000740	RP11-307N16.6	Head and Neck Cancer
ACH-000930	SPATA13	Breast Cancer
ACH-002044	LTB4R	Head and Neck Cancer
ACH-000981	NEDD8-MDP1	Leukemia
ACH-000980	NEDD8-MDP1	Lung Cancer
ACH-000155	RP11-468E2.1	Pancreatic Cancer
ACH-000928	RP11-468E2.4	Endometrial/Uterine Cancer
ACH-000757	ALDH5A1	Lung Cancer
ACH-001339	AQP4	Skin Cancer
ACH-002156	RP11-507M3.1	Lung Cancer
ACH-001113	RP11-507M3.1	Lung Cancer
ACH-001321	UBXN2A	Thyroid Cancer
ACH-000662	CST3	Lung Cancer
ACH-000901	SCNN1G	Lung Cancer
ACH-000757	TNFRSF10C	Lung Cancer
ACH-000872	ADGRA3	Lung Cancer
ACH-000992	ADGRA3	Head and Neck Cancer
ACH-001719	ADGRA3	Ovarian Cancer
ACH-000998	ADGRA3	Colon/Colorectal Cancer
ACH-000871	RP11-294C11.1	Lung Cancer
ACH-001347	RP11-145E5.5	Head and Neck Cancer
ACH-001736	RP11-145E5.5	Leukemia
ACH-000002	RP11-145E5.5	Leukemia
ACH-000546	RP11-145E5.5	Head and Neck Cancer
ACH-002256	RP11-145E5.5	Leukemia

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000796	RP11-145E5.5	Ovarian Cancer
ACH-000935	RP11-145E5.5	Colon/Colorectal Cancer
ACH-002275	RP11-145E5.5	Leukemia
ACH-002164	RP11-145E5.5	Pancreatic Cancer
ACH-001616	RP11-145E5.5	Lymphoma
ACH-001377	RP11-145E5.5	Pancreatic Cancer
ACH-000034	RP11-145E5.5	Leukemia
ACH-001740	RP11-145E5.5	Sarcoma
ACH-001196	RP11-145E5.5	Sarcoma
ACH-000715	RP11-145E5.5	Head and Neck Cancer
ACH-000303	RP11-145E5.5	Gastric Cancer
ACH-002310	RP11-145E5.5	Liposarcoma
ACH-000274	CDKN2A	Skin Cancer
ACH-000993	OSBPL1A	Endometrial/Uterine Cancer
ACH-000982	RP11-145E5.5	Colon/Colorectal Cancer
ACH-001345	RP11-145E5.5	Colon/Colorectal Cancer
ACH-000847	SPX	Gastric Cancer
ACH-000993	DNAH11	Endometrial/Uterine Cancer
ACH-000014	XRN2	Skin Cancer
ACH-002275	RP11-545J16.1	Leukemia
ACH-002275	SLCO1B7	Leukemia
ACH-002215	RP11-545J16.1	Lymphoma
ACH-002215	SLCO1B7	Lymphoma
ACH-000228	RP11-545J16.1	Head and Neck Cancer
ACH-000621	RP11-545J16.1	Breast Cancer
ACH-000593	ANG	Bladder Cancer
ACH-000593	RP11-903H12.5	Bladder Cancer
ACH-001002	RP11-545J16.1	Skin Cancer
ACH-000827	RP11-545J16.1	Skin Cancer
ACH-000997	HACD4	Colon/Colorectal Cancer
ACH-000988	LDAH	Endometrial/Uterine Cancer
ACH-000458	ABCB5	Skin Cancer
ACH-001513	CFAP61	Cervical Cancer
ACH-000961	CFAP61	Endometrial/Uterine Cancer
ACH-000874	CFAP61	Leukemia
ACH-002222	CFAP61	Leukemia
ACH-000754	CFAP61	Lymphoma
ACH-001127	CFAP61	Leukemia
ACH-000556	MALRD1	Cervical Cancer
ACH-000780	MALRD1	Lung Cancer
ACH-001282	YJEFN3	Bone Cancer
ACH-000911	BORCS8-MEF2B	Gastric Cancer
ACH-000911	BORCS8	Gastric Cancer
ACH-000953	RP13-279N23.2	Leukemia
ACH-000279	RP13-279N23.2	Bone Cancer
ACH-000735	ADGRG2	Head and Neck Cancer
ACH-000978	AC002985.3	Endometrial/Uterine Cancer
ACH-001639	ADGRG2	Leukemia
ACH-002094	SAXO1	Bone Cancer
ACH-000537	AC002985.3	Liver Cancer
ACH-000999	ARL6IP1	Colon/Colorectal Cancer
ACH-000696	NT5C1B	Ovarian Cancer
ACH-002222	NT5C1B	Leukemia
ACH-000715	NT5C1B	Head and Neck Cancer
ACH-000400	NT5C1B	Colon/Colorectal Cancer
ACH-001208	AC007192.4	Kidney Cancer
ACH-000953	AC007192.4	Leukemia
ACH-000999	SAA2-SAA4	Colon/Colorectal Cancer
ACH-000349	AC007192.4	Breast Cancer
ACH-001034	SAA2-SAA4	Bone Cancer
ACH-000375	GEN1	Kidney Cancer
ACH-000608	ATPAF2	Ovarian Cancer
ACH-000988	CTD-2278I10.6	Endometrial/Uterine Cancer
ACH-000876	CTD-2278I10.6	Breast Cancer
ACH-000989	CTD-2278I10.6	Colon/Colorectal Cancer
ACH-000386	CTD-2278I10.6	Leukemia
ACH-000953	CTD-2278I10.6	Leukemia
ACH-000977	AC010646.3	Prostate Cancer
ACH-001091	RP11-45M22.4	Colon/Colorectal Cancer
ACH-000072	RP11-45M22.4	Leukemia

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000626	RP11-45M22.4	Myeloma
ACH-000304	RP11-45M22.4	Skin Cancer
ACH-000887	RP11-45M22.4	Brain Cancer
ACH-000834	CTD-3222D19.2	Bladder Cancer
ACH-002166	CTD-3222D19.2	Skin Cancer
ACH-000998	CTD-3222D19.2	Colon/Colorectal Cancer
ACH-000720	CALR3	Bladder Cancer
ACH-000875	CTD-3222D19.2	Lung Cancer
ACH-000105	LRRC75A	Leukemia
ACH-000458	DNAJC16	Skin Cancer
ACH-000438	RP11-385D13.1	Lung Cancer
ACH-000662	EPHX3	Lung Cancer
ACH-000960	EPHX3	Leukemia
ACH-000929	RBSN	Lung Cancer
ACH-000969	RBSN	Colon/Colorectal Cancer
ACH-002022	RBSN	Colon/Colorectal Cancer
ACH-000663	RBSN	Ovarian Cancer
ACH-000452	C12orf60	Esophageal Cancer
ACH-000941	ADGRE2	Endometrial/Uterine Cancer
ACH-000988	ADGRE2	Endometrial/Uterine Cancer
ACH-002217	ADGRE2	Neuroblastoma
ACH-000506	ADGRE3	Lung Cancer
ACH-000838	RP11-140L24.4	Myeloma
ACH-000633	RP11-140L24.4	Gastric Cancer
ACH-000485	RP11-140L24.4	Gastric Cancer
ACH-000990	RP11-140L24.4	Endometrial/Uterine Cancer
ACH-000217	RP11-140L24.4	Liver Cancer
ACH-000983	RP11-140L24.4	Leukemia
ACH-000773	RP11-140L24.4	Lymphoma
ACH-000858	RP11-140L24.4	Lung Cancer
ACH-000227	RP11-140L24.4	Neuroblastoma
ACH-000524	RP11-140L24.4	Ovarian Cancer
ACH-000604	RP11-140L24.4	Leukemia
ACH-000255	RP11-140L24.4	Gastric Cancer
ACH-000321	RP11-140L24.4	Leukemia
ACH-000203	RP11-140L24.4	Neuroblastoma
ACH-000409	RP11-140L24.4	Ovarian Cancer
ACH-000372	RP11-140L24.4	Leukemia
ACH-000030	RP11-140L24.4	Lung Cancer
ACH-000779	RP11-140L24.4	Lung Cancer
ACH-000513	RP11-140L24.4	Kidney Cancer
ACH-000537	RP11-140L24.4	Liver Cancer
ACH-000197	RP11-140L24.4	Leukemia
ACH-000488	RP11-140L24.4	Esophageal Cancer
ACH-000917	RP11-140L24.4	Esophageal Cancer
ACH-000146	RP11-140L24.4	Leukemia
ACH-000630	RP11-140L24.4	Head and Neck Cancer
ACH-000570	RP11-140L24.4	Brain Cancer
ACH-000269	ADGRE5	Brain Cancer
ACH-001539	ADGRE5	Lymphoma
ACH-002256	ADGRE5	Leukemia
ACH-000734	ADGRE5	Liver Cancer
ACH-001061	ADGRL1	Colon/Colorectal Cancer
ACH-000997	ADGRL1	Colon/Colorectal Cancer
ACH-000594	ADGRL1	Lung Cancer
ACH-000938	ADGRL1	Leukemia
ACH-000278	TCEANC	Ovarian Cancer
ACH-000398	BTBD10	Lymphoma
ACH-000781	FAM234B	Lung Cancer
ACH-000999	FAM234B	Colon/Colorectal Cancer
ACH-000993	FAM234B	Endometrial/Uterine Cancer
ACH-001856	WDR83OS	Bile Duct Cancer
ACH-001366	MKRN2OS	Neuroblastoma
ACH-000727	CTD-3105H18.14	Lung Cancer
ACH-000982	ZNF625-ZNF20	Colon/Colorectal Cancer
ACH-001345	ZNF625-ZNF20	Colon/Colorectal Cancer
ACH-000873	CTD-2006C1.10	Esophageal Cancer
ACH-000025	DNAH9	Brain Cancer
ACH-001328	DNAH9	Skin Cancer
ACH-002275	C19orf80	Leukemia

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000278	SMARCA4	Ovarian Cancer
ACH-001709	PRH1-PRR4	Lymphoma
ACH-001786	RP11-637O19.3	Colon/Colorectal Cancer
ACH-000638	RP11-637O19.3	Lung Cancer
ACH-000993	MRVI1	Endometrial/Uterine Cancer
ACH-001318	RP11-277P12.6	Liver Cancer
ACH-001318	KLRC2	Liver Cancer
ACH-000991	RP11-277P12.6	Colon/Colorectal Cancer
ACH-000991	KLRC3	Colon/Colorectal Cancer
ACH-001111	KLRC4-KLRK1	Lymphoma
ACH-000812	KLRC4-KLRK1	Skin Cancer
ACH-002309	APITD1-CORT	Sarcoma
ACH-000269	CTD-2369P2.10	Brain Cancer
ACH-000999	PPAN	Colon/Colorectal Cancer
ACH-001369	ARPC4-TTLL3	Ovarian Cancer
ACH-001377	GSG1L2	Pancreatic Cancer
ACH-000940	CFAP52	Endometrial/Uterine Cancer
ACH-000288	CFAP52	Breast Cancer
ACH-000990	ZNF559-ZNF177	Endometrial/Uterine Cancer
ACH-000996	CFAP52	Endometrial/Uterine Cancer
ACH-000991	CFAP52	Colon/Colorectal Cancer
ACH-001550	ZNF559-ZNF177	Skin Cancer
ACH-000943	MTCL1	Colon/Colorectal Cancer
ACH-000705	ANOS1	Lung Cancer
ACH-002156	ANOS1	Lung Cancer
ACH-001113	ANOS1	Lung Cancer
ACH-001857	ANOS1	Bile Duct Cancer
ACH-000965	ELAVL1	Endometrial/Uterine Cancer
ACH-000977	ELAVL1	Prostate Cancer
ACH-000608	C3AR1	Ovarian Cancer
ACH-000655	ELAVL1	Brain Cancer
ACH-001203	ELAVL1	Lymphoma
ACH-000809	ELAVL1	Esophageal Cancer
ACH-000095	ELAVL1	Brain Cancer
ACH-000620	ELAVL1	Liver Cancer
ACH-002234	ELAVL1	Head and Neck Cancer
ACH-002207	ELAVL1	Skin Cancer
ACH-000614	ELAVL1	Skin Cancer
ACH-002461	ELAVL1	Skin Cancer
ACH-000943	ELAVL1	Colon/Colorectal Cancer
ACH-002188	ELAVL1	Kidney Cancer
ACH-000804	ELAVL1	Neuroblastoma
ACH-000512	ELAVL1	Myeloma
ACH-000890	BLOC1S5	Lung Cancer
ACH-001819	BLOC1S5	Breast Cancer
ACH-000752	RP11-1099M24.7	Lung Cancer
ACH-000269	MCEMP1	Brain Cancer
ACH-002214	CTD-2207O23.10	Lymphoma
ACH-001321	TP53	Thyroid Cancer
ACH-000376	CTD-2207O23.3	Brain Cancer
ACH-000990	CTD-2207O23.3	Endometrial/Uterine Cancer
ACH-000867	SENP3-EIF4A1	Lung Cancer
ACH-000660	SENP3-EIF4A1	Lymphoma
ACH-000393	RP11-542C16.2	Liver Cancer
ACH-000855	RP11-542C16.2	Esophageal Cancer
ACH-000674	RP11-542C16.2	Gastric Cancer
ACH-000409	RP11-542C16.2	Ovarian Cancer
ACH-002287	RP11-542C16.2	Colon/Colorectal Cancer
ACH-000994	RP11-542C16.2	Endometrial/Uterine Cancer
ACH-000985	RP1-4G17.5	Colon/Colorectal Cancer
ACH-001458	RP11-146B14.1	Colon/Colorectal Cancer
ACH-000937	RNASEK-C17orf49	Leukemia
ACH-000914	ADGRE1	Lymphoma
ACH-001081	ADGRE1	Colon/Colorectal Cancer
ACH-000398	C3	Lymphoma
ACH-000993	DCHS1	Endometrial/Uterine Cancer
ACH-000695	ZBTB48	Lung Cancer
ACH-001328	GRID2IP	Skin Cancer
ACH-000274	SLC25A23	Skin Cancer
ACH-000537	MCPH1	Liver Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-000946	CHD5	Endometrial/Uterine Cancer
ACH-000917	CTB-54O9.9	Esophageal Cancer
ACH-000525	RIC1	Lung Cancer
ACH-000999	RIC1	Colon/Colorectal Cancer
ACH-000963	RIC1	Colon/Colorectal Cancer
ACH-002005	RIC1	Skin Cancer
ACH-000984	RIC1	Endometrial/Uterine Cancer
ACH-000806	RIC1	Lymphoma
ACH-000992	RIC1	Head and Neck Cancer
ACH-001239	TRIM6-TRIM34	Skin Cancer
ACH-002002	TRIM34	Skin Cancer
ACH-002091	TRIM6-TRIM34	Skin Cancer
ACH-000685	ICE1	Pancreatic Cancer
ACH-001489	HBG2	Lung Cancer
ACH-000277	EEF2KMT	Breast Cancer
ACH-000988	RP11-234B24.6	Endometrial/Uterine Cancer
ACH-000025	DPP9	Brain Cancer
ACH-000994	MYDGF	Endometrial/Uterine Cancer
ACH-000989	MYDGF	Colon/Colorectal Cancer
ACH-000925	CTB-50L17.14	Lung Cancer
ACH-000970	TIGAR	Colon/Colorectal Cancer
ACH-002124	CORO7-PAM16	Lung Cancer
ACH-000677	CHAF1A	Lung Cancer
ACH-000860	CORO7-PAM16	Lung Cancer
ACH-000988	CORO7-PAM16	Endometrial/Uterine Cancer
ACH-000981	CORO7-PAM16	Leukemia
ACH-001791	CORO7-PAM16	Liposarcoma
ACH-000990	CORO7-PAM16	Endometrial/Uterine Cancer
ACH-000921	CRACR2A	Lung Cancer
ACH-000998	NCBP3	Colon/Colorectal Cancer
ACH-000893	NCBP3	Lung Cancer
ACH-000662	TRAP1	Lung Cancer
ACH-000941	CRACR2A	Endometrial/Uterine Cancer
ACH-000981	NCBP3	Leukemia
ACH-001127	NCBP3	Leukemia
ACH-000468	PRMT8	Pancreatic Cancer
ACH-001127	P2RX5	Leukemia
ACH-000491	P2RX5	Colon/Colorectal Cancer
ACH-000990	RP11-235E17.2	Endometrial/Uterine Cancer
ACH-000977	SMIM24	Prostate Cancer
ACH-000123	ITFG2	Ovarian Cancer
ACH-000426	PUM3	Myeloma
ACH-000921	CSMD1	Lung Cancer
ACH-000695	C5orf38	Lung Cancer
ACH-000649	RP11-20I23.1	Kidney Cancer
ACH-000649	RP11-20I23.3	Kidney Cancer
ACH-000806	RP11-20I23.1	Lymphoma
ACH-000247	RP4-734P14.4	Gastric Cancer
ACH-001421	CFAP99	Eye Cancer
ACH-000960	PLCH2	Leukemia
ACH-001151	TRPM5	Ovarian Cancer
ACH-000517	ZFYVE28	Pancreatic Cancer
ACH-000949	CFAP74	Gastric Cancer
ACH-001390	CRAMP1	Breast Cancer
ACH-001390	LA16c-431H6.6	Breast Cancer
ACH-000786	WNT5B	Lymphoma
ACH-000998	CRAMP1	Colon/Colorectal Cancer
ACH-001610	AC005943.2	Brain Cancer
ACH-000025	ATAD3C	Brain Cancer
ACH-000953	RP11-314N13.10	Leukemia
ACH-001091	RP11-314N13.10	Colon/Colorectal Cancer
ACH-000757	TPSD1	Lung Cancer
ACH-000925	CBARP	Lung Cancer
ACH-000993	MUC2	Endometrial/Uterine Cancer
ACH-000941	AL645608.1	Endometrial/Uterine Cancer
ACH-001613	AL645608.1	Leukemia
ACH-001719	CRACR2B	Ovarian Cancer
ACH-000988	PLPPR3	Endometrial/Uterine Cancer
ACH-001517	PLPPR3	Endometrial/Uterine Cancer
ACH-000788	PLPPR3	Skin Cancer

Continued on next page

Table C.1 – continued from previous page

DepMap ID	SYMBOL	DISEASE
ACH-002349	PLPPR3	Unknown
ACH-000996	PIDD1	Endometrial/Uterine Cancer
ACH-000987	PIDD1	Skin Cancer
ACH-000985	PIDD1	Colon/Colorectal Cancer
ACH-001391	PIDD1	Breast Cancer
ACH-001302	PIDD1	Neuroblastoma
ACH-000918	PIDD1	Leukemia
ACH-002256	PIDD1	Leukemia
ACH-001151	KDM5A	Ovarian Cancer
ACH-000111	AHRR	Breast Cancer
ACH-001339	SLC6A12	Skin Cancer
ACH-002285	PLPP2	Neuroblastoma
ACH-000008	PLPP2	Skin Cancer
ACH-000691	SCGB1C2	Breast Cancer

C.3 Functional screening - Drug sensitivity

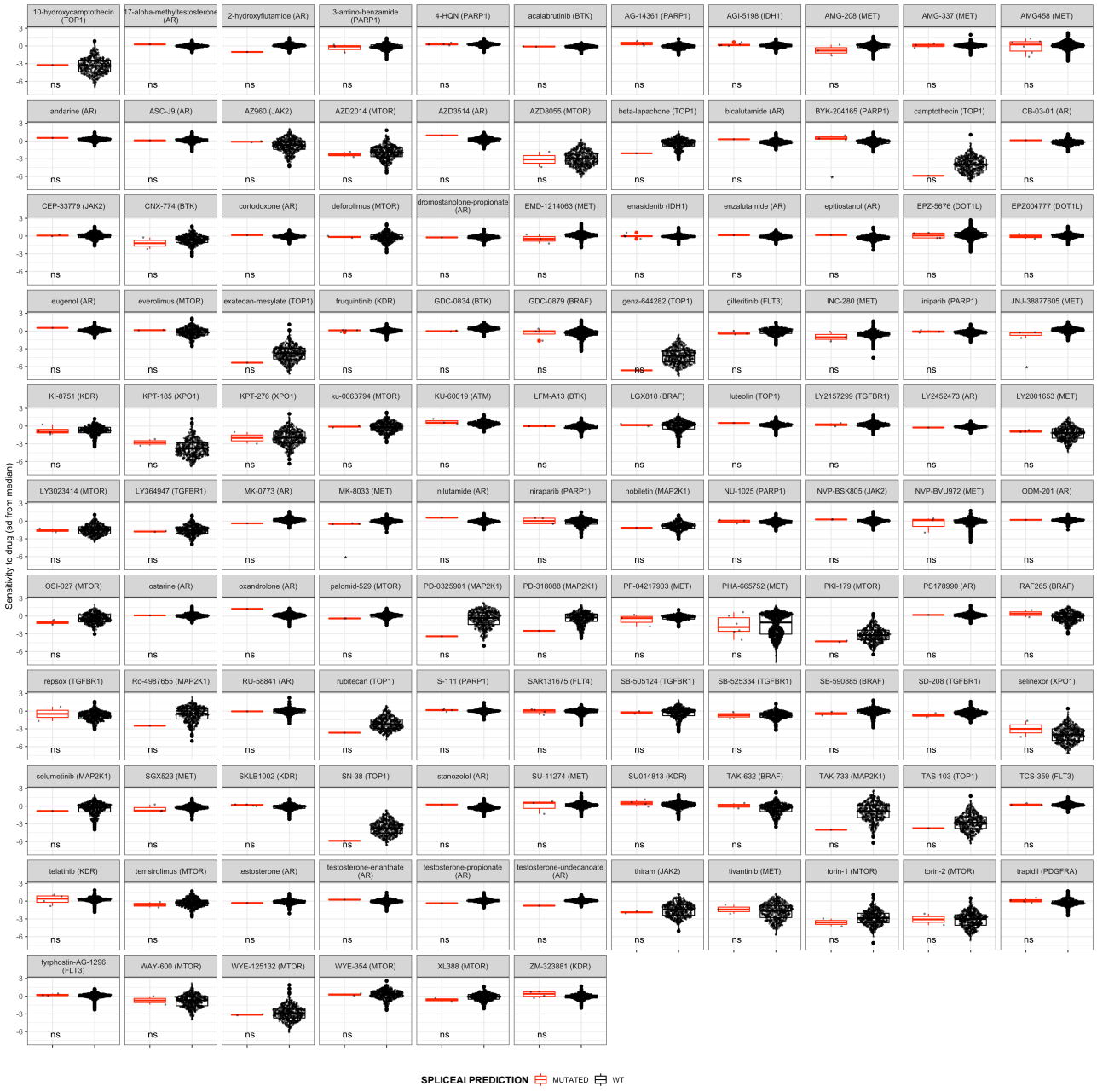

Figure C.2: Difference in sensitivity to drugs from PRISM per gene comparing cell lines where the gene is predicted to be splice-altered according to SpliceAI vs. cell lines where SpliceAI did not predict a mutation. We selected those genes where a splice-altering variant occurred in at least 5 cell lines.In the context of this analysis WT means that the gene does not carry a variant predicted by SpliceAI. Significance assessed using Welch's t-test.

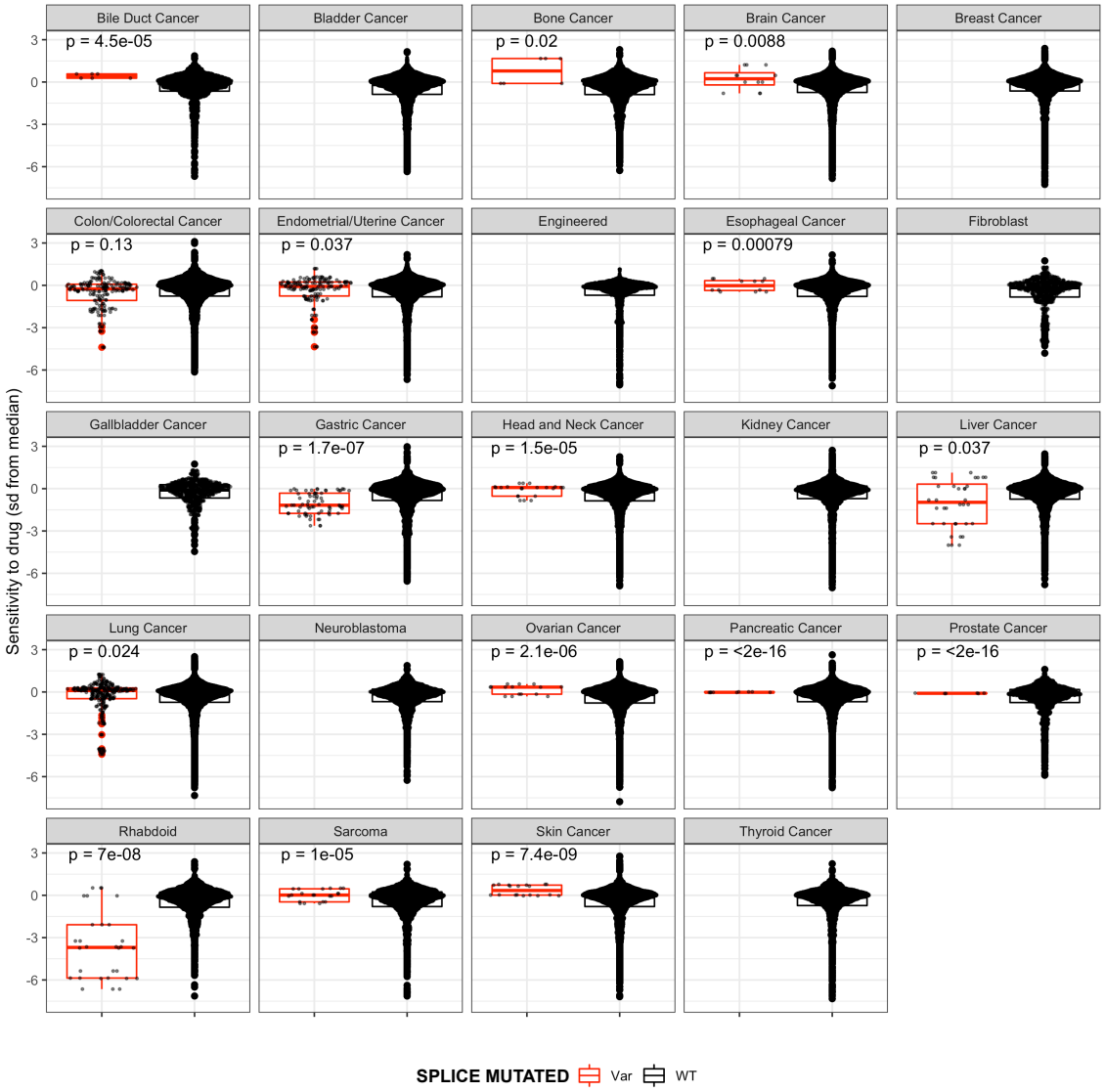

Figure C.3: Comparison of sensitivity to PRISM repurposing drugs per cancer. Not filtered for MSK-IMPACT panel genes. Significance estimated using Welch's t-test. Note that some cancers do not present any splice altering mutation, while in Figure C.1 we can see that most of the cancers present off-target mutations and therefore must carry splice altering mutations. The difference is because not all cell lines present in the somatic mutation dataset (CCLE) were studied in the PRISM repurposing project.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY