Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics,
Engineering and Natural Sciences
of the
Ruprecht - Karls - University
Heidelberg

Presented by
Nicholas Allen Baclig Abad

Born in: California, United States of America
Oral Examination: 11 October 2024

REMIND-Cancer:
A Recurrence-Agnostic Workflow for Identifying and Prioritizing Functional Promoter Single Nucleotide Variants

Referees: Prof. Dr. Carl Herrmann

Prof. Dr. Benedikt Brors

Abstract

The discovery of functional cancer drivers has traditionally focused on mutations within coding regions of DNA, despite these regions comprising only about 2% of the entire genome. However, studies have shown that the remaining 98% of the genome, known as non-coding DNA or "junk" DNA, also harbors mutations that can alter gene functionality, most notably the two hotspot SNVs within the promoter of *TERT*. Beyond these two SNVs, however, the catalog of known functional mutations within promoter regions remains limited. This scarcity can primarily be attributed to existing methods relying on a singular SNV being observed in a large amount of patients (i.e. high recurrence) thereby having high statistical power for their detection. Since the vast majority of SNVs in current datasets do not meet this criterion and are instead categorized as singletons or lowly-recurrent SNVs, there is a need for new approaches to identify this underrepresented class of mutations.

To address this gap, I developed the REMIND-Cancer workflow, which is a recurrence-agnostic approach designed to identify and prioritize functional SNVs within promoter regions of protein-coding genes. This workflow, which follows a filtering-ranking-inspection-validation process, was applied to two pan-cancer datasets, resulting in the identification of 10 promoter SNVs that activate their corresponding promoters, as validated *in vitro* using a luciferase assay. Aiming to have a translational impact, this workflow was also applied to two ongoing precision oncology programs as a pilot study to evaluate the effectiveness of my approach and its computational efficiency.

Broadly, this thesis highlights the importance of identifying functional mutations within the non-coding genome, beyond those that are highly recurrent, in order to advance personalized oncology.

Zusammenfassung

Die Entdeckung funktioneller Krebstreiber hat sich traditionell auf Mutationen in den kodierenden Bereichen der DNA konzentriert, obwohl diese Bereiche nur etwa 2% des gesamten Genoms ausmachen. Studien haben jedoch gezeigt, dass die verbleibenden 98% des Genoms, die als nicht codierende DNA oder "Junk"-DNA bekannt sind, ebenfalls Mutationen enthalten, die die Genfunktionalität verändern können, insbesondere die beiden Hotspot-SNVs im Promotor von *TERT*. Abgesehen von diesen beiden SNVs ist der Katalog der bekannten funktionellen Mutationen in Promotorregionen jedoch begrenzt. Diese Seltenheit ist in erster Linie darauf zurückzuführen, dass die vorhandenen Methoden darauf angewiesen sind, dass eine einzelne SNV bei einer großen Anzahl von Patienten beobachtet wird (d. h. eine hohe Aufkommensrate) und somit eine hohe statistische Aussagekraft für ihre Entdeckung hat. Da die überwiegende Mehrheit der SNVs in den aktuellen Datensätzen dieses Kriterium nicht erfüllt und stattdessen als Singletons oder selten aufkommende SNVs eingestuft werden, besteht ein Bedarf an neuen Ansätzen zur Identifizierung dieser unterrepräsentierten Klasse von Mutationen.

Um diese Lücke zu schließen, habe ich den REMIND-Cancer Arbeitsablauf entwickelt, einen Aufkommensrate-agnostischen Ansatz zur Identifizierung und Priorisierung funktioneller SNVs in Promotorregionen von proteinkodierenden Genen. Dieser Arbeitsablauf, der Filter-, Priorisierungs-, Inspizierungs- und Validierungsprozesse beinhaltet, wurde auf zwei Pan-Krebs-Datensätze angewandt, was zur Identifizierung von 10 Promotor-SNVs führte, die ihre entsprechenden Promotoren aktivieren, wie *in vitro* mit einem Luciferase-Assay validiert wurde. Mit dem Ziel, eine translationale Wirkung zu erzielen, wurde dieser Arbeitsablauf auch auf zwei laufende Präzisionsonkologie-Programme als Pilotstudie angewandt, um die Wirksamkeit meines Ansatzes und seine rechnerische Effizienz zu bewerten.

Zusammenfassend unterstreicht diese Arbeit die Relevanz, funktionelle Mutationen innerhalb des nicht-kodierenden Genoms zu identifizieren, die über die häufig aufkommenden Mutationen hinausgehen, um die personalisierte Onkologie voranzubringen.

The above text was translated from English to German using AI-assisted technology, particularly DeepL (https://www.deepl.com/en/translator).

List of Figures

List of Tables

Contents

1 Introduction to REMIND-Cancer

1.1 Cancer

In 2022, nearly 20 million new cancer cases were diagnosed worldwide with 4.4 million in Europe alone (Ferlay et al., 2024). Cancer also caused approximately 10 million deaths globally, making it responsible for one in six deaths (Ferlay et al., 2024) (WorldHealthOrganization, 2024). Despite the common misconception of cancer as a single disease, it comprises over 200 distinct types (CancerResearchUK, 2023) each of which may have unique subtypes, incidence (i.e. number of new cases) rates, mortality (i.e. number of deaths) rates, therapeutic optionss and patient responses. Consequently, studying the various types of cancer and their specific characteristics is essential for developing targeted treatments, improving patient outcomes and advancing our collective understanding of this complex set of diseases.

Cancer is a heterogeneous disease that arises from alterations or mutations in the DNA of cells, thereby being categorized as a genetic disease. These changes potentially cause normal cells to become abnormal and grow uncontrollably, which is similar to Darwinian evolution where abnormal cells undergo clonal expansion. This implies that these abnormal cells replicate and create more cells like themselves by outcompeting normal cells.

This cellular transformation from normal to abnormal occurs when DNA acquires one or more mutations or alterations, which can be caused by a number of different factors. These factors include genetic predispositions (i.e. inherited tendencies to develop certain types of cancer), environmental exposures (e.g. ultraviolet radiation) and lifestyle choices (e.g. smoking, alcohol consumption, diet) (Panno, 2005).

As these abnormal cells continue to multiply in subsequent generations, they can form a mass of tissue known as a tumor. If the tumor were to grow and spread, it can thus invade nearby tissues, disrupt their normal function and metastasize to other parts of the body. This spread of cancer cells to other organs or tissues is what often makes cancer particularly dangerous, as it can affect the function of vital parts of the body.

Mutations inherited from germ cells (sperm and egg) are known as *germline* mutations whereas those occurring after conception are termed *somatic* mutations. Unlike germline mutations, somatic mutations are not inherited by offspring. Somatic mutations, in particular, are extensively studied for their role in cell behavior during proliferation and mitosis. Somatic mutations in cancer cells are passed to their cellular descendants, forming clones that trace back to the original mutated cell. If a mutation provides a growth advantage, it leads to the expansion of mutated cells, fueling cancer progression.

The term *mutation* encompasses a variety of genomic events. Insertions or deletions (indels) can result in frame shifts while single nucleotide variants (SNVs), also known as point mutations or single base substitutions, involve the substitution of one nucleotide for another. Other mutation types include inversions (reversal of a DNA segment's orientation), translocations (transfer of a DNA segment within or between chromosomes), duplications (copying of a DNA segment), and deletions (deletion of a DNA segment). Typically, genomic alterations larger than 50 basepairs (bps) are considered to be structural variants (SVs) though the size sometimes varies and is somewhat arbitrary in the literature (Mahmoud et al., 2019).

Although various mutation types contribute to cancer, this thesis primarily focuses on the most common type, single nucleotide variants (SNVs) (Spencer, Zhang, & Pfeifer, 2015).

1.1.0.1 Drivers and Passengers

Though tens of thousands of somatic alterations are within a typical cancer genome (McFarland, Korolev, Kryukov, Sunyaev, & Mirny, 2013), only a small number are considered to drive tumor progression through a selective advantage (Martincorena & Campbell, 2015). It is estimated that only about four to five of these mutations, typically referred to as *driver mutations*, *driver events* or *drivers*, are present within a typical cancer genome (PCAWG Consortium, 2020)(Vogelstein et al., 2013) whereas the other mutations are classically referred to as *passenger mutations* or *passengers*. However, this number of driver events varies depending on the cancer type. For instance, sarcomas, thyroid and testicular cancer each harbor about one driver mutation per tumor while endometrial and colorectal cancers harbor about 10 driver events (Martincorena et al., 2017).

Though driver events were thought to work individually, recent findings have shown that drivers can work with other drivers (i.e. driver-driver) and even other passengers (i.e. driver-passenger) to amplify their oncogenic potential (Saito et al., 2020) (Hanker et al., 2021). This would thus allow weak and infrequent drivers to display strong effects if in the proper context but be neutral in another environment (Ostroverkhova, Przytycka, & Panchenko, 2023). Moreover, characterizing a mutation, in a binary sense, to be either a driver or a passenger is not straightforward as some mutations become drivers at a later stage of cancer evolution, which are referred to as *latent drivers* or *mini-drivers* (Ostroverkhova et al., 2023) (Nussinov & Tsai, 2015) (Yavuz, Tsai, Nussinov, & Tuncbag, 2023).

Whether acting independently or synergistically, multiple studies have aggregated lists of driver mutations such as IntOGen (Gonzalez-Perez et al., 2013), OncoVar (T. Wang et al., 2021), CNCDatabase (E. M. Liu, Martinez-Fundichely, Bollapragada, Spiewack, & Khurana, 2021), and, most notably, Catalogue Of Somatic Mutations In Cancer (COSMIC) (Tate et al., 2019) though more and more mutations are constantly being added to these catalogs.

Biologically, driver mutations influence cellular pathways by altering the functionality of their associated genes. These mutations belong to the broader category of functional mutations, which affect gene activity. Demonstrating this functionality can be achieved by introducing a mutation into a model organism or cell line and directly measuring the resulting changes in activity. Therefore, observing a statistically significant change confirms the mutation's functional activity. Consequently, the discovery of novel functional mutations constitutes a major goal in modern genetics and genomics studies (Cline & Karchin, 2011).

1.1.0.2 Non-Coding Region

The discovery of functional cancer drivers has traditionally focused on genes that directly code for proteins (i.e. coding region) (Khurana et al., 2016) (Rheinbay et al., 2020). Protein-coding mutations have the ability to alter the codon that determines the amino acid sequence of a protein and could be split into three categories: *silent* (i.e. same amino acid with/without the mutation), *missense/nonsynonymous* (i.e. different amino acid after the mutation), or *nonsense* (i.e. stop codon is created rather than an amino acid codon). Notably, the Kirsten rat sarcoma viral oncogene homologue (*KRAS*), arguably the most well-known oncogene and the most common oncogenic driver in human cancers (L. Huang, Guo, Wang, & Fu, 2021) (Cox, Fesik, Kimmelman, Luo, & Der, 2014) (Prior, Lewis, & Mattos, 2012), has a high frequency of protein-coding SNVs. These mutations lead to the activation of the KRAS protein, resulting in uncontrolled cell division and growth in cancers such as colorectal, pancreatic, and lung cancer (L. Huang et al., 2021) (Wood, Hensing, Malik, & Salgia, 2016).

While the majority of these discoveries have focused on the coding region of DNA, this region only constitutes about one to two percent of the human genome. The remaining 98% is non-coding DNA, historically considered "junk DNA". However, our current understanding of this vast region is limited due to the lack of reliable annotations and lack of effective tools to analyze these regions (Patel & Wang, 2018) (Weinhold, Jacobsen, Schultz, Sander, & Lee, 2014) (Khurana et al., 2016) (Lochovsky, Zhang, Fu, Khurana, & Gerstein, 2015). Despite being understudied, mutations in non-coding regions can disrupt the function of regulatory elements, leading to abnormal gene expression and contributing to cancer development. These non-coding regions contain critical regulatory elements, such as enhancers, silencers, and promoters, which are discussed in the following section. As research advances, our understanding of non-coding DNA in cancer biology is evolving, revealing that it is far from "junk" but rather an important component in the complex regulation of the genome.

1.2 Non-Coding Elements

As mentioned previously, several commonly studied non-coding elements play crucial roles in gene regulation through the recruitment of regulatory proteins known as transcription factors (TFs). Within this section, three distinct non-coding elements will be discussed in detail: enhancers (Section 1.2.1), silencers (Section 1.2.2) and promoters (Section 1.2.3). Briefly, enhancers and silencers are relatively short DNA sequences that can modulate gene transcription from considerable distances by interacting with the promoter in order to increase or decrease transcription, respectively. Promoters are also relatively short DNA sequences located near the gene they regulate. They initiate the transcription process by providing a binding site for RNA polymerase, which is essential for the start of transcription. Additionally, promoters can attract other TFs to regulate gene expression more precisely. Together, these non-coding elements are integral to the intricate regulation of gene expression, highlighting the complexity of genomic regulation beyond just the coding sequences.

1.2.1 Enhancers

Enhancers, discovered approximately 40 years ago (Moreau et al., 1981) (Banerji, Rusconi, & Schaffner, 1981), are DNA sequences that can enhance or increase the transcription of one or multiple genes. The genomic position of these regulatory elements are extremely diverse due to being able to be upstream or downstream of the transcription start site (TSS), with some located as far as 100,000 base pairs away from the TSS (Riethoven, 2010).

To help regulate gene activity, enhancers harbor short, evolutionary-conserved DNA motifs, otherwise known as transcription factor binding sites (TFBSs) (See Section 1.3), in which TFs are able to bind and help regulate gene activity (Shlyueva, Stampfel, & Stark, 2014). Importantly, a single enhancer can accommodate multiple binding sites for multiple TFs, allowing for a complex network of gene regulation as this enhancer element could theoretically influence multiple genes.

Research has shown that enhancers can also be involved in long-range chromatin interactions, forming loops that bring them into close proximity with their target promoters, despite being distant (Calo & Wysocka, 2013). This looping mechanism is facilitated by protein complexes, particularly the mediator complex (Ramasamy et al., 2023) (Allen & Taatjes, 2015), which play crucial roles in the spatial organization of the genome and in the regulation of gene expression. Additionally, enhancers can undergo various modifications, such as histone acetylation (Pradeepa, 2017) and methylation (Sharifi-Zarchi et al., 2017), which can further affect their activity and the accessibility of TFBSs.

1.2.2 Silencers

Silencers, a counterpart to enhancers, act as regulatory DNA elements with *repressive* functions, aiming to reduce the activity of linked promoters. Like enhancers, silencers exhibit position- and orientation-independent actions (Segert, Gisselbrecht, & Bulyk, 2021) and recruit TFs. However, unlike enhancers that recruit activators, silencers recruit transcriptional repressors (Segert et al., 2021). An example of a repressive element binding to a silencer is the TF Snail, which has been associated with "antilooping" thereby disrupting enhancer-promoter interactions to silence gene expression (Chopra, Kong, & Levine, 2012).

Despite their discovery nearly 40 years ago (Brand, Breeden, Abraham, Sternglanz, & Nasmyth, 1985), silencers remain significantly understudied compared to enhancers (Segert et al., 2021) (Pang, van Weerd, Hamoen, & Snyder, 2023). However, research into this non-coding element is ongoing and gaining momentum in the field of genomics (T. Zhang, Li, Sun, Xu, &

Wang, 2023).

1.2.3 Promoters

A promoter is a region of DNA that initiates and regulates the transcription of a particular gene. This region serves as the binding site for RNA polymerase and other TFs, which are necessary for transcription to begin and are typically found near the TSS of the gene that they regulate. In general, the two main types of promoters are the *core* promoter and *proximal* promoter, which differ in their functionality. Briefly, the core promoter, which is located near the TSS, contains essential elements necessary for binding the basic transcription machinery (i.e. RNA polymerase II and general transcription factors) thereby initiating transcription. On the other hand, proximal promoters lie upstream of the core promoter and contain binding sites for specific TFs that can activate/enhance or inhibit/repress the transcription of the associated gene. Within this section, these two classes of promoters will be described in detail.

1.2.3.1 Core Promoters

The core promoter region of a gene is a short, conserved DNA sequence in both prokaryotes and eukaryotes that typically ranges from upstream (5' flanking region) of the TSS to potentially past the TSS and into the first exon of the gene. In this region, RNA polymerase and other necessary proteins are able to bind and begin the transcription process to synthesize RNA from the DNA template, which ultimately leads to the production of mRNA. By allowing these necessary proteins, otherwise known as *general transcription factors* (GTFs), to bind, the core promoter specifically determines the location of the TSS and direction of transcription (Andersson & Sandelin, 2020). GTFs, as well as other regulatory elements within the core promoter (i.e. TATA-box, initiator), that bind to the core promoter region are detailed in Section 1.3.1.

Core promoters are typically defined as being within $\pm$ 50-100 bp of the TSS (Roy & Singer, 2015) (Sloutskin, Shir-Shapira, Freiman, & Juven-Gershon, 2021) and have been observed from bacteria to metazoans (Roy & Singer, 2015). Common core promoter elements include the TATA box, Intiator (Inr), and TFIIB recognition elements (BRE) that help in the initiation of transcription. Without a functional core promoter, the entire transcription machinery would not properly assemble at the TSS and the transcription of the subsequent gene would not begin.

1.2.3.2 Proximal Promoters

The proximal promoter is another part of the promoter that modulates the efficiency and rate of transcription but is not necessary to start transcription. Similar to the core promoter, TFs bind to this region of DNA but are primarily specific or gene regulatory TFs rather than GTFs. This class of TFs are further detailed in Section 1.3.2. These DNA-binding TFs can then directly or indirectly (i.e. through the recruitment of co-activators (Andersson & Sandelin, 2020)) influence core promoter activity.

Though the typical proximal promoter length ranges between 100 (i.e. TSS $\pm$ 50) to 1,000 (i.e. TSS $\pm$ 500) bps long (Le, Yapp, Nagasundaram, & Yeh, 2019), no length is consistently used as this distance differs between genes and tissue context (X. Chen, Wu, Hornischer, Kel, & Wingender, 2006). Currently, there is no universal definition of promoter sequences though statistical modeling and machine learning approaches have attempted to predict these exact lengths (Ohler, Harbeck, Niemann, & Reese, 1999) (Umarov & Solovyev, 2017) (K. Song, 2012) (Le et al., 2019).

1.2.3.3 Annotating Promoters

Determining the specific sequence and subsequent length of a promoter is important due to the inclusion or exclusion of TFBSs. If incorrectly specified, a TFBS may be missed completely or not fully recognized, which can thus drastically affect the expression of its adjacent gene.

Although different promoter databases have been generated such as the Tissue Specific Promoter Database (TiProD) (X. Chen et al., 2006) and the Eukaryotic Promoter Database (EPD) (Périer, Praz, Junier, Bonnard, & Bucher, 2000), verifying the accuracy of these predictions is challenging. These databases often do not agree with one another due to several reasons such as diferences in tissue specificity, number of genes included and variations in gene reference annotation. As a result, studies characterizing the effects of mutations have stuck to larger, more general definitions of promoters (e.g. TSS $\pm$ 500) to be as inclusive as possible (Rheinbay et al., 2020). This inclusivity helps capture a wider range of potential regulatory elements, reducing the risk of missing critical TFBSs.

1.3 Transcription Factors

As previously-noted, TFs are proteins that recognize TFBSs in order to bind to DNA and therefore regulate the expression of adjacent genes. There are two primary classes of TFs: general (or basal) TFs and gene-specific regulatory TFs. Broadly, general transcription factors (GTFs), otherwise known as basal TFs, are a part of the transcription pre-initiation complex (PIC), facilitating the recruitment and positioning of RNA Polymerase II (RNA Poly II) at the TSS, enabling transcription (Sikorski & Buratowski, 2009). This transcriptional complex assembles at the core promoter.

On the other hand, gene-specific regulatory TFs are proteins that still bind to TFBSs but rather than forming the PIC, they modulate the transcription of their target genes directly or by interacting with other regulatory elements (i.e. enhancers, silencers). Both classes of TFs play important roles in gene regulation and will be described in more detail below.

1.3.1 General TFs

GTFs are proteins that assist in initiating transcription by helping RNA Poly II bind to the core promoter region of genes in all eukaryotes. These six GTFs, namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, along with RNA Poly II assemble, into the PIC whose formation and function is critical in regulating transcription (Freiman, 2009) (Sikorski & Buratowski, 2009).

The most important GTF is TFIID, which contains multiple subunits such as the TATA-binding protein (TBP) and approximately 13 to 14 TBP-associated factors (TAFs) (Cler, Papai, Schultz, & Davidson, 2009). The TBP is critical because it recognizes the TATA-box, which is a conserved sequence (TATAAA 5' to 3') located 25-30 base pairs upstream of the TSS in both eukaryotes and archaea (Burley, 1996). Although TBP alone can bind to TATA boxes, the inclusion of its associated factors allows the complex to recognize other TFBSs such as the initiator (Inr) and the downstream promoter element (DPE) (Akhtar & Veenstra, 2011).

The other GTFs support the binding process in various ways (Cler et al., 2009) (Freiman, 2009). Briefly, TFIIA structurally stabilizes TFIID binding to the TATA box and helps recruit TFIID to the promoter if the TATA box does not exist. TFIIH unwinds the DNA around the TSS to allow chromatin to be accessible, thus allowing RNA Poly II to access the template strand

and begin RNA synthesis. TFIIB binds to both TBP (within TFIID) and DNA, bridging the interaction between TFIID and RNA Poly II. This TF also helps in aligning RNA Poly II at the correct start site. Lastly, TFIIF stabilizes the binding of RNA Poly II to the DNA and TFIIE recruits and modulates the activity of TFIIH. Together, these GTFs and RNA Poly II bind to the promoter to initiate the transcription process (Latchman, 1997) (Figure 1.3.1).

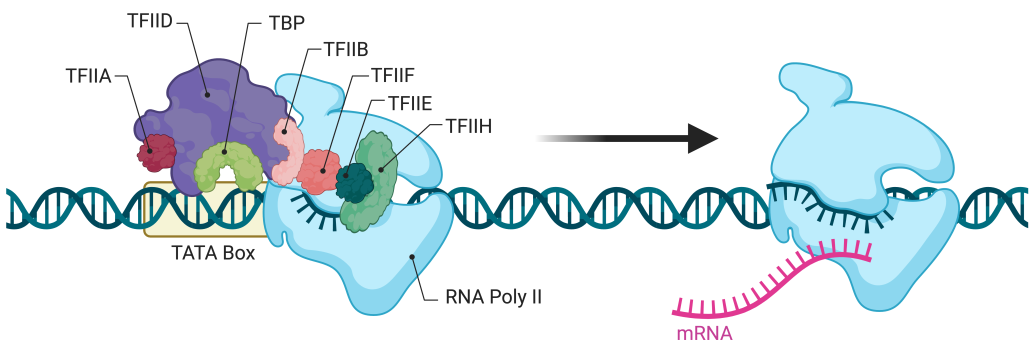

Figure 1: *An overview of the pre-initiation complex (PIC) in which general transcription factors (GTFs) and RNA Polymerase II (RNA Poly II) bind to the core promoter via the TATA Box in order to initiate the transcription process. Created with* `BioRender.com`.

1.3.2 Gene-Specific Regulatory TFs

The other class of TFs are known as gene-specific regulatory TFs. Unlike GTFs, this class of proteins do not participate directly in PIC formation. Instead, these TFs modulate the transcription rate of specific genes by binding to promoters, enhancers, or silencers. With approximately 1,400 to 1,900 total human TFs (Ignatieva, Levitsky, & Kolchanov, 2015) (Göös et al., 2022) and only several GTFs, a majority of TFs fall into this category. In typical nomenclature, this abundantly represented category of TFs is generally referred to simply as *transcription factors* while GTFs are specifically designated as *general transcription factors*. This thesis will adopt this convention: unless otherwise specified, any reference to a TF will pertain to a gene-specific regulatory TF.

Upon binding to DNA, the impact that a specific TF has on transcription is highly variable and context specific (Lambert et al., 2018) though it is thought that a majority of TFs act by recruiting co-factors (Reiter, Wienerroither, & Stark, 2017) (Z. Wang et al., 2021). These co-factors do not necessarily bind to DNA but, instead, modify and influence TF activity through various mechanisms. The two main types of TF co-factors are co-activators and co-repressors. Co-activators increase the activity of TFs by modifying the chromatin structure to make DNA more accessible for transcription or by recruiting RNA Poly II to the core promoter (Näär, Lemon, & Tjian, 2001). Co-repressors do the opposite and make the DNA less accessible thereby repressing transcription (Perissi, Jepsen, Glass, & Rosenfeld, 2010).

Adding to the complexity of TFs, one key feature of transcriptional regulation is that genes are often regulated by more than one TF (Wagner, 1999) (M.-J. M. Chen et al., 2012). At times, groups or clusters of very closely spaced TFBSs occur thereby indicating that these TFs are involved in the gene's regulation (Wagner, 1999). Regardless, TFs typically may work in tandem with many other regulatory elements thus creating a large and complex network of TFs influencing the activity of other TFs as well as regulating multiple genes.

1.3.3 Detecting Transcription Factor Binding Sites

Understanding TFs and their importance is not possible without first understanding their corresponding binding sites or motifs.

Experimentally, motifs can be detected using chromatin immunoprecipitation sequencing assays along with sequencing techniques, which is referred to as ChIP-seq (Mundade, Ozer, Wei, Prabhu, & Lu, 2014). However, other methods, such as using gel electrophoresis mobility shift assay (EMSA) (Hellman & Fried, 2007) (Gurevich, Zhang, & Aneskievich, 2010), Systematic Evolution of Ligands by EXponential enrichment (SELEX) and protein binding microarrays (PBMs) are also utilized (Hombach, Schwarz, Robinson, Schuelke, & Seelow, 2016).

Using these methods of identification, several databases have been created to catalog different motifs such as TRANSFAC (Matys et al., 2003), JASPAR (Fornes et al., 2020), HT-SELEX (Jagannathan, Roulet, Delorenzi, & Bucher, 2006), and others (Lambert et al., 2018). In these databases, the specific sequence in which a certain TF could bind is typically displayed in the form of a sequence logo (Schneider & Stephens, 1990), which are nucleotide letters stacked on top of one another for each position in the aligned sequence, as exemplified in Figure 2. The height of each letter is made proportional to its frequency and normalized for context whereas the letters are sorted in descending order.

In particular, the JASPAR database (Fornes et al., 2020) is an openly-accessible resource that curates TFs and their binding profiles in the form of position frequency matrices (PFMs), which are then used to create a sequence logo. This database aggregates data across various taxonomic groups and species, including Homo sapiens, derived from ChIP-seq, SELEX and PBM experiments. Moreover, JASPAR offers users the option to select from highly confident TFs, termed JASPAR CORE, or unverified TFs, termed JASPAR UNVALIDATED, along with their corresponding binding profiles. As of 28 February 2024, the JASPAR CORE database contains 727 unique TFs and their corresponding sequence logos.

Through the utilization of these plots, the evolutionary conservation of a particular TFBS could be assessed in which large letters represent a high propensity of a nucleotide to occur at this position. As TFBSs are well-conserved, certain mutations, such as SNVs, can disrupt their normal process by either creating a new binding site for a TF to recognize or by destroying a pre-existing binding site, thus not allowing the TF to bind. Several studies have specifically attempted to predict the impact of SNVs on these motifs through various mechanisms (Yiu Chan, Gu, Bieg, Eils, & Herrmann, 2019) (Carrasco Pro, Bulekova, Gregor, Labadorf, & Fuxman Bass, 2020) (Fornes et al., 2018).

1.3.4 Transcription Factor Families

In addition to understanding TFBSs and their detection, another significant aspect of TFs is the concept of a TF family. A TF family comprises related TFs sharing common structural traits and typically perform similar functions in regulating gene expression. These families are grouped based on similarities within their DNA-binding domain (DBD), which is the region of the TF that specifically binds to the TFBS (Lambert et al., 2018). The number of TF families within humans is estimated to be around 50 to 100 (Fornes et al., 2020) (Jolma et al., 2013) but this varies on the individual classification method used as their is no fully agreed-upon strategy.

Several families, such as the E-twenty six (E26) family and the Nuclear

Factor kappa-light-chain-enhancer of activated B cells (NF-κB) family, have been highly implicated in multiple different cancers including prostate cancer, colorectal cancer, and melanoma (T. Hsu, Trojanowska, & Watson, 2004)(Dolcet, Llobet, Pallares, & Matias-Guiu, 2005). These two families, in particular, will be discussed briefly below.

1.3.4.1 E26 (ETS)

The E26 family of proteins, sometimes also known as ETS factors, is considered one of the largest TF families due to having 28 genes within 12 different subfamilies (Qian, Li, & Chen, 2022). TFs within this family, such as ETS1, ETV5, and ELK4, share a highly-conserved DBD, otherwise known as the ETS domain, that binds to GGA(A or T) (5' to 3' direction) motifs in DNA as exemplified by Figure 2.

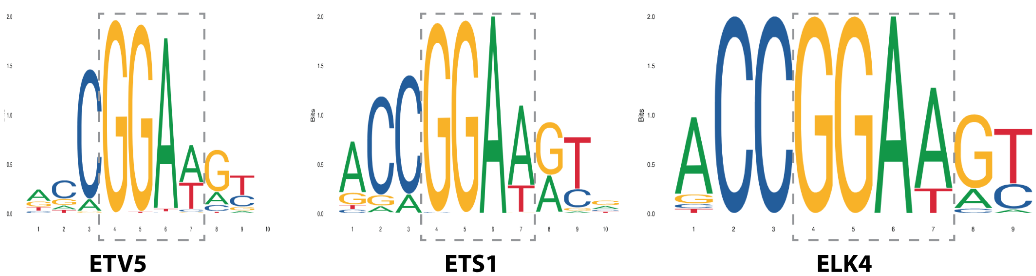

Figure 2: *Sequence logos of three ETS factors: ETV5, ETS1 and ELK4. The ETS domain, which is represented by a GGAA or GGAT sequence, is denoted by the dashed grey box around each plot.*

ETS1 is the founding member of the E26 family. It is expressed in different cell types and is reported to play a number of different roles in both physiological and pathological conditions (Adler & Wernert, 2012). However, this TF has been implicated in a number of different cancers, particularly through its overexpression within breast cancer (particularly the triple-negative subtype) (G.-C. Kim et al., 2018), head/neck squamous cell carcinoma (particularly within the mesenchymal subtype) (Gluck et al., 2019), and prostate cancer (Adler & Wernert, 2012).

ELK4 has been shown to promote tumorigenesis and tumor progression, particularly within prostate cancer (Makkonen et al., 2008). Studies have observed that this TF is overexpressed in prostate cancer samples (S. Edwards et al., 2005) and is frequently involved in gene fusions, particularly with the *SLC45A3* gene. This fusion is correlated with higher levels of ELK4 expression and has been shown to promote prostate cancer cell proliferation, indicating its role in tumor growth (Y. Zhang et al., 2012)(Rickman et al., 2009). It has additionally been found that in response to stimulation by a growth factor, ELK4 activates the expression of oncogenes such as *EGR1* and *FOS* (Zhu et al., 2023). However, though the research into ELK4 is increasing, the role of ELK4 is not fully known as it is dependent on the cellular context (Zhu et al., 2023).

ETS variant transcription factor 5 (ETV5) has an important role in cell development, differentiation, proliferation and apoptosis (Sementchenko & Watson, 2000). Particularly within ovarian cancer, the significant upregulation of this TF leads to the transcriptional increase of other oncogenes involved in the resistance of programmed cell death (i.e. cell apoptosis), formation of new blood vessels (i.e. angiogenesis), migration and invasion (L. Zhang et al., 2021) (Bullock et al., 2019) (Alonso-Alconada et al., 2014). ETV5 has also been implicated within colorectal cancer (X. Cheng et al., 2019) and thyroid carcinoma (Puli et al., 2018).

Though being implicated within cancer, many ETS-family TF, particularly the likes of ETS1, ELK4, and ETV5, have been used as molecular targets for clinical treatment studies for precision oncology (L. Zhang et al., 2021).

1.3.4.2 Nuclear Factor-κB (NF-κB)

Nuclear Factor kappa-light-chain-enhancer of activated B cells (NF-κB) is another well-known TF family that actively plays a large role in regulating immune and inflammatory responses, cell proliferation, and survival (Giuliani, Bucci, & Napolitano, 2018). The activation of this family, however, has been well-described in the literature, particularly within breast, lung, colorectal, and pancreatic cancers (Naugler & Karin, 2008) as well as multiple myeloma (Yu, Lin, Zhang, Zhang, & Hu, 2020), leukemia (Kordes, Krappmann, Heissmeyer, Ludwig, & Scheidereit, 2000) and lymphoma (Weniger & Küppers, 2016).

This family consists of five subunits: p50 (derived from p105; encoded by *NF-κB1* gene), p52 (derived from p100; encoded by the *NF-κB2* gene), p65 (otherwise known as RelA; encodes *RELA* gene), c-Rel (encoded by the

REL gene) and RelB (encoded by the *RELB* gene) (Zinatizadeh et al., 2021). Similar to how ETS-factors share the ETS domain, these proteins share the highly conserved Rel homology domain (RHD) (Zinatizadeh et al., 2021).

Briefly, NF-κB proteins are located in the cytoplasm where they are kept inactive by a family of inhibitory proteins called IκB (inhibitor of NF-κB) (Naugler & Karin, 2008). When cells receive signals such as pro-inflammatory cytokines (e.g. TNF-α and IL-1β), microbial infections or stress signals (Pickering & O'Connor, 2007), the IκB kinase (IKK) enzyme complex phosphorolates IκB proteins, which leads to their degradation (D'Acquisto, May, & Ghosh, 2002). As a result, NF-κB proteins are no longer retained in the cytoplasm and are able to translocate to the nucleus where they can bind to the promoter and regulate the gene expression of target genes (Naugler & Karin, 2008) (Solt & May, 2008) (D'Acquisto et al., 2002).

1.3.5 Chromatin Structure

While a TF recognizing the proper TFBS is *necessary* for binding to DNA, this alone is not *sufficient* (Pop et al., 2023). Chromatin structure, in particular, plays a large role in regulating TF access to DNA. Short segments of DNA, typically around 147 base pairs in length, wrap around histone proteins to form nucleosomes, which are the basic repetitive units of chromatin (Richmond & Davey, 2003) (Peterson & Laniel, 2004). These nucleosomes further compact with multiple nucleosomes stacking to create chromatin fibers, which then condense and intertwine to form chromosomes. Generally, nucleosomes are less abundant at certain genomic locations, particularly within regulatory elements such as promoters, resulting in accessible chromatin (Minnoye et al., 2021). This accessibility allows TFs to bind to DNA more easily. Conversely, in tightly packed heterochromatin, DNA is less accessible, making it difficult for TFs to bind.

Given the significant impact of chromatin structure on TF binding, several studies have focused on predicting chromatin accessibility both *in vitro* and *de novo*. To detect open chromatin *in vitro*, DNase I hypersensitive site sequencing (DNase-seq) and Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) are among the most commonly used chromatin accessibility profiling methods (Minnoye et al., 2021)(L. Song & Crawford, 2010).

Briefly, DNase-Seq identifies open chromatin regions by first isolating cells and extracting nuclei, which are then treated with the DNase I enzyme to cut DNA at regions where the chromatin is open and accessible. These DNA fragments are then isolated, purified, sequenced, and mapped back to the reference genome thereby identifying the locations of DNase I hypersensitive sites, which correspond to open chromatin regions (Cockerill, 2011). Similarly, ATAC-Seq provides insights into chromatin accessibility by using a transposase enzyme to insert sequencing adapters into open regions of chromatin (Buenrostro, Wu, Chang, & Greenleaf, 2015). This method is quicker and requires fewer cells than DNase-seq, making it a popular choice for studying chromatin dynamics and TF binding.

In contrast, several *de novo* approaches and tools have been developed to predict open chromatin regions based on various types of genomic data. These methods leverage machine learning models, sequence motifs, and epigenetic marks to enhance prediction accuracy. For instance, machine learning algorithms, such as DeepSEA (J. Zhou & Troyanskaya, 2015), are trained on large datasets of known open chromatin regions to learn sequence features associated with chromatin accessibility. Additionally, predictive models often incorporate known sequence motifs for TFBSs and epigenetic marks such as histone modifications, which are strong indicators of active chromatin.

Integrative approaches that combine multiple types of genomic data, such as DNA methylation, histone modifications, and chromatin conformation, provide a more comprehensive and accurate prediction of open chromatin regions. ChromHMM (Ernst & Kellis, 2012) is a tool that exemplifies this integrative strategy by utilizing hidden Markov models to segment the genome into different chromatin states based on combinations of histone modifications. This tool generates a map of chromatin states, identifying regions that are likely to be open and accessible for TF binding. ChromHMM works by using ChIP-seq data as input for various histone modifications and learning patterns that correspond to distinct chromatin states, such as active promoters, enhancers, or repressed regions (Ernst & Kellis, 2012). These states are then used to annotate the genome, providing insights into the regulatory landscape and helping to predict regions of open chromatin with high accuracy. Consequently, many studies have used this tool for various annotations (Trapnell et al., 2014) (Watanabe, Taskesen, Van Bochoven, & Posthuma, 2017) (B. Zhang et al., 2022).

1.3.6 Predicting TF Activity

TF activity refers to its ability to control the transcription of its corresponding genes. Numerous studies aim to infer TF activity levels based on changes observed in the expression levels of the TF's target genes (Badia-i Mompel et al., 2022) (Trescher & Leser, 2019).

1.3.6.1 decoupleR

A popular tool for predicting TF activities is *decoupleR*, available in both Python and R. This unified framework includes 12 computational methods to infer TF activities solely from bulk RNA-Seq data. Due to the lack of consensus on the "optimal" algorithm, prior studies have primarily used methods such as the univariate linear model (ULM), multivariate linear model (MLM), weighted sum (WSUM), VIPER (Alvarez, Giorgi, & Califano, 2014), or an aggregation of these tools, known as a consensus approach (Arriojas, Patalano, Macoska, & Zarringhalam, 2023)(Whitlock, Wilk, Howton, Clark, & Lasseigne, 2024)(Perez & Sarkies, 2023), (Hosseini-Gerami et al., 2023)(Arriojas et al., 2023) (Tudose, Bond, & Ryan, 2023)(Er-Lukowiak et al., 2023).

Briefly, ULM and MLM both fit linear models to assess TF activity for each sample. In ULM, the expression of each gene in a sample is treated as the response variable while the TF-gene weight, which represent the influence or regulatory strength of the TF on its target genes, is used as the single predictor. This approach evaluates the effect of each TF independently. In contrast, MLM fits a linear model using multiple TF-gene weights simultaneously as predictors. In both cases, the activity of a TF is represented by the t-value obtained from the fitted linear model. This t-value is used as the score indicating the activity level of the TF in this individual sample where a positive t-value represents an active TF and a negative t-value represents an inactive TF.

Conversely, the weighted sum (WSUM) approach multiplies each gene expression value by its associated TF-gene weight. The weighted target features are then summed to produce an enrichment score, which is used as the WSUM estimate. This score is then normalized to reflect the relative activity of the regulator based on the weighted contributions of its target features.

Lastly, Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) (Alvarez et al., 2014) infers TF activity by analyzing the expression patterns of the TF's target genes, collectively known as its regulon, rather than relying on the expression level of the TF itself as the previous methods do. It assesses whether these target genes are collectively upregulated or downregulated using statistical enrichment analysis to determine if the observed gene expression changes are consistent with activation or repression of the TF. VIPER then computes a "normalized enrichment score" (NES) for each TF, reflecting the inferred activity.

1.3.6.2 DoRothEA and Collectri

The *decoupleR* tool utilizes pre-defined gene regulatory networks that incorporate prior biological knowledge of TF-gene interactions. Two primary databases used for this purpose are DoRothEA (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019) and Collectri (Müller-Dott et al., 2023).

DoRothEA generated human TF-gene interactions using four strategies: manually-curated literature resources, ChIP-seq interactions, TFBS predictions on gene promoters, and transcriptional regulatory interactions inferred from expression data across normal human tissues and cancer types (Garcia-Alonso et al., 2019). This data is compiled into a table with columns for the transcription factor name (source), confidence level (A for highest, D for lowest), gene name (target), and interaction weight (ranging from -1 for repression to +1 for activation).

Collectri, an extension of DoRothEA curated by the same research lab, incorporates only high-confidence TF-gene interactions (DoRothEA confidence level A) but also includes additional experimental resources and a curated subset from their text mining method ExTRI, resulting in 43,178 TF-gene associations corresponding to 1,186 distinct TFs across humans (Müller-Dott et al., 2023).

In total, both DoRothEA and Collectri contain a large number of TF-gene interactions, categorized into activators and repressors. As of 21 July 2024, DoRothEA contains 32,275 unique TF-gene interactions (31,637 activators and 638 repressors) in comparison to Collectri's 43,178 unique interactions (37,704 activators and 5,474 repressors). In benchmarking evaluations for transcription factor activity, the DoRothEA study (Garcia-Alonso et al., 2019) utilized VIPER while the Collectri study (Müller-Dott et al., 2023) employed the ULM algorithm.

1.4 Methods to Identify Functional pSNVs

Within this section, a review of the existing literature on the most common tools used to detect functional pSNVs will be introduced. Generally, these tools can be split into two distinct categories: recurrence-based approaches and non-recurrence-based approaches. Here, the term *recurrence* refers to identifying a mutation that occurs in multiple samples though this term could further be characterized into "highly recurrent" mutations (i.e. being observed in more than three samples) and "lowly recurrent" mutations (i.e. being observed within one to three samples). In contrast, those mutations that only occur in a single sample are referred to as "singletons".

Historically, driver identification has relied on mutations being recurrent, particularly since this is thought of as a strong signal of positive selection in cancer cells (Hess et al., 2019) (Miller et al., 2015). The overarching assumption is that if each genomic position has a relatively equal propensity of being mutated, those positions that harbor a specific mutation within multiple genomes must be under some type of selective pressure. This concept is further exemplified by some of the most well-known oncogenes, such as *BRAF* (Chapman et al., 2011) (Ascierto et al., 2012) and *KRAS*, harboring highly recurrent mutations across multiple cancer types. Following in these footsteps, several methods to identify functional mutations within the non-coding region of DNA have attempted to leverage this concept of recurrence. Several of the most well-known approaches will be detailed below.

1.4.1 Recurrence-based Approaches

FunSeq2 (Fu et al., 2014) is a commonly-used computational framework that was developed to annotate and prioritize non-coding variants that may have functional significance by first creating a *data context* and then using this data context to prioritize variants within their pipeline. This data context is built by integrating six key features: functional annotations (i.e. knowledge of cancer-related genes), prediction of gain-of-function or loss-of function events, assessment of variants in conserved or non-conserved regions, linking variants with their target genes, incorporating user annotations, and importantly, identifying recurrent elements from both user-input and publicly-available cancer datasets. Once this context is built, a scoring algorithm is then applied to prioritize putative high-impact variants. Through this methodology, they cite a high predictive power for the detection of *recurrent* somatic regulatory variants.

OncoDriveFML (Mularoni, Sabarinathan, Deu-Pons, Gonzalez-Perez, & López-Bigas, 2016) also aims to identify and prioritize mutations by estimating a functional impact (FI) score for each mutation. In essence, the functional impact of all mutations within a specific region (i.e. promoter) is computed through the use of a number of different annotations (i.e. conservation scores, predicted effect on protein structure) and user-specified algorithms. Then, this impact is compared to a local background distribution in order to detect regions or exact genomic positions under positive selection, which is then used as a score for each mutation.

ncDriver (Hornshøj et al., 2018) employs a two-step approach to identify and prioritize non-coding variants. Initially, it identifies recurrent mutations (SNVs and indels), and subsequently assesses their significance based on cancer specificity and conservation. In the first step, mutations are evaluated to determine if their frequency exceeds expectations, employing a background mutation rate derived from a binomial distribution, to retain only significant mutations. In the second stage, prioritization is based on cancer specificity, local conservation, and global conservation tests.

Lochovsky et al. (Lochovsky et al., 2015) introduced LARVA, a tool designed to pinpoint highly mutated non-coding regulatory elements by detecting mutations that occur more frequently than anticipated. By contrasting the observed mutation count with a background mutation distribution modeled as a β-binomial random variable, accounting for various factors like replication timing, mutations were then ranked and prioritized.

MutSigNC (Rheinbay et al., 2017) is an additional computational method based upon the MutSig suite of tools (Lawrence et al., 2014) with a goal of identifying significantly mutated promoters, particularly within breast cancer. For each genomic element (i.e. promoter), MutSigNC compares the observed mutation rate against a patient-specific background mutation rate, which is modeled by a β-binomial distribution, in order to determine the significance of this region.

In addition to these tools, other recurrence-based studies include those of Weinhold et al. (Weinhold et al., 2014), Nik-Zainal et al. (Nik-Zainal et al., 2016), and Melton et al. (Melton, Reuter, Spacek, & Snyder, 2015).

1.4.1.1 Strategy of Recurrence-Based Methods

In summary, recurrence-based methods, such as FunSeq2 (Fu et al., 2014), OncodriveFML (Mularoni et al., 2016), ncDriver (Hornshøj et al., 2018), LARVA (Lochovsky et al., 2015), and MutSigNC (Rheinbay et al., 2017), all have a single guide for detecting functional mutations: if a single base pair or region of base pairs is mutated more frequently than randomly expected, the mutation must have been positively selected during tumor progression (Elliott & Larsson, 2021) (Martínez-Jiménez et al., 2020).

This methodology could be broken down into four distinct steps:

1. Calculate the observed mutation rate within a dataset for their regions of interest (i.e. promoters).

2. Construct a statistical background model that represents the expected mutation rate, which is typically done via a binomial, β-binomial, or Poisson distribution.

3. Compare the two distributions of the observed mutation rate to the expected mutation rate.

4. If there is a significant difference between these two distributions, then it could be hypothesized that this position (or region) is positively selected during tumor progression.

However, determining what should be expected is predicated on having a reliable background distribution to serve as a null model, which has been shown to be difficult to estimate (Zhao, Martin, & Gordân, 2022)(Lawrence et al., 2013). Furthermore, this background model must incorporate the mutational heterogeneity of different tumor and tissue types as this has a large effect on driver detection (Lawrence et al., 2013). If the underlying background model is misspecified, calculations and discoveries of functional mutations may not necessarily be correct.

Furthermore, this recurrence-based strategy of detecting functional pSNVs requires high statistical power to produce reliable results. By definition, statistical power is the probability of rejecting a null hypothesis given that it is truly false, which, in the context of functional pSNV discovery, is the probability of correctly identifying a true, functional pSNV. If high statistical power is not achieved, however, this implies that methods are unable to discern the difference between truly functional mutations and passenger mutations (E. Kim et al., 2016).

Consequently, one of the keys of obtaining high statistical power is to increase the sample size in order to have enough samples harboring the event of interest. The more samples harboring this event (i.e. higher recurrence level) leads to higher statistical power, which makes the detection of these functional pSNVs possible. Therefore, to reliably detect functional pSNVs, these recurrence-based methods not only depend on a correctly-specified background mutation rate but also inherently depend on having a high recurrence rate.

1.4.2 Singletons and Lowly-Recurrent Events

Though having high statistical power through a high recurrence rate is desirable, the majority of functional events occur at low frequencies (E. Kim et al., 2016) (Garraway & Lander, 2013) (Rheinbay et al., 2020). This makes recurrence-based methods statistically underpowered, leaving a significant blindspot in the identification of functional singletons and lowly-recurrent mutations.

Rheinbay et al. (Rheinbay et al., 2020) illustrated this effect by comparing the number of samples harboring a specific pSNV needed to achieve adequate statistical power across different sample sizes (Figure 3). In a pan-cancer dataset of 2,278 total samples, it was reported that at least 15 samples need to harbor a specific pSNV in order achieve a statistical power of 90%, a threshold commonly used in pSNV detection. In smaller cohorts such as bladder transitional cell carcinoma (n=23), 6 (26%) of those samples would need to harbor the same pSNV to reach the same power threshold. The smallest required recurrence level to achieve sufficient power was 4, which was calculated for the myeloproliferative neoplasms (Myeloid-MPN; n=23) cohort, the general myeloid cohort (Myeloid; n=38), and the central nervous system pilocytic astrocytomas (CNS-PiloAstro; n=89).

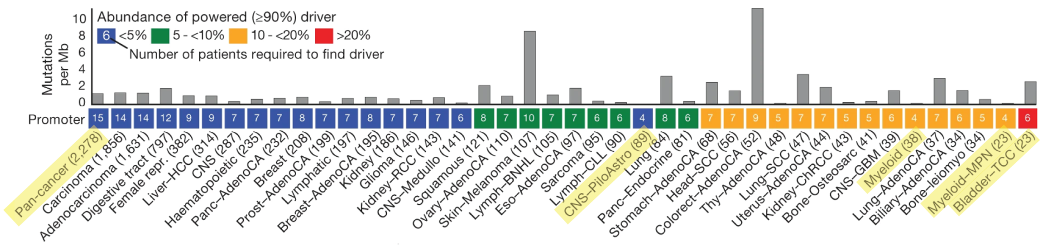

Figure 3: *A heatmap showing the minimal frequency needed for cohorts to obtain at least 90% statistical power. For each cohort, the mutational rate is displayed as a bar graph whereas the number within each colored box represents the number of patients required to find a driver. The pan-cancer dataset, along with the CNS-PiloAstro, Myeloid, Myeloid-MPN, and Bladder-TCC cohorts, are highlighted in yellow. This figure is a modified version of Figure 4A within Rheinbay et al. (Rheinbay et al., 2020), adapted and reprinted under the Creative Commons Attribution 4.0 International License.*

Given that most mutations are not recurrent enough to meet even this lenient threshold of four, detecting singletons and lowly-recurrent mutations, particularly pSNVs, using a recurrence-based approach is unattainable.

This would not be an issue if these types of mutations were inconsequential. However, several studies have implicated non-recurrent and lowly-recurrent mutations as having oncogenic potential (Scholl & Fröhling, 2019) (Zhao et al., 2022) (E. Kim et al., 2016) (Ostroverkhova et al., 2023) (Nussinov, Tsai, & Jang, 2019). Consequentially, as these mutation types have been missed and lie in the "long tail of infrequent molecular alterations" (Scholl & Fröhling, 2019), this implies that, from a clinical perspective, many potentially actionable mutations are yet to be discovered.

To address the limitations of recurrence-based approaches, only a few studies have explored non-recurrence-based methods to identify functional pSNVs (Zhao et al., 2022). However, these studies typically rely solely on the binding of specific transcription factors (TFs) or TF families and often lack experimental validation to support their computational findings. Consequently, non-recurrence-based methods are not only preferred but necessary to discover a broader range of functional pSNVs, particularly those that are singletons or lowly-recurrent.

1.5 *In vitro* validation via luciferase assays

Luciferase reporter assays are commonly used in biology to measure a promoter's ability to drive the expression of a reporter gene (Yin, Xiang, & Li, 2005)(Horn et al., 2013)(F. W. Huang et al., 2013). In a dual reporter luciferase assay system, the promoter region of interest, which may or may not contain the SNV of interest, is cloned upstream of both the Renilla luciferase gene and the firefly luciferase gene in a circular plasmid. These plasmids are then transfected into the target cells, which are allowed to grow for a sufficient amount of time for luciferase expression. A substrate, such as luciferin, is added to detect and measure luciferase activity, which is measured in relative light units (RLU). The firefly luciferase activity reflects the experimental condition while the Renilla luciferase activity serves as an internal control for normalization, ensuring that variations in transfection efficiency and cell viability are accounted for (McNabb, Reed, & Marciniak, 2005).

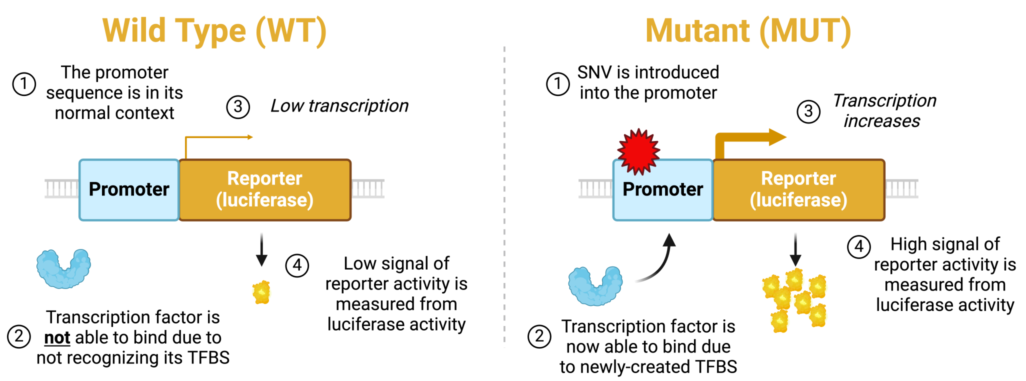

Figure 4: *Diagram of a luciferase reporter assay system to detect the effect of a pSNV. In the wild type configuration (left), the promoter exists in its native sequence context without the presence of the SNV. Consequently, a specific TF is unable to bind to the promoter leading to low transcription of the reporter and low luciferase signal. Conversely, in the mutant configuration (right), the pSNV is now introduced into the promoter, which now creates a TFBS allowing a TF to bind. After binding, the TF increases the transcription of the reporter gene and this high signal is measured. Created with* `BioRender.com`.

By measuring the normalized luciferase activity for both the wild type (WT) promoter sequence (i.e. normal sequence) and the mutant (MUT) promoter sequence (i.e. pSNV introduced into sequence), it is then possible to observe the effect that the pSNV has *in vitro*. An example of a pSNV leading to the creation of a new TFBS thereby recruiting a new TF and increasing transcription can be seen in Figure 4.

1.5.1 Known Functional Promoter Mutations

Although there exists a significant blindspot for functional pSNVs within the literature, there are, however, several well-known pSNVs, which will be described within this section. In particular, I will initially describe the two well-known *TERT* hotspot mutations, followed by pSNVs within *CDC20*, *RALY* and *LEPROTL1*.

1.5.1.1 $TERT_{C228T}$ and $TERT_{C250T}$

Independently, Huang et al. (F. W. Huang et al., 2013) and Horn et al. (Horn et al., 2013) were the first to describe the oncogenic potential of pSNVs with their discoveries of the highly recurrent $TERT_{C228T}$ (**C→T** mutation at chr5:1,295,**228**) and $TERT_{C250T}$ (**C→T** mutation at chr5:1,295,**250**) mutations. These mutations occurred within the core promoter of telomerase reverse transcriptase (*TERT*), which is a catalytic subunit of telomerase. In order to increase *TERT* transcription and therefore promote telomerase activation, both of these point mutations create new TFBSs for the E26-family TF GABPA, which acts as a "master regulator" of *TERT* transcription (Yuan, Dai, & Xu, 2020).

Furthermore, both seminal studies found that $TERT_{C228T}$ and $TERT_{C250T}$ were highly recurrent in their respective studies: Horn et al. (Horn et al., 2013) found recurrence rates of 46 (27.4%) and 64 (38.1%) of $TERT_{C228T}$ and $TERT_{C250T}$ pSNVs, respectively, whereas Huang et al. (F. W. Huang et al., 2013) found a recurrence rate of 27 (39%) and 23 (33%), respectively. Consequentially, these two pSNVs are commonly referred to as *hotspot* pSNVs.

To functionally validate their findings *in vitro*, Huang et al. used a luciferase reporter assay to compare the activities between the WT construct and the two MUT constructs (i.e. one construct with $TERT_{C228T}$ and the other with $TERT_{C250T}$) independently. These mutations resulted in a 100% to 300% increase in activity over the WT, depending on the cell lines used and were all considered to be statistically significant (p-value < 0.05). Since their initial discovery, nearly all functional pSNV detection methods (e.g. FunSeq2 (Fu et al., 2014), OncodriveFML (Mularoni et al., 2016), ncDriver (Hornshøj et al., 2018)) use the detection of $TERT_{C228T}$ and $TERT_{C250T}$ as a minimum requirement and common baseline.

1.5.1.2 $CDC20_{G529A}$

Prior studies have shown that Cell Division Cycle 20 (*CDC20*) is overexpressed in various human cancers including that of colorectal cancer and ovarian cancer (Xi et al., 2022) (Wu et al., 2013) (S. Cheng, Castillo, & Sliva, 2019). However, Godoy et al. (He et al., 2021) further elucidated how four distinct mutations, namely $CDC20_{G529A}$, $CDC20_{G52AA}$, $CDC20_{G525A}$, and the dinculeotide variant $CDC20_{GG528/9AA}$, all have a functional consequence. Using a recurrence- and window-based approach that defines a background model to determine frequently-mutated regions, He et al. found that at least one of these *CDC20* pSNVs were present in 26 samples using a 7 bp window (chr1:43824522-43824532), though this slightly differs from the classical definition of recurrence of one single base pair.

Furthermore, the hotspot mutations in *CDC20* were predicted to affect the binding site for several E26 transformation-specific (ETS) family transcription factors, similar to that of both previously reported *TERT* pSNVs. Using a short hairpin RNA (shRNA) to knock down the expression of 6 identified ETS-family TFs, He et al. observed that ELK4 was the only one that resulted in a signficant up-regulation. Furthermore, they used the publicly-available ENCODE (Consortium et al., 2012) dataset to verify the binding of ELK4 to the *CDC20* promoter.

Using the melanoma cell line M14 as well as the kidney cell line HEK293FT for the *in vitro* validation of $CDC20_{G529A}$ via luciferase reporter assays, a slight, statistically-significant upregulation of about 50% in both cell lines was observed. It should be noted, however, that the specific details (i.e. number of replicates, specific mean value, specific p-value) of the experiments were not described in the text or the supplement and this approximation of their mean upregulation value was gathered from one of their figures.

Godoy et al. (Godoy et al., 2023) similarly identified $CDC20_{G529A}$ through FunSeq2, which is a recurrent-based approach detailed in a future section (See Section 1.4). However, upon also using the same cell line (HEK293FT) as He et al, they observed that this mutation actually *decreases* the transcriptional activity in comparison to the wildtype construct.

1.5.1.3 ***LEPROTL1***$_{C921T}$

Through the use of the previously-described tool *ncDriver* (see Section 1.4.1), Rheinbay et al. (Rheinbay et al., 2017) analyzed 360 primary tumor breast cancer samples and identified 9 specific elements as having a statistically-significant increase in mutations. Of these 9 elements, one in particular was within the promoter of Leptin Receptor Overlapping Transcript Like 1 (*LEPROTL1*), whose overexpression has previously been observed within endometrial cancer (Boroń, Nowakowski, Grabarek, Zmarzły, & Opławski, 2021).

In particular, the *LEPROTL1*$_{C921T}$ pSNV was identified due to its high recurrence with 5 samples harboring this specific mutation in their dataset. Consequently, this pSNV was validated *in vitro* via a luciferase assay within the HEK293FT kidney cell line. Their validation efforts, however, observed *LEPROTL1*$_{C921T}$ as being functional - though in a negative direction (i.e. mutant promoter activity is less than the wild type activity) by observing approximately a 50% decrease in promoter activity when comparing it to the WT activity.

In addition to the high recurrence status of *LEPROTL1*$_{C921T}$ within this particular dataset, additional studies have identified this pSNV as being recurrent in 33 of 302 urothelial bladder cancer samples though no functional validation was conducted (Jeeta et al., 2019).

1.5.1.4 ***RALY***$_{C927T}$

In addition to *TERT*$_{C228T}$, *TERT*$_{C250T}$, *CDC20*$_{G529A}$ and *LEPROTL1*$_{C921T}$, other pSNVs have been nominated as being functional but rarely do these non-coding studies validate these results *in vitro* through a luciferase reporter assay or other system.

A notable example of this is of Hayward et al. (Hayward et al., 2017) when, particularly through the analysis of melanoma samples, multiple non-coding driver mutations were proposed within the promoters of genes such as Heterogeneous Nuclear Ribonucleoprotein (*RALY*). This gene, which encodes the RNA-binding protein Raly, is particularly interesting due to its overexpression being previously observed in ovarian, lung, bladder, brain, and breast cancers in addition to multiple myelomas and melanomas (Tsofack et al., 2011). This overexpression has led to poorer survival rate and has been known to promote the invasiveness of cancer cells (Bondy-Chorney et al., 2017). Though the $RALY_{C927T}$ pSNV was been implicated by this study, no experimental studies were conducted to validate these computation results *in vitro*.

1.5.2 Melanoma and Mutational Signatures

Within their respective original studies, $TERT_{C228T}$, $TERT_{C250T}$, $CDC20_{G529A}$ and $RALY_{C927T}$ were all identified within melanoma samples, which raises the question as to why this is. One possible hypothesis is that there are certain mutational processes that only act within certain cancer types, melanoma being one of them.

Melanoma is a malignant tumor that arises from melanocytes, which are cells that produce the melanin pigment found in skin, eyes, ears and gastrointestinal tract among others (Long, Swetter, Menzies, Gershenwald, & Scolyer, 2023). Melanoma is one of the fastest growing cancer types worldwide with significantly different incidence rates across the world (Eggermont, Spatz, & Robert, 2014). For example, there are approximately 25 new melanoma cases per 100,000 people within the United States of America yearly, which is slightly less than the 30 new cases per 100,000 per year within Europe. On the other hand, both Australia and New Zealand have roughly double those amounts with approximately 60 new cases per 100,000 (Long et al., 2023).

These high incidence rates in warm countries may be partly due to melanoma being primarily attributed to ultraviolet (UV) radiation exposure (Long et al., 2023) (Hayward et al., 2017) (Eggermont et al., 2014). UV radiation exposure mainly affects sun-exposed parts of the body, such as the skin. Consequently, one subtype of melanoma, which is related to the skin, is referred to as cutaneous melanoma. The other two distinct melanoma subtypes are acral melanoma (occurring on the palms of the hands, nails, and soles of the feet) and mucosal melanoma (occurring within the mucosal epithelium) (Rabbie, Ferguson, Molina-Aguilar, Adams, & Robles-Espinoza, 2019).

Though there are distinct mutational differences between these three subtypes (i.e. cutaneous melanoma has an 18-fold higher increase in SNV frequency than mucosal and acral combined (Hayward et al., 2017)), melanoma, as a whole, is still the most frequently mutated cancer (Hayward et al., 2017). Previous studies have shown that there is a direct relationship between UV exposure and melanoma, particularly by enriching the amount of

C>T/G>A mutations at certain locations (Hodis et al., 2012) (Eggermont et al., 2014) (Hayward et al., 2017). However, it has been shown that of those rare melanoma cases that do not have this specific mutation, these lead to fewer point mutations (Hayward et al., 2017).

The concept of mutational signatures was first introduced in 2013 by Alexandrov et al. (Alexandrov et al., 2013), aiming to identify common mutational patterns within the genome. Their analysis encompassed approximately 7,000 genomes comprising 4,938,362 SNVs and indels combined across 30 different cancer types. At the time, they identified 21 distinct mutational signatures, particularly focusing on single nucleotide variants (SNVs) or single base substitutions (SBS). Several SBS signatures are associated with specific factors such as SBS1 with age, SBS4 with smoking, and SBS7 with ultraviolet light exposure, though many factors remain to be elucidated (Alexandrov et al., 2013).

Today, there are now 99 SBS mutational signatures according to the Catalog of Somatic Mutations in Cancer (COSMIC) database (Tate et al., 2019). These signatures are typically defined by their dinucleotide or trinucleotide context and are denoted only in the 5' to 3' direction. Additionally, this database has further expanded to include patterns within dinucleotide variants (doublet base substitutions: DBS), indels (small insertions and deletions; ID), copy number variations (CNVs), and structural variants (changes exceeding 1 kb in length/SV) (Tate et al., 2019).

The aforementioned C>T/G>A mutation is a major component of the mutational signature Single Base Substitution 7 (SBS7), which has three distinct "subsignatures" that are all associated with UV radiation exposure. SBS7a is typically characterized as a C>T mutation in a TpC dinculeotide context (i.e. TC>TT / GA>AA) whereas SBS7b is typically characterized as a C>T mutation but at a CpC dinucleotide context (i.e. CC>CT or CC>TC / GG>AG or GG>AG), both of which incorporate C>T/G>A mutation. Lastly, SBS7c is characterized as either a T>C or T>A mutation in any context.

Specific characterizations according to the COSMIC website of these three mutational signatures can be seen in Figure

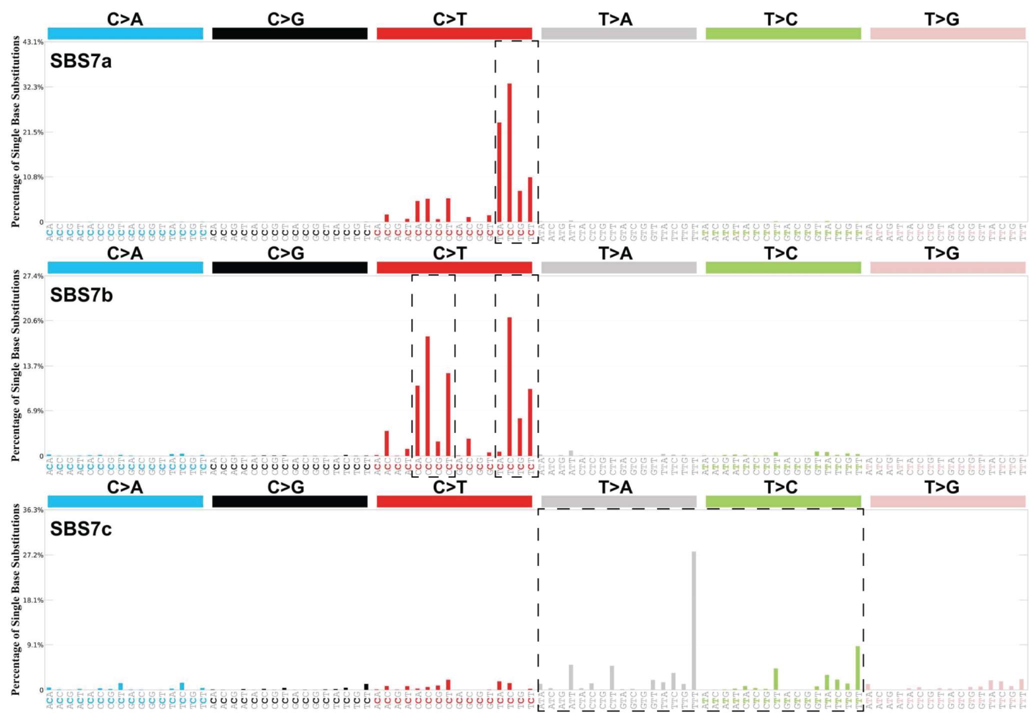

Figure 5: *Mutational signatures of SBS7 as defined by COSMIC (Alexandrov et al., 2013). Each original mutational signature was downloaded from* *https://cancer.sanger.ac.uk/signatures/sbs*

In summary, COSMIC employed 4,938,362 somatic substitutions sourced from 30 diverse cancer types to generate the plots. Through the utilization of their tools *SigProfiler* and *SigProfilerExtractor* for mutational signature extraction, 21 distinct mutational signatures were revealed, each associated with unique samples and validated through lab results across different tiers. Detailed information can be seen within the supplement of Alexandrov et al. (Alexandrov et al., 2013).

1.6 Statistical Testing

Of those pSNVs described in Section 1.5.1, the WT activity and MUT activity were compared using a student t-test, which assumes that both activities follow a Gaussian distribution.

A Gaussian or Normal distribution is a fundamental statistical distribution characterized by its mean μ and standard deviation σ. It describes a symmetric bell-shaped curve where data points are clustered around the mean, tapering off the further it is away from the center.

The student t-test, commonly known as the t-test, is a parametric statistical test used to determine if there is a significant difference between the means of two independent groups that are assumed to be normally distributed. It is particularly suitable when sample sizes are small and the population standard deviation is unknown. The t-test can be performed in two variations: one-sided and two-sided tests. In both cases, the null hypothesis H_0 suggests no difference between the means of the group while the alternative hypothesis H_1 varies. In a two-sided t-test, H_1 indicates a statistical difference in *either* direction of the distribution while a one-sided t-test specifies a difference in a *singular* direction (either positive or negative).

In hypothesis testing, a p-value quantifies the probability of observing the test statistic if the null hypothesis were to be true. Typically, a p-value less than or equal to 0.05 indicates strong evidence against the null hypothesis, suggesting a significant difference between the group means while a large p-value suggests insufficient evidence to reject the null hypothesis.

In the context of measuring the effect of a pSNV with a luciferase reporter assay, t-tests have commonly been used (F. W. Huang et al., 2013) (Horn et al., 2013) (He et al., 2021) (Godoy et al., 2023) to test whether there is a statistical difference between the wild type and promoter activity. Since these two groups are independent, it is reasonable to assume that the 'true' luciferase activity for each group corresponds to a single value from which the observed data is drawn from. Minor deviations from this point would then be expected to mirror a normal distribution.

If the aim is to detect only a specific increase in mutant activity compared to the wild type, conducting a one-sided t-test would be suitable. However, if any change, whether increase or decrease, is of interest, a two-sided t-test would be preferable. In both scenarios, rejecting the null hypothesis with a significant p-value would indicate a statistically significant difference between the two groups.

1.7 Next-Generation Sequencing (NGS) Techniques

The cost of sequencing a single genome has reduced nearly 50,000 fold (Goodwin, McPherson, & McCombie, 2016) as of 2016. Moreover, data from the National Human Genome Research Institute (NHGRI) indicates an even greater reduction from 95,263,072 US Dollars to $525 from September 2001 to May 2022 (Wetterstrand, 2019).

This reduction in price has been primarily attributed to the rise of next generation sequencing (NGS) technologies, which has also vastly increased data output. In contrast to traditional Sanger Sequencing (Valencia et al., 2013), which was at the forefront of sequencing methods for three decades since its conception in 1977 and only allows for the sequencing of a single genome at a time, NGS platforms simultaneously sequence millions of DNA fragments at various lengths (i.e. short read sequencing or long read sequencing).

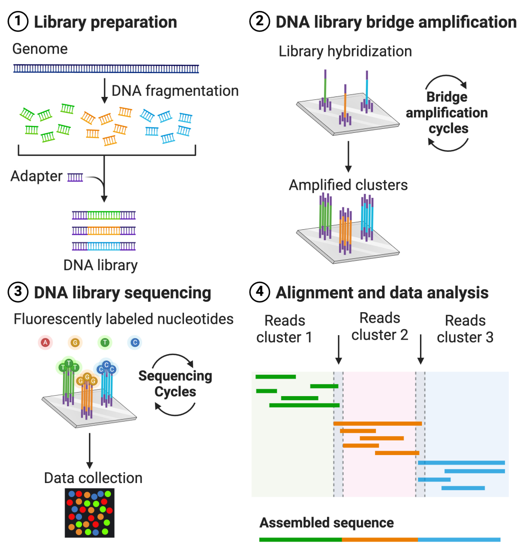

Figure 6: *The four major steps of using the Illumina HiSeq Platform for whole genome sequencing. Image adapted from Ona, S. (2020) on* `BioRender.com`.

Short read sequencing methods include that of Illumina Sequencing, Ion Torrent Sequencing, and Pyrosequencing whereas long read sequencing includes PacBio Sequencing and Nanopore Sequencing. In particular, Illumina's HiSeq (Caporaso et al., 2012) platform has been used within many research institutes and core facilities, such as that within the German Cancer Research Center (DKFZ), Harvard Medical School and Johns Hopkins.

To describe how a sample is sequenced using Illumina's WGS approach, the process begins with fragmenting the DNA sample into smaller pieces, approximately 500 base pairs long. Adapters are then attached to both the 5' and 3' ends of these fragments. Next, the DNA fragments undergo bridge amplification where they are attached to a solid surface and amplified to form clusters of identical DNA fragments. Subsequently, sequencing-by-synthesis is performed, during which nucleotides labeled with distinct fluorescent markers (one color for A, another for G, etc.) are incorporated into the DNA clusters. The emitted fluorescent signals are captured to determine the nucleotide sequence of each fragment. Finally, the sequenced reads are aligned to a reference genome using algorithms such as Burrows-Wheeler Aligner Maximal Exact Matches (BWA-MEM) to reconstruct the original DNA sequence. Once the alignment is complete, the reconstructed genome is typically annotated with various features, including genes, transcripts, and regulatory elements. This process can be seen in Figure 6 (L. Liu et al., 2011).

Conversely, Illumina's HiSeq Platform can also be used for RNA-Seq. This involves several key steps, which are similar to its WGS approach. Initially, RNA is extracted from the sample and enriched for mRNA, which is then fragmented into smaller pieces. Reverse transcription is performed to generate complementary DNA (cDNA) and thus attach adapters to their ends, enabling them to bind to the flow cell and be amplified during sequencing. The cDNA fragments are amplified using PCR to create a cDNA library. Next, the cDNA library is loaded onto a flow cell where the cDNA fragments bind to complementary adapter sequences on the flow cell surface. Through bridge amplification, clusters of identical cDNA fragments are formed. The flow cells then undergo sequencing-by-synthesis where fluorescently-labeled nucleotides are incorporated into the cDNA strands. Images of the fluorescence emitted by each incorporated nucleotide is then used to determine the sequence of bases in each cDNA fragment. Finally, that raw sequencing data is processed to generate high-quality reads (Kumar et al., 2012).

1.8 Precision Oncology

Traditional oncology treatments have relied on aggregate-based decisions derived from population averages obtained from randomized clinical trials involving selected patient groups (Fountzilas, Tsimberidou, Vo, & Kurzrock, 2022). This approach has been the foundation of drug approvals for decades (Warner et al., 2020). However, given the heterogeneous nature of cancer, applying a single treatment strategy to all patients with the same type of cancer can significantly limit treatment effectiveness (Fountzilas et al., 2022) (Wahida et al., 2023).

To address this issue, precision oncology, also known as personalized oncology or targeted gene therapy, considers the individual patient's genomic biomarkers and predispositions when considering treatment options

(Fountzilas et al., 2022). The significant rise of precision oncology can be largely attributed to advancements in NGS technologies (Weymann, Pataky, & Regier, 2018), which lowers the cost of sequencing an individual genome. Consequently, individualized genetic information allows for the development of personalized treatment plans tailored to an individual's cancer. This approach improves treatment efficacy, minimizes side effects, guides treatment decisions and is becoming more cost-effective and accessible as time goes on.

To exemplify its utility, precision oncology has successfully identified specific subtypes of cancers that can be treated with targeted therapies. For instance, human epidermal growth factor receptor 2 (HER2)-positive breast cancer can be effectively treated with targeted treatments (Swain et al., 2015), as can *BRAF* mutant melanoma (Chapman et al., 2011), and lung cancers with alterations in the epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK), or ROS1 (C. Zhou et al., 2011)(Shaw et al., 2013).

However, analysis of non-coding elements has yet to yield many druggable targets with the exception of the long non-coding RNA MALAT1 (Amodio et al., 2018). While promoter mutations, most notably in *TERT*, have been cited as potential therapeutic targets (Yang et al., 2021)(J. Chen et al., 2021), treatments have yet to be developed though this has been partially attributed to the paucity of known non-coding drivers (Rheinbay et al., 2020).

To enable a more individualized approach, many hospitals are now incorporating these methods into their clinical trials. For instance, the National Cancer Institute (NCI) sponsored the Molecular Analysis for Therapy Choice trial (NCI-MATCH; Clinical Trials Identifier: NCT02465060), which was a precision cancer trial from 2015 to 2023, designed to match patients with advanced solid tumors, lymphoma, or myeloma to appropriate therapies (O'Dwyer et al., 2023)(Flaherty et al., 2020). NCI-MATCH enrolled roughly 6,0000 individuals whose cancers had progressed on standard treatments or who had rare cancers with no standard treatment options. 1,567 of these patients were assigned to separate substudies in which seven of these substudies showed a positive result (O'Dwyer et al., 2023).

Another significant clinical trial is the Profile-Related Evidence Determining Individualized Cancer Therapy (I-PREDICT; Clinical Trials Identifier: NCT02534675) at the University of California, San Diego Moores Cancer Center and Avera Cancer Institute. In this study, 149 patients with metastatic cancer were enrolled. Genetic profiling of 236 to 405 genes was performed on these patients with 83 receiving treatment and being considered for analysis.

Of these, 73 patients were administered personalized therapy. Although the trial is still in progress, a key takeaway is the value of NGS technology and the need to implement personalized precision medicine approaches earlier in the treatment process (Sicklick et al., 2019).

In addition to these two trials, many other clinical trials, such as the NCT-MASTER program (Horak et al., 2017), COGNITION (Pixberg et al., 2022), and INFORM (Van Tilburg et al., 2021), have been created to match patient-specific and tumor-specific mutations with targeted therapies.

1.9 Reproducibility, Interpretability and Computational Efficiency

With the continuous increase of available biological data and the development of bioinformatics tools to analyze it, three major challenges have emerged within the filed of bioninformatics that must shape how new tools are designed: reproducibility, interpretability and computational efficiency.

1.9.1 Reproducibility

Reproducibility in bioinformatics and in science as a whole is missing (Baker, 2016). It is estimated that only approximately 5.9% to 26% of code stemming from bioinformatics studies can be re-ran to get the originally reported results (Ioannidis et al., 2009) (Samuel & Mietchen, 2024) (Trisovic, Lau, Pasquier, & Crosas, 2022), which highlights reproducibility as one of the most significant challenges facing the field.

Reproducibility can be broken down into three distinct categories: code, data, and results. Code reproducibility refers to other researchers (i.e. bioinformaticians or clinicians) being able to have an up-to-date, working version of a tool's codebase. Data reproducibility refers to the original input data of the tool being documented and verified. Result reproducibility refers to the that given the code and original input data, the same exact results can be reproduced.

Code reproducibility involves having clear documentation, version control, and dependency management systems to facilitate the implementation and execution of the code in other working environments (i.e. different institutions, operating systems, etc.). As the results of computational experiments can be highly sensitive to changes in software versions, parameter settings, and reference annotations, ensuring consistency in these aspects is crucial (Wratten, Wilm, & Göke, 2021). Repositories such as GitHub or GitLab for code sharing and collaboration help ensure that the code is well-commented and includes instructions for setup and use.

Data reproducibility involves documenting and verifying the original input data used by the tool. This requires detailed metadata, data provenance information (de Paula, Holanda, Gomes, Lifschitz, & Walter, 2013) and ensuring that the datasets are accessible and maintained in a stable manner. Additionally, the versioning of datasets is essential, particularly when running pipelines, in order to trace original data back to its source and re-perform any analysis if necessary.

Lastly, *result* reproducibility means that given the code and the original input data, the same exact results can be consistently reproduced. This involves rigorous testing, validation, and often the use of standardized computational environments to minimize discrepancies.

1.9.2 Interpretability

In addition to reproducibility, bioinformatics tools must also be interpretable not only for bioinformaticians but also for clinicians who make patient care decisions, such as those within molecular tumor boards (MTBs). From a clinical perspective, the widespread adoption of a tool primarily depends on being able to accurately assess how a result was made (Couckuyt et al., 2022). Guidelines, such as those proposed by SPIRIT (Chan et al., 2013), CONSORT-AI (X. Liu et al., 2020) and MI-CLAIM (Norgeot et al., 2020), have been proposed to support the use of bioinformatic analyses in clinical trials, emphasizing that interpretability and transparency must be considered from the outset of any project. Furthermore, interpretable bioinformatics tools facilitate better communication between interdisciplinary teams. Clinicians, bioinformaticians, and other healthcare professionals can collaboratively evaluate and refine these tools, leading to more accurate and personalized patient care.

1.9.3 Computational Efficiency

As the amount of generated biological data increases, the need for tools and algorithms to process this data in a timely manner is not only preferred but becoming increasingly more important year by year. Particularly within an MTB setting where the turnaround time from biopsy to therapeutic decision is short (e.g. 6 weeks), any time saved in tool runtime or data aggregation can be extremely important, which could allow healthcare professionals focus more on result interpretation and potential therapeutic options. Consequently, several studies have cited computational efficiency as being of one of the biggest challenges of today (Hanussek, Bartusch, & Krüger, 2021) (Fernald, Capriotti, Daneshjou, Karczewski, & Altman, 2011) (Kleftogiannis, Kalnis, & Bajic, 2016).

1.10 Datasets

1.10.1 Pan-Cancer Analysis of Whole Genomes (PCAWG)

Following the completion of the Human Genome Project (Collins & Fink, 1995), The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) was established in 2005 with the primary goal of sequencing whole cancer genomes. TCGA collected and analyzed sequencing data from over 11,000 cancer samples, primarily focusing on exome sequencing, within 33 different cancer types. Although the main focus was on variant analysis, other topics such as gene expression profiling and copy number alterations were also analyzed. A parallel effort by many of the same TCGA members resulted in the formation of the International Cancer Genome Consortium (ICGC) (ICGC Consortium, 2010), further expanding global collaboration in cancer genomics research. The publicly-available data from these projects formed the foundation for numerous cancer genomics studies (Martínez-Jiménez et al., 2020) (Hoadley et al., 2018) (Sanchez-Vega et al., 2018) (Alexandrov et al., 2013).

To advance beyond the TCGA and ICGC datasets, the Pan-cancer Analysis of Whole Genomes (PCAWG) consortium was formed (PCAWG Consortium, 2020). PCAWG aimed to improve genomic analysis by examining the entire genome, including the non-coding regions of DNA, which has been largely overlooked. This was achieved by WGS data rather than focusing solely on the exome and offered three significant advantages over its predecessors.

Firstly, the PCAWG dataset comprised mainly of WGS data rather than WES data, providing a more holistic view of the genomic landscape of cancer. Through this data, specialized working groups focused on specific aspects of cancer genomics. Notably, these investigations explored non-coding somatic drivers which helped elucidate the role of non-coding regions in cancer progression (Rheinbay et al., 2020), characterized mutational signatures across various resolutions (i.e. single/doublet-base substitutions and small-insertions-and-deletions) (Alexandrov et al., 2020) and attempted to identify both coding and non-coding drivers in cancer using computational tools like DriverPower (Shuai, Gallinger, & Stein, 2020).

Secondly, WGS tumor samples within the PCAWG dataset were only included if they were of high quality according to the PCAWG Consortium's Technical Working Group (PCAWG Consortium, 2020). Initially, WGS data from 2,834 donors were collected. After quality assurance measures, 176 samples were excluded and an additional 75 contained minor issues and were categorized as "grey-listed" samples (PCAWG Consortium, 2020). The remaining 2,583 were labeled as "white listed" or "included" and were deemed to be of optimal quality.

Lastly, all grey and white listed samples underwent reprocessing using a standardized computational workflow to ensure reliability and mitigate biases (PCAWG Consortium, 2020). The WGS data from different data centers was reprocessed via Illumina HiSeq platforms, yielding paired-end sequencing reads with an average coverage of 38 reads for tumor samples and approximately 60 reads for control samples (PCAWG Consortium, 2020). Furthermore, tumor and normal samples were aligned to human genome build 19 (hg19) using the BWA-MEM algorithm. To detect somatic SNVs, three distinct pipelines (EMBL and DKFZ (Rimmer et al., 2014), Sanger (Jones et al., 2016) and Broad Institute (Ramos et al., 2015)) were utilized and aggregated to ensure reliability. The RNA-Seq data were also reanalyzed by the PCAWG consortium. By having access to the original BAM files from re-analysis, the gene expression, measured in fragments per kilobase of transcript per million mapped reads (FPKM), was then quantified for most patients.

Currently, the PCAWG dataset is publicly-available and could be accessed within the PCAWG Data Portal: `https://dcc.icgc.org/pcawg`.

1.10.2 NCT-MASTER

The Molecularly Aided Stratification for Tumor Eradication Research (NCT-MASTER) program is an ongoing personalized oncology program hosted by the NCT, DKFZ, and DKTK whose focus is treating young adults with late stage cancer across all histologies and patients with rare tumors (Horak et al., 2017). As a personalized oncology program, NCT-MASTER not only gathers data for research purposes but, more importantly, provides clinically-relevant diagnostics and therapeutic options for those enrolled in the trial, many of whom have been treated previously.

Since the program's inception in 2017, the NCT-MASTER Molecular Tumor Board (MTB), which consists of biologists, bioinformaticians, clinicians, and pathologists alike, meets weekly to discuss diagnoses and potential therapeutic options on a patient-by-patient basis. A turnaround time from biopsy to final decision of less than 6 weeks was reported at the time of the primary research paper in 2017 (Horak et al., 2017). Though originally focused on collecting WES data, WGS samples are now regularly collected. Similar to the PCAWG project, these WGS samples are primarily sequenced using Illumina HiSeq platforms and aligned to the human genome using the BWA-Mem algorithm. Tumor samples are paired with matching control samples taken from either blood or buffy coat. Additionally, RNA-Seq data are aligned using STAR Version 2.5.3a. Unlike the PCAWG data, which included only samples from primary tumors, the NCT-MASTER dataset comprises biopsies from both metastatic sites and primary tumors.

As this is an ongoing precision oncology program, patient data is annonymized, stored internally within the DKFZ cluster system and therefore cannot be accessed by the general public. Though each patient has a vast amount of collected data, specific cancer types/cohorts are not available.

1.10.3 COGNITION

The COGNITION (Comprehensive assessment of clinical features, genomics and further molecular markers to identify patients with early breast cancer for enrolment on marker driven trials) (Pixberg et al., 2022) trial aims to identify biomarkers in patients with early breast cancer with a high risk for relapse. During the pilot phase of this trial, which lasted from April 2019 to September 2020, 213 patients were deemed to be fit for their study (Pixberg et al., 2022).

Samples were sequenced via whole genome or whole exome sequencing approaches approaches and were annotated according to one of the four main breast cancer subtypes: (1) triple-negative breast cancer (TNBC), (2) hormone receptor-positive/human epidermal growth factor receptor 2-negative (HR+HER2-), (3) hormone receptor-positive/human epidermal growth factor receptor 2-positive (HR+HER2+), and (4) hormone receptor-negative/human epidermal growth factor receptor 2-positive (HR-HER2+). Furthermore, samples were prepared using TruSeq Nano DNA Kit (Illumina) and were sequenced on a HiSeqX or NovaSeq 6000 platform (Illumina) yielding 151 bp long reads with median 83x coverage. In terms of RNA-sequencing, libraries were prepared using the TruSeq Stranded mRNA Library Kit (Illumina) and were paired-end sequenced on an Illumina-patterned flowcell v2.5 generating 101 bp long reads. Sequencing data was further processed and analyzed using the DKFZ computational pipelines described in Section 1.10.2 (Horak et al., 2021).

1.11 Aims of the REMIND-Cancer Workflow

Within current cancer research, there is a significant blindspot in the identification of activating promoter SNVs (pSNVs), which can be at least partially attributed to current methods requiring events to be highly recurrent (i.e. occurring in a large number of samples) in order to be detected. However, considering that most mutations are either singletons (i.e. only occur in a single patient sample) or lowly-recurrent, many of these events have been systematically neglected and thus have not yet been implicated in cancer.

Due to this blindspot, the primary aim of my thesis was to establish a workflow that identifies, prioritizes and validates *in vitro* activating pSNVs. This workflow, otherwise known as the Regulatory Mutation Identification 'N' Description in Cancer (REMIND-Cancer) *workflow*, encapsulates multiple aspects, which I will discuss in detail throughout my thesis:

- The creation of the REMIND-Cancer *pipeline*, which is a computationally efficient framework used to filter, score, and rank sample-specific pSNVs based on their genomic, transcriptomic, and annotations data

- The creation of *pSNV Hunter*, which is a data aggregation and visualization dashboard used to manually investigate pSNVs that have successfully passed the REMIND-Cancer pipeline

- The selection of candidate pSNVs that I believe are most likely to be activating within subsequent *in vitro* validation experiments

Using the results of the *in vitro* validation experiments conducted by Sabine Karolus and Dr. Cindy Körner as part of the REMIND-Cancer workflow, my work highlights the importance of looking beyond highly-recurrent mutations, pariticlarly within the promoter region, in order to add to the catalog of known functional mutations.

2 Introduction to the DREAM Challenge

As a complementary project to the REMIND-Cancer workflow, I also participated in a three-month bioinformatics challenge where the primary objective was to detect gene expression in yeast using only promoter sequences through a neural network.

Consequently, the following chapter will introduce the key concepts related to this task whereas details on the specific methods used are provided in Section 3.2. Furthermore, I will then present the results of this challenge within Section 5.9 while concluding with a discussion about how these results can have future applications within the REMIND-Cancer computational pipeline in Section 6.2.

2.1 Predicting Gene Expression From DNA Sequences

The prediction of gene expression given only a DNA sequence has been an ongoing area of bioinformatics research in recent years (Agarwal & Shendure, 2020) (Kelley, 2020) (Kelley et al., 2018) (Ding, Dixit, Parker, & Wen, 2023), with recent studies specifically focusing on achieving this task using only short promoter sequences (Vaishnav et al., 2022). Particularly when using only the promoter where other distal non coding elements (e.g. enhancers, silencers) cannot be fully accounted for, this task implicitly involves identifying TFBSs within the sequence as TFs play a major role in regulating gene expression. Consequently, a majority of efforts have utilized machine learning algorithms, particularly neural networks, to address this sequence-to-expression task (Vaishnav et al., 2022) (Zrimec et al., 2022). Neural networks have been used because of their ability to model non-linear relationships, automatically detect and transform useful features, and perform non-traditional ML tasks such as image and language processing.

2.1.1 Neural Networks

Neural networks are a general class of non-linear machine learning models that were originally inspired by how the brain works where neurons (or nodes) are responsible for the flow of information across the network. These nodes are arranged in layers, initially starting with the *input layer* that receives the data. This input layer is then followed by a number of user-defined *hidden* *layers* in which data is transformed by mathematical functions (i.e. activation functions, weights, biases). Outputs from these layers are propagated forward until it reaches the final *output layer* where an output is predicted depending on the task at hand (classification or regression). This approach is often called deep learning as many hidden layers are included.

The difference between the predicted output and the true output is measured by a *loss function*, which is then used to update the weights and biases of each node to minimize the loss, typically through a process called backpropagation. Neural networks generate accurate predictions by passing data through the network (forward pass) and subsequently updating weights to refine the predictions (backpropagation).

The strength of neural networks lies in their learning phase, where the weights and biases are refined using optimization techniques such as stochastic gradient descent (SGD) (Robbins & Monro, 1951), Root Mean Square Propagation (RMSProp) (Hinton, Srivastava, & Swersky, 2012), Adam (Kingma & Ba, 2014), and Nadam (Dozat, 2016). These methods aim to minimize the loss function by iteratively adjusting the model's parameters during training. This iterative process theoretically reduces the difference between the predicted outputs and actual targets, though this does not always happen in practice. Through each iteration or epoch, neural networks automatically construct features and complex structures from high-dimensional data (Alzubaidi et al., 2021) (LeCun, Bengio, & Hinton, 2015).

Several key factors significantly affect the performance of neural networks, most notably the network architecture (i.e. how the different layers are organized), hyperparameters, data quality, and compute resources. In particular, hyperparameters, such as the learning rate, activation function, and the number of hidden layers, are not updated by the algorithm and are manually assigned prior to training (Baydilli & Atila, 2018). Choosing the "best" hyperparameter for each of these values has been shown to play a crucial role in neural network performance (Koutsoukas, Monaghan, Li, & Huan, 2017) (Jaisswal & Naik, 2021) (Bardenet, Brendel, Kégl, & Sebag, 2013). However, hyperparameter optimization algorithms such as Grid Search, Random Search (Bergstra & Bengio, 2012) and Bayesian Optimization (Snoek, Larochelle, & Adams, 2012) have been proposed to find the best combination of these parameters given a neural network architecture.

2.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) (LeCun et al., 2015) have proven effective for natural language processing tasks such as sequence-to-expression (Alipanahi, Delong, Weirauch, & Frey, 2015). Rather than utilizing layers in which all nodes in one layer are connected to all other nodes in the subsequent layer (i.e. dense layer), CNNs utilize a *sliding window* approach to efficiently capture local dependencies within data. By sliding windows or filters across one-hot encoded nucleotide sequences (i.e. sequences represented as a two-dimensional vector) and by applying additional normalization layers (i.e. pooling layers, dropout layers), CNNs can identify important motifs and patterns that potentially influence gene expression (Figure 7).

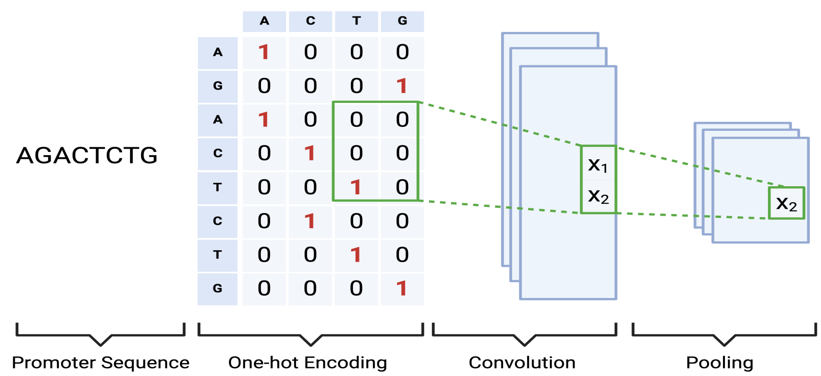

Figure 7: *An overview of Convolutional Neural Networks (CNNs). Starting from the left-most image, a promoter sequence is one-hot encoded such that the sequence is represented by a two-dimensional matrix. A filter (green box) iteratively slides throughout the one-hot encoded matrix and performs mathematical operations that result in a smaller matrix, otherwise known as a feature map. This process is known as a convolution. Finally, a pooling layer selects values from the feature map to further downsample or reduce the size of the matrix.*

2.1.3 Transformers

Transformers, which are another type of neural network, have been at the forefront of language-based modeling since its creation in 2017 (Vaswani et al., 2017). In the context of the sequence-to-expression task, the core of the transformer model is the "self-attention" mechanism, which enables the model to weigh the importance of different nucleotide sequences within a DNA strand. Unlike its sequential predecessors such as recurrent neural networks (RNNs) (Rumelhart, Hinton, & Williams, 1986) and LSTMs (Hochreiter & Schmidhuber, 1997), transformers are able to capture both short-term and long-term relationships and dependencies across the entire genetic sequence while being able to efficiently process sequences in parallel. Several studies utilize the combination of both transformer blocks (i.e. multiple transformer layers) and convolutional blocks (i.e. multiple convolutional layers) to accurately predict gene expression given a sequence (Vaishnav et al., 2022).

2.2 DREAM Challenge: "Predicting Gene Expression Using Millions of Random Promoter Sequences"

Each year, the non-profit Dialogue on Reverse Engineering Assessment and Methods (DREAM) partners with academic researchers and companies in order to host bioinformatics challenges that are open to the general public. Specifically, a challenge was hosted in 2022 to predict the gene expression within yeast using only promoters sequences (Rafi et al., 2023). Within this specific challenge, billions of promoter sequences within yeast and their approximate expression were given to participants as training data using a Gigantic Parallele Reporter Assay (GPRA) developed previously in their lab (de Boer et al., 2020).

Prior to this challenge, several studies by the organizers as well as other participants of this challenge have attempted this sequence-to-expression task albeit with different data. In particular, Vashinav et al. (Vaishnav et al., 2022) applied two different neural networks, a CNN and a transformer, to do so and achieved state-of-the-art results at the time using their transformer model.

With a long-term goal of translating the neural network architecture and lessons learned to humans at a later time point (Rafi et al., 2023), this challenge could provide initial insights as to how the REMIND-Cancer workflow can incorporate neural network approaches into its TFBS detection in the future given a promoter sequence and a sample's corresponding RNA-Seq data.

3 Methodology

Within this section, the methodologies of the REMIND-Cancer Workflow (See Section 3.1) and of the DREAM Challenge (See Section 3.2) will be detailed, with a significant focus on the former.

3.1 REMIND-Cancer Workflow

To recapitulate the aim of the REMIND-Cancer Workflow, my goal was to identify and characterize activating promoter SNVs *in silico*, irrespective of the mutation's recurrence frequency and driver status. To do so, this approach follows a filtering-ranking-inspection-validation paradigm in which somatic SNVs are (1) filtered to focus solely on putative functional mutations, (2) scored and subsequently ranked according to my heuristic scoring algorithm, (3) manually inspected via *pSNV Hunter* to assess their selection for further *in vitro* validation and (4) validated *in vitro* by a luciferase assay (Figure 8).

Figure 8: *Overview of the REMIND-Cancer workflow, which follows a filtering-ranking-inspection-validation paradigm.*

To outline the remaining section, the first subsection will detail the datasets used within this study, namely the PCAWG dataset (See Section 3.1.1.1), the early onset prostate cancer (EOPC-DE) dataset (See Section 3.1.1.2), the NCT-MASTER dataset (See Section 3.1.1.3) and finally, the COGNITION dataset (See Section 3.1.1.5). After explaining the datasets used, the following three subsections will then detail how I reduce the number of pSNVs through the filtering steps:

- Section 3.1.2.1 details how the promoter region was defined and determines what SNV is considered a subsequent pSNV

- Section 3.1.2.2 initially describes JASPAR (Fornes et al., 2020), a database of curated TFs and their binding profiles, and FIMO (Grant, Bailey, & Noble, 2011), which is a tool to search for motifs in DNA. Using these two, I then describe how a relative binding affinity is calculated for each *de novo* created or destroyed TFBS and how mutations are kept for further downstream analysis.

- Section 3.1.2.3 describes how the upregulation of a pSNV's adjacent gene expression was identified

The following three subsections will then detail additional features that were considered, how the REMIND-Cancer prioritization score is computed, two important quality control measures, and *pSNV Hunter*:

- Section 3.1.3 details how five additional features, namely recurrence, presence as a known cancer gene, open chromatin, tumor purity and allele frequency, were annotated per each pSNV. Unlike the previous subsections, no filtering or removing of pSNVs takes place during this time.

- Section 3.1.4 describes how the REMIND-Cancer prioritization score was computed for sunbsequent candidate ranking and *in vitro* validation

- Section 3.1.6 describes the use of two previously-established quality control measures: *DeepPileup* (Rheinbay et al., 2020) and *GenomeTornadoPlots* (Hong, Thiele, & Feuerbach, 2022)

- Section 3.1.7 describe the main features of *pSNV Hunter*, which was a tool created to assist in the selection of candidates for *in vitro* validation

The remaining subsections will then describe the luciferase assay system used (Section 3.1.8, the statistical testing employed to determine a significant upregulation in promoter activity (Section 3.1.9, how mutational signatures were detected (Section 3.1.10) and lastly, how TF activity is predicted via decoupleR (Badia-i Mompel et al., 2022), DoRothEA (Garcia-Alonso et al., 2019), and Collectri (Müller-Dott et al., 2023) (Section 3.1.11).

3.1.1 Data

The data used within the REMIND-Cancer workflow can be delineated into two categories: retrospective analysis and prospective analysis. The retrospective analysis utilized two datasets, namely PCAWG and *retrospective* NCT-MASTER data (i.e. NCT-MASTER data from November 2016 to October 2020), in order to fine tune the pipeline as well as prioritize pSNVs for subsequent *in vitro* validation.

Conversely, the prospective analysis also includes two datasets, namely the COGNITION and *prospective* NCT-MASTER datasets (i.e. new NCT-MASTER from 9 January 2023 to 13 March 2023), that were used as pilot studies to assess the applicability of the REMIND-Cancer workflow to true clinical trial programs. These two categories of data are detailed below.

3.1.1.1 Pan-Cancer Analysis of Whole Genomes (PCAWG)

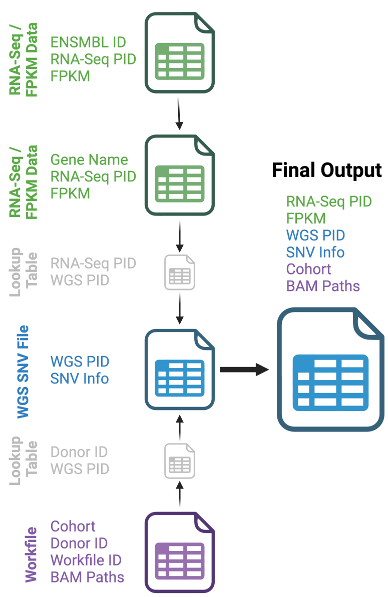

Figure 9: Schematic of how the PCAWG files were aggregated into a single file for analysis

Both the WGS and RNA-Seq data used in this study are from an internal version of the publicly-available PCAWG dataset, which can be found at at https://docs.icgc-argo.org/. This dataset consists of 2,520 WGS primary tumor samples of which 1,029 have corresponding RNA-Seq data available and belong to 45 distinct cohorts.

Prior to the analysis of this dataset, WGS, RNA-Seq, and cohort information data needed to be matched in order to fully leverage each patient's molecular profile. In particular, the WGS data did not share the same patient identifier (PID) as the matching RNA-Seq data, despite being from the same patient. Additionally, neither the WGS or RNA-Seq PIDs matched those in an annotation file containing the cohort and paths to their BAM files. Therefore, several preprocessing steps were necessary to aggregate these files into a usable format for the REMIND-Cancer workflow, which is exemplified in Figure 9.

SNV files in variant call format (.vcf) were identifiable by their WGS PID. Dr. Chen Hong, a former member of the Division of Applied Bioinformatics at the DKFZ, created a look-up table linking WGS PIDs (blue) to RNA-Seq PIDs (green). In particular, Dr. Chen Hong created an expression table where each column represented the genes (originally in ENSMBL format), each row represented an RNA-Seq sample and each value within the table represented an FPKM value. After transforming the columns of this table such that the ENSMBL gene identifier was transformed into gene names, this allowed for the direct integration of gene expression into the SNV .vcf files using the gene name as a key.

Furthermore, cohort information in addition to their bam files (purple) was given for each donor ID and workfile ID, each of which did not match either the WGS PID or the RNA-Seq PID. Consequently, Dr. Lars Feuerbach created an additional look-up table that mapped donor IDs to WGS PIDs, enabling the linkage of WGS PIDs, SNV files, RNA-Seq PIDs, cohort, donor ID, workfile ID, and BAM file paths (final output). Only the data files of WGS PIDs within an inclusion list provided by PCAWG were kept, which ensures optimal quality of the data.

In total, the PCAWG dataset comprise 2,520 total patients, 1,029 of which have RNA-Seq data available. An overview of the percentage of samples with and without RNA-Seq data available can be seen in Figure 10 for each cohort.

3.1.1.2 PCAWG: Early Onset Prostate Cancer (EOPC-DE) Cohort

As part of the PCAWG project, only the WGS data for the early onset prostate cancer (EOPC-DE) cohort was made publicly available. However, within my research division (DKFZ's Division of Applied Bioinformatics, led by Prof. Dr. Benedikt Brors), we have access to an internal version of this cohort's RNA-Seq data. I integrated this RNA-Seq data with the WGS data to create a complete cohort that could be analyzed by the REMIND-Cancer pipeline. Consequently, this EOPC-DE cohort included 56,608 SNVs from 23 unique samples, all of which had corresponding RNA-Seq data available.

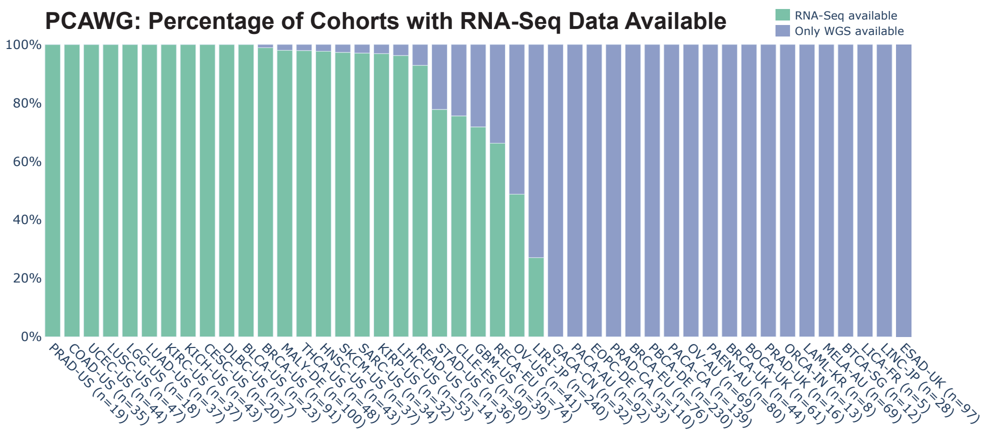

Figure 10: *A bar chart representing the percentage of WGS samples with and without corresponding RNA-Seq data available for the 44 PCAWG cohorts. The cohort abbreviations and the total number of WGS samples can be seen on the x-axis whereas the green bar represents WGS samples with RNA-Seq data available and blue represents only having WGS available (i.e. no corresponding RNA-Seq data).*

3.1.1.3 *Retrospective* NCT-MASTER

The *retrospective* NCT-MASTER dataset is from a data freeze that contains samples sequenced from November 2016 to October 2020. This dataset comprises 2,378 total WGS samples corresponding to 40,660,325 SNVs. Of these samples, 1,957 have available RNA-Seq data, which corresponds to 34,172,384 SNVs. Unlike the PCAWG dataset, the WGS and RNA-Seq data were organized within the same patient folder in the DKFZ cluster system, simplifying the matching process. Regardless, samples were categorized as being within 17 distinct cohorts and were given to me by Dr. Jennifer Hüllein of the NCT-MASTER Molecular Tumor Board. Of these 1,957 samples with RNA-Seq data, 1,718 (88% of samples with RNA-Seq data; 72% of all samples) came from the primary tumor whereas the remaining came from a metastatic sample (Figure 11). Additionally, samples within the "Anderes / Other" cohort as well as the "Sarkom / Sarcom" cohort were omitted, which will be explained in Section 4.4. Therefore, after determining the dates of the data freeze and considering only those mutations coming from the primary tumor with RNA-Seq data available and belonging to a well-defined cohort, 862 (36% of all samples) samples corresponding to 16,501,084 (40% of all SNVs) remained. To denote these remaining samples and for brevity, I will thus be calling these *analyzable samples*.

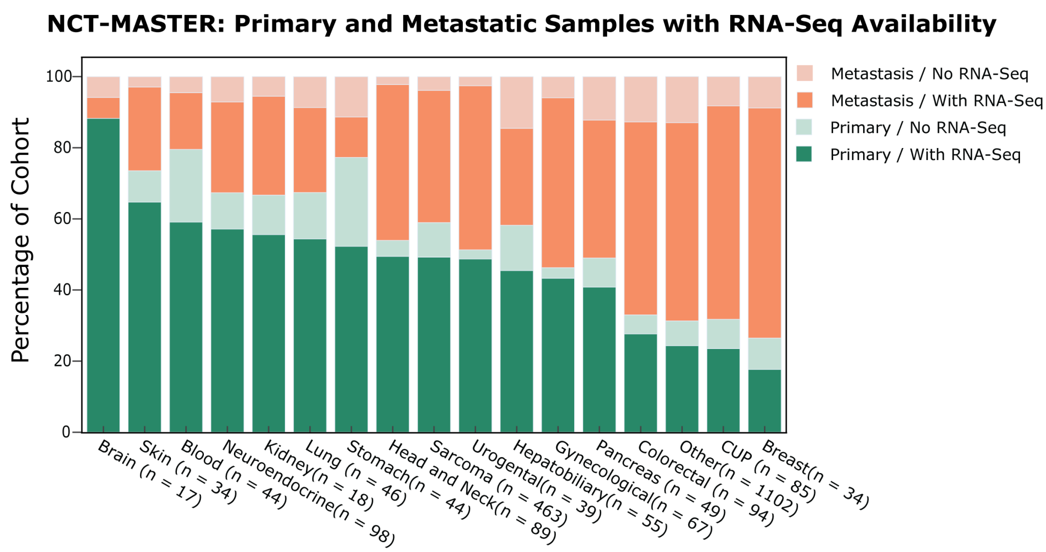

Figure 11: *A bar chart of each cohort representing the percentage of samples belonging to each of the four unique categories: (1) WGS samples from the primary tumor with matching RNA-Seq data (dark green), (2) WGS samples from the primary tumor without matching RNA-Seq data (light green), (3) WGS samples from a metastatic site with RNA-Seq data (dark orange), and (4) WGS samples from a metastatic site without RNA-Seq data (light orange). Only those WGS samples from the primary tumor with RNA-Seq data (dark green) were completely analyzable by the REMIND-Cancer Workflow.*

3.1.1.4 *Prospective* NCT-MASTER

The *prospective* NCT-MASTER data consists of newly enrolled patients between 9 January 2023 and 13 March 2023, which encompasses a 9-week period. Information regarding these patients was sent to me weekly by Dr.

Barbara Hutter of the NCT-MASTER MTB.

Each week, new information came via email in a tabular format that contained information such as the unique patient identifier, indication of whether the sample originated from the primary tumor or a metastatic site and annotations by the clinicians among other information. Importantly, however, specific cohort information was not given within this email. Consequently, prior to analysis, I manually inspected the clinician comments in order to assign each sample to one of the 17 cohorts used in the *retrospective* NCT-MASTER program. This brought forth multiple challenges, however, which I describe in detail at a later time in Section 6.1.6.3.

Using the unique patient identifier given in the email, the WGS and RNA-Seq files for each patient were extracted from the DKFZ cluster system and combined to create a single file. In total during this 9-week period, 105 samples corresponding to 1,160,518 mutations were extracted for analysis.

3.1.1.5 COGNITION

In addition to the *prospective* NCT-MASTER dataset, I also received data from the COGNITION (Pixberg et al., 2022) clinical trial (See Section 1.10.3) by Dr. Mark Zapatka. Though 213 samples were reported (Pixberg et al., 2022), only 89 (42%) WGS samples were transferred, all of which had RNA-Seq data available. Presumably, the remaining 124 (58%) samples were not provided as they were WES samples and may not include the entire promoter region.

Of the 89 samples that were given, 63 (71%) had a specifically-labeled subtype, some subtypes of which had a low sample size. Within the labeling of these subtypes, *HR+* and *HR-* refers to Hormone Receptor positive and negative, respectively, whereas *HER2+* and *HER2-* refers to human epidermal growth factor receptor 2 positive and negative, respectively. Specifically, HR+HER2- had 26 samples, TNBC had 24 samples, HR+HER2+ had 9 samples, HR-HER2+ had 4 samples whereas the remaining 26 samples were unlabeled. Particularly due to low sample size for the HR+HER2+ and HR-HER2+ subtypes as well as many samples being unlabeled, I considered all samples to be from the same cohort.

Utilizing the WGS and RNA-Seq data files given to me, I combined the patient information into a single file where all necessary data was available (i.e. FPKM, VAF) except for tumor purity, which was thereby set to 0 (or not available) when computing a prioritization score that will be detailed in

Seciton 3.1.4.

3.1.2 Filtering Steps

As part of the REMIND-Cancer computational pipeline, three filtering steps, namely (1) the promoter filter, (2) the TFBS motif and TF expression filter and (3) the gene expression filter were applied to the datasets described above. This section will detail how each filtering step was conducted.

3.1.2.1 Promoter Filter

Promoter regions are defined as spanning from 1,000 base pairs upstream to 500 base pairs downstream of a gene's canonical transcription start site (TSS), considerate of the strand. This annotation follows the GRCh37 (hg19) genome assembly and GENCODE Annotation V19 (Frankish et al., 2019), which can be downloaded here: `https://www.gencodegenes.org/human/release_19`. Only SNVs within a promoter region were used for subsequent analysis.

3.1.2.2 TFBS Motif and TF Expression Filter

The TFBS motif and TF expression Filter consists of two parts: (1) predicting the creation and/or destruction of a TFBS and (2) considering the expression of the corresponding TF.

To determine the impact of a pSNV on the creation or deletion of TFBSs, the Finding Individual Motif Occurrences (FIMO) (Grant et al., 2011) tool was utilized to search for specific DNA motifs by comparing them against the JASPAR (Fornes et al., 2020) database of TFs. Two 15 bp sequence contexts were considered: the wildtype (WT) sequence, which corresponds to the natural sequence according to hg19 (e.g. ACACT), and the mutant (MUT) sequence, which includes the pSNV (e.g. AC<u>**T**</u>CT for an A>T mutation).

For both WT and MUT sequences, FIMO compares these sequences to all motifs within the JASPAR database, represented as position frequency matrices (PFMs), to compute a log-likelihood ratio score for finding the motif within the given sequence. If the motif has a high ratio score, it is determined that this TF was found in the given sequence. Consequently, both the WT and MUT sequences will have a list of found motifs with some motifs present in both and others unique to only one.

To determine the creation and/or destruction of a TFBS introduced by the pSNV, the score of a motif in the MUT sequence is divided by its corresponding score in the WT sequence to obtain a MUT-to-WT ratio (Equation 1). If a score is missing (i.e. the motif is found only in one sequence), a default score of 1 is used to facilitate the calculation. This MUT-to-WT ratio is then used to determine TFBS creation and destruction.

$$\text{Ratio}_{\text{MUT/WT}} = \frac{\text{Score}_{\text{MUT}}}{\text{Score}_{\text{WT}}} \tag{1}$$

where

- $\text{Score}_{\text{MUT}}$ is the score of the motif in the mutant sequence if it exists; 1 otherwise

- Score_{WT} is the score of the motif in the wildtype sequence if it exists; 1 otherwise

A TFBS is determined to be created if $\text{Ratio}_{\text{MUT/WT}} > 11$. Conversely, a TFBS is predicted to be destroyed if $\text{Ratio}_{\text{MUT/WT}} < \frac{1}{11} = 0.09$. The rationale behind these thresholds can be found in Section 4.3. Subsequently, only pSNVs that result in the creation and/or destruction of at least one TFBS where the corresponding TF also shows detectable expression (i.e. FPKM > 0) are kept for anaysis.

3.1.2.3 Gene Expression Filter

To identify upregulated genes, the FPKM of the pSNV's corresponding gene was zscore normalized relative to the expression of this gene in samples of the same cohort. The z-score calculation is depicted in Equation 2.

$$\text{z-score} = \frac{x - \mu^{\text{gene}}_{\text{cohort}}}{\sigma^{\text{gene}}_{\text{cohort}}} \tag{2}$$

where

- x is the FPKM of the gene

- $\mu^{\text{gene}}_{\text{cohort}}$ is the average (mean) FPKM of the gene within the cohort

- $\sigma^{\text{gene}}_{\text{cohort}}$ is the standard deviation of the gene's FPKM within the cohort

The rationale for normalizing the FPKM relative to the gene and cohort could be found in Section 4.2. A zscore greater than 0 would indicate that the associated gene exhibits elevated expression and is thus kept for further analysis.

3.1.3 Additional Annotations / Features

The following subsection details the optional features, namely recurrence, presence within the Cancer Gene Census (CGC) database, chromatin accessibility, tumor purity, and variant allele frequency (VAF), that were added for each sample-specific pSNV within the REMIND-Cancer workflow.

3.1.3.1 Recurrence

The recurrence of each pSNV was calculated relative to its specific dataset (i.e. PCAWG pSNV recurrence was calculated only against other PCAWG pSNVs). Additionally, when calculating recurrence, every WGS sample, regardless of RNA-Seq availability and biopsy location, was considered.

To compute the recurrence of each pSNV, a Python dictionary was employed as a database to keep track of each mutation. Every line of every WGS data file was sequentially parsed to extract various details such as chromosomal location, gene name, reference nucleotide and alternate nucleotide. These details comprised the dictionary key while additional information such as patient ID, cohort, and expression (if available) were stored as a tuple within the associated list, forming the value for the key. Therefore, the dictionary contained keys representing chromosomal locations, gene names, reference nucleotides, and alternate nucleotides and the corresponding value of each key stored specific patient information for the entire dataset.

Using this dictionary, I added the mutational recurrence number for each remaining pSNV, ensuring to track mutations not belonging to the same patient in the dictionary and having the same alternate nucleotide.

3.1.3.2 Cancer Gene Census Database

To identify whether a pSNV occurs in a known oncogene, the COSMIC Cancer Gene Census (CGC) (Sondka et al., 2018) database from November 11, 2020, which includes 723 unique genes, was used. A gene was considered a known oncogene and thus given a boolean value of "True" only if it was listed in this database regardless of its tier classification within COSMIC.

3.1.3.3 Chromatin Accessibility

I assessed whether each pSNV's chromosomal location had an open or closed chromatin state using ChromHMM (Ernst & Kellis, 2012) annotations. These annotations were derived from Ernst et al.'s universal full-stack model, which was developed by training a hidden Markov model on over 1,000 datasets across more than 100 cell types. As a result, each 200 bp interval in the genome was assigned one of their specific characterizations. Intervals labeled as bivalent promoters or containing histone modifications H3K4me1, H3K4me2, or H3K4me3 were considered to have open chromatin and were thus assigned a boolean value of "True" for open chromatin.

3.1.3.4 Tumor Purity

For the PCAWG dataset, an estimate of each sample's tumor purity was found within the publicly-available PCAWG data portal. Each sample's WGS ID was then matched to its corresponding purity estimate. This purity estimate was derived from six individual copy number callers, which individually analyzed the copy number profiles obtained from their WGS data, and were combined to establish a census purity value for each sample (Gerstung et al., 2020). Conversely, tumor purity estimates from the NCT-MASTER (both retrospective and prospective) were provided within a separate data file that was integrated with their corresponding SNV VCF file. However, purity estimates from the COGNITION dataset were not available and thus did not contribute to the prioritization score, which will be detailed in Section 3.1.4.

3.1.3.5 Variant Allele Frequency

The variant allele frequency (VAF), which represents the proportion of sequencing reads that support a particular mutation compared to the total number of reads covering that position, was provided for each pSNV within the PCAWG dataset. However, for the NCT-MASTER (both retrospective and prospective) dataset as well as the COGNITION dataset, I manually calculated this by dividing the number of reads supporting the variant allele by the total number of reads at that position thus resulting in a fraction between 0 and 1.

3.1.4 REMIND-Cancer Prioritization Score and Ranking

To prioritize mutations for *in vitro* validation, an empirically-calibrated weighted-sum scoring function was used to compute a prioritization score, which was then used for subsequent ranking of a sample-specific pSNV. This weighted sum score took into account the genomic, transcriptomic, and annotationed-based features that have been described in this section. The specific formula to calculate the prioritization score can be seen in Equation 3 and Equation 4. The maximum attantainble score for a pSNV was 114 where the minimum was 0.

Since these prioritization scores are specific to each sample, it is important to note that the same mutation (i.e. C>T mutation at chr5:1,295,250) will yield unique scores depending on the sample's normalized gene expression, tumor purity, and variant allele frequency. All other features will contribute equally to the score, as their values will remain identical across samples.

The specific features that were used and their corresponding weights can be found in Table 1. For recurrence and the number of created/destroyed TFBSs, a maximum contribution value to the score is given. Additionally, for tumor purity and variant allele frequency, a boolean feature weight was added to the prioritization score if the sample met or exceeded the specific threshold value. Lastly, when scoring both open chromatin at a pSNV position and presence of the adjacent gene within the CGC list, a fixed weight was added to the score if the features were "True".

$$\text{Prioritization Score} = \sum_{i=1}^{\text{all features}} g(x_i, w_i, m_i, t_i) \tag{3}$$

$$g(x_i, w_i, m_i, t_i) = \begin{cases} w_i x_i & m_i = \emptyset \quad \text{and} \quad t_i = \emptyset \\ min(w_i x_i, m_i) & m_i \neq \emptyset \quad \text{and} \quad t_i = \emptyset \\ w_i x_i & m_i = \emptyset,\ t_i \neq \emptyset \text{ and } x_i \geq t_i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

such that

- *Prioritization Score* is the prioritization score of a sample-specific pSNV
- x_i is feature i's value (e.g. zscore of 5)

- w_i is feature i's weight (e.g. normalized FPKM / zscore weight of 5)
- m_i is feature i's maximum value (e.g. maximum value of 25 for the normalized FPKM / zscore)
- t_i is feature i's threshold (e.g. *None* for normalized FPKM / zscore)

Feature Type	Feature Name / Value (x_i)	Weight (w_i)	Maximum (m_i)	Threshold (t_i)
Genomic	Number of Created TFBSs with FPKM > 0	2	6	None
Genomic	Number of Destroyed TFBSs with FPKM > 0	2	6	None
Genomic	Tumor Purity (Boolean Value)	10	None	0.25
Genomic	Variant Allele Frequency (Boolean Value)	10	None	0.3
Transcriptomic	Normalized FPKM	5	25	None
Annotations	Open Chromatin (Boolean Value)	20	None	None
Annotations	Cancer Gene (Boolean Value)	15	None	None

Table 1: *A table consisting of the 7 features considered for the weighted sum REMIND-Cancer prioritization score. Specifically, 'Feature Type' refers to one of the three major types (genomic, transcriptomic, and annotations) of features considered, 'Feature Name' refers to the name/description of the feature, 'Weight' refers to the weight given during the computation, 'Maximum' refers to the maximum contribution (i.e.* $max(maximum\ value\ noted, x_i \times w_i)$*) that the feature could contribute to the prioritization score (if applicable), and 'Threshold' refers to the threshold needed to contribute to the prioritization score (if applicable).*

3.1.5 Computational Pipeline

The filtering steps described in Section 3.1.2 and the annotations outlined in Section 3.1.3 were implemented in Python 3.11.0 using libraries such as Pandas v.2.1.1 (Reback et al., 2020), NumPy v.1.26.1 (Harris et al., 2020), and SciPy v.1.11.3 (Virtanen et al., 2020). All packages and versions used can be seen within the publicly-available REMIND-Cancer Computational Pipeline GitHub Repository page (Abad, Körner, & Feuerbach, 2024d).

At each stage of the process (e.g. gene expression filter, recurrence annotation), intermediate results were saved as a JSON file to ensure interpretability and data provenance. This JSON file was continuously updated after each step with the filtering/annotation steps as keys and the paths (including the PIDs) as values. An example of this workflow is illustrated in Figure 12.

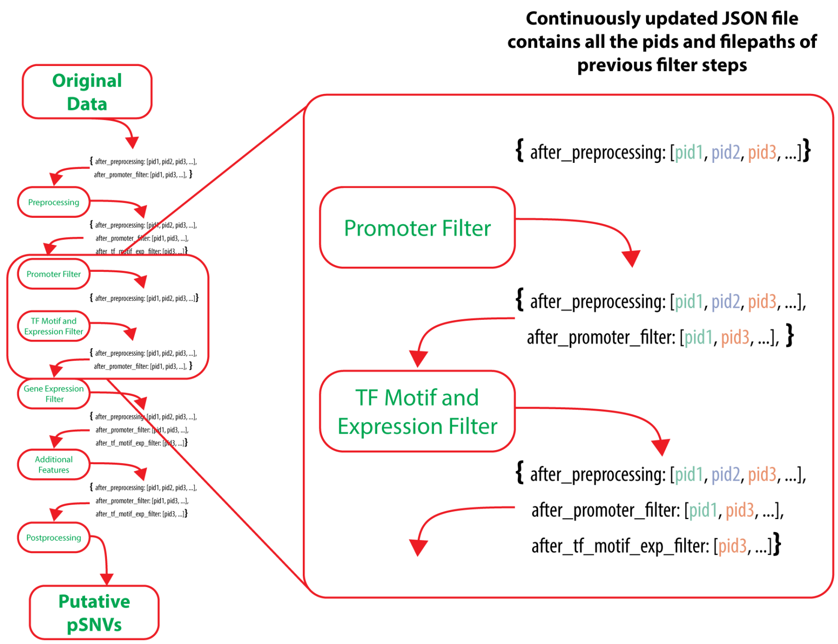

Figure 12: *An outline of the REMIND-Cancer pipeline's workflow, illustrating the transfer of information between stages (i.e. from the promoter filter to the TFBS motif and TF expression filter). The pipeline utilizes a JSON file to track the patients and mutations that progress to each subsequent phase.*

3.1.6 Quality Control Tools

During the inspection phase of the REMIND-Cancer workflow, two quality control tools, namely *DeepPileup* (Rheinbay et al., 2020) and *GenomeTornadoPlots* (Hong et al., 2022), were applied to putative pSNVs to assist in their selection for *in vitro* validation.

3.1.6.1 *DeepPileup*

DeepPileup (Rheinbay et al., 2020) was originally designed by Dr. Lars Feuerbach in order to distinguish true signals from potential alignment artifacts at individual genomic positions.

By considering all raw sequencing files for each cohort in the form of binary alignment map (BAM) files from both tumor and normal samples, two plots are generated. The first plot shows the percentage of samples per cohort in which the non-reference allele reaches a frequency greater than 25%. An example of a valid position could be seen in Figure 13a where nearly no noise (i.e. ∼0% of samples do not reach a minor allele frequency of 25%) is detected within control samples but signal is detected in a fraction of tumor samples (e.g. ∼8% of tumor samples in the MELA-AU cohort reach a minor allele frequency of 25%). Consequently, this position would be considered a technically valid position.

However, within Figure 13b for a different pSNV, there is an abundance of signal coming from normal samples, which in many cohorts (e.g. BOCA-UK) is greater than the signal coming from the tumor sample. Consequently, this genomic position could be considered to be noisy due to Figure 13b.

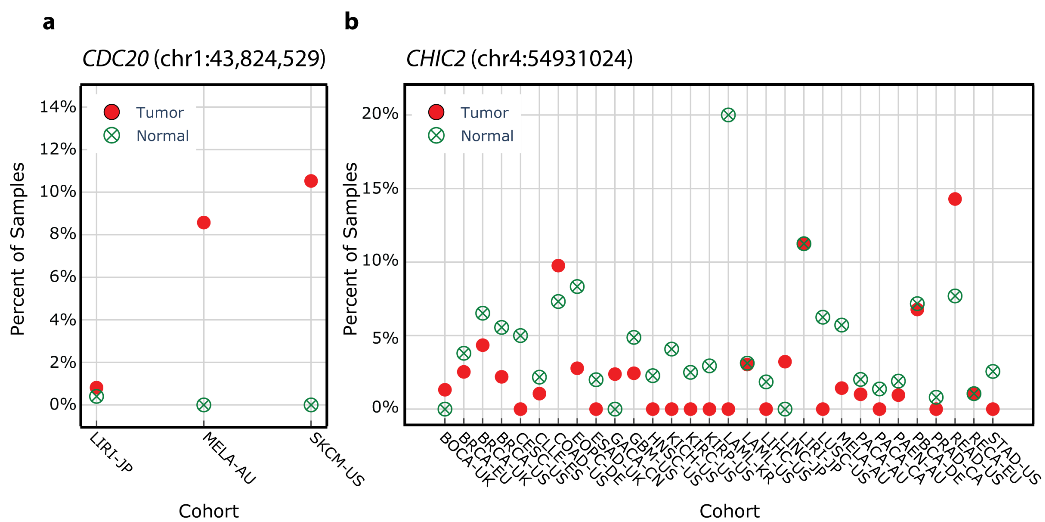

Figure 13: *The percentage of samples per cohort in which the minor allele frequency (MAF) is greater than 25% separated by tumor samples and normal samples. (a) A technically valid genomic position due to being almost no signal within control samples but signal within tumor samples (b) A noisy genomic position due to observing signal in both the control and tumor in a majority of cohorts*

Furthermore, a secondary plot is also generated that shows the percentage of samples per cohort with two or more variant alleles. For all pSNVs in both the PCAWG and retrospective NCT-MASTER dataset passing the filtering steps, both plots were generated.

Though this tool was originally designed by Dr. Lars Feuerbach, I further expanded this tool by updating the original code to upgrade from Python 2 to Python 3, vastly improving upon its computational efficiency, developing new forms of visualization, and making this publicly-available for general use on GitHub (Abad, Körner, & Feuerbach, 2024a).

3.1.6.2 *GenomeTornadoPlots*

GenomeTornadoPlots (Hong et al., 2022) was used in order to observe potential implications of convergent tumor evolution for specific genes within the PCAWG dataset. The original *GenomeTornadoPlot* package allows users to visualize copy number variations (CNVs), particularly focal deletions and amplifications, at specific gene locations.

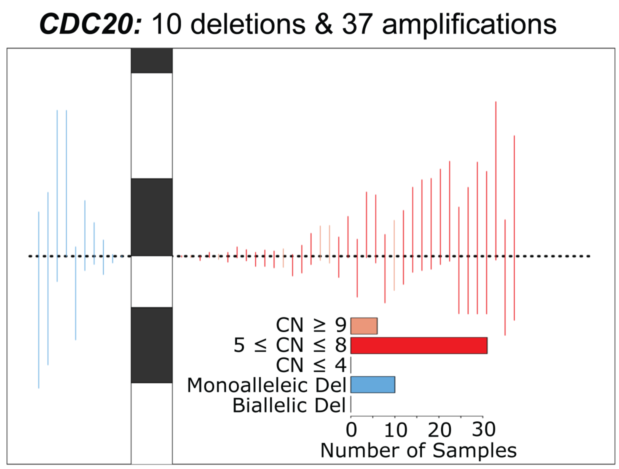

Figure 14: *An example of a GenomeTornadoPlot for CDC20 in which 10 focal deletions and 37 amplifications are visualized.*

To allow for easy integration into *pSNV Hunter*, which will be described in the following section, a wrapper script was created and can be found on GitHub (Abad, Körner, & Feuerbach, 2024b).

3.1.7 *pSNV Hunter*

To assist in the selection of putative pSNVs for functional validation during the inspection phase, I developed a general-purpose data aggregation and visualization tool called *pSNV Hunter*. This tool allows for the examination of individual pSNVs, enabling users to interactively assess factors such as adjacent gene expression, specific gene function, recurrence level, expression levels of transcription factors, sample-specific information, as well as view quality control plots from *DeepPileup* and *GenomeTornadoPlots*.

Additionally, users have the ability to take notes regarding their preferences or concerns about specific pSNVs and can conveniently extract candidates into a separate CSV file for further analysis. An overview of the most important aspects of pSNV Hunter along with actual screenshots of the tool can be seen in Figure 15.

A detailed description of each of these 12 features can be found below:

1. Individual pSNVs can be chosen for comprehensive analysis within a filterable (e.g. "only show pSNVs with a recurrence level of at least 5") and sortable (i.e. sort by prioritization score ranking, sort by gene name) data table of the remaining pSNVs

2. The expression levels (normalized z-score, FPKM, or natural log) of the adjacent gene along with the expression of this gene in the rest of the cohort and in cohorts with recurrent samples, are visualized through an interactive violin plot.

3. The function of the adjacent gene was extracted from National Center for Biotechnology Information (NCBI) (Brown et al., 2015) and can be read in its entirety along with a hyperlink to where this information came from.

4. Select which of the *de novo* created or destroyed TFBSs and their corresponding TFs to analyze further

5. The expression levels (normalized z-score, FPKM, or natural log) of the selected TF, as well as its expression in the remainder of the cohort, are depicted using a violin plot.

6. Information regarding the function of the chosen TF, obtained from NCBI, provides initial insights into its potential roles

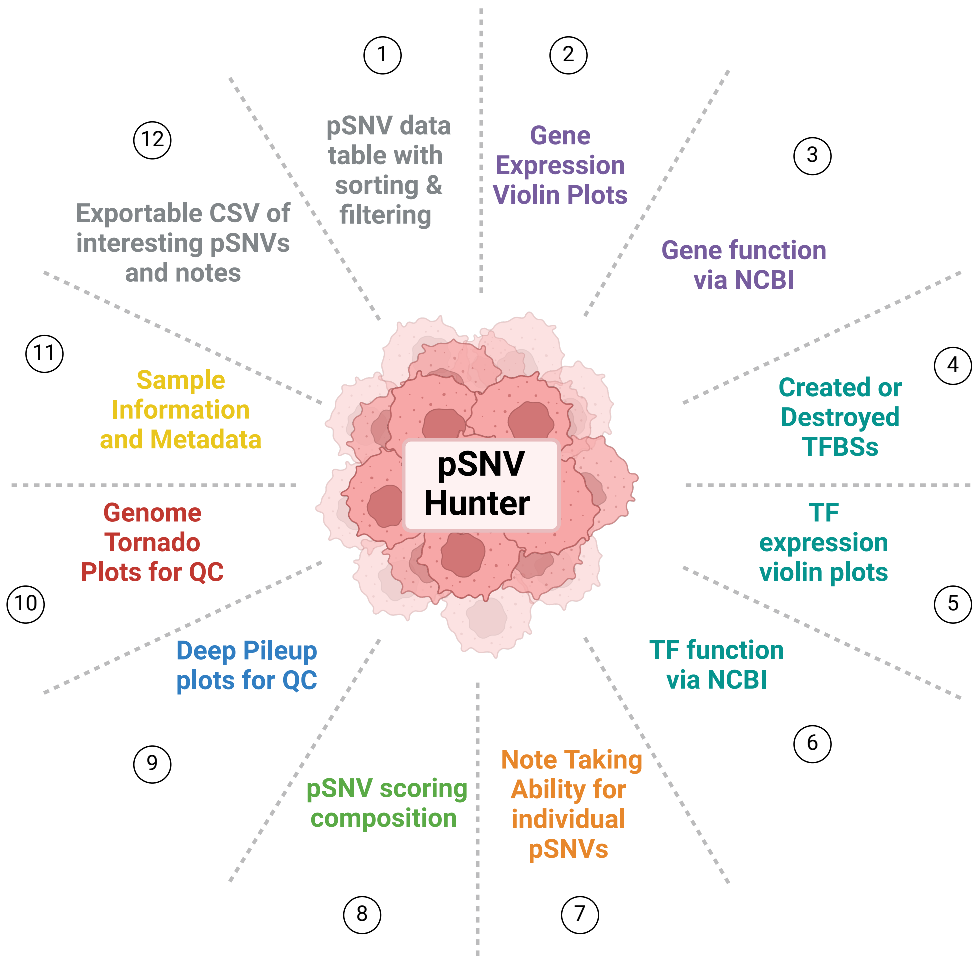

Figure 15: *An overview of the 12 most important features of pSNV Hunter. The different colors of text represent different tabs within pSNV Hunter. For example, the purple features (2. Gene Expression Violin Plots and 3. Gene function via NCBI) can be seen within a single tab that's related to the gene whereas the teal features (4. Created or Destroyed TFBSs, 5. TF expression violin plots relative to the different cohorts, and 6. TF function via NCBI) can be seen in a different tab related to the transcription factors.*

7. Users can take notes for individual pSNVs that can be exported individually or along with the results exported in (12)

8. A sunburst chart visualizes the scoring composition of the selected pSNV with the inner ring representing the three major scoring categories (genomic, transcriptomic, and annotations), and the outer ring detailing specific aspects (e.g., tumor purity under genomic category, gene expression level under transcriptomic category)

9. The two interactive *DeepPileup* plots as described in 3.1.6 can be seen and used for quality control purposes.

10. The *GenomeTornadoPlots* as described in 3.1.6 can be seen and used for quality control purposes.

11. Other important (meta)information about the sample can be seen such as the number of original SNVs in this sample, the number of remaining pSNVs and cohort information.

12. Users can export a CSV file containing only the pSNVs they find interesting, presented in a readable format.

pSNV Hunter was developed in Python 3.11.0 with both the backend and frontend constructed using Plotly Dash (Inc., 2015c) version 2.1.14. Plot generation is facilitated by Plotly (Inc., 2015a) version 5.18.0 while other essential libraries include Dash Bootstrap Components (AI, n.d.) version 1.5.0 and Dash Bio (Inc., 2015b) version 1.0.2. *pSNV Hunter* is fully-available online on GitHub (Abad, Körner, & Feuerbach, 2024c).

3.1.8 Luciferase Assay

The *in vitro* validation of candidate pSNVs was conducted by Sabine Karolus and Dr. Cindy Körner using a luciferase reporter assay in HEK293FT cells. The signal obtained from the firefly luciferase was normalized to the Renilla luciferase signal and the median of all conducted technical replicates was recorded. Furthermore, the percentage of upregulation (i.e. relative change) was calculated as the change in normalized signal between the mutant and its corresponding wildtype vectors in percentage. Further details (e.g. sequences used, restriction enzymes for subcloning, number of cells per well) can be found in our pre-print (Abad, Glas, et al., 2024).

3.1.9 Statistical Testing

The statistical comparison between WT and MUT luciferase activity focused on the percentage of upregulation using a one-sided t-test with an expected value of 0. The null hypothesis (H_0) states that the mean percentage upregulated in the mutant is less than or equal to that in the wild type. Therefore, a significant p-value (i.e., p-value ≤ 0.05) would lead to rejecting the null hypothesis, indicating a positive upregulation in the mutant compared to the wild type. I conducted this analysis using the SciPy (Virtanen et al., 2020) version 1.11.3 Python library.

3.1.10 Mutational Signature Detection

The three common UV-induced mutational signatures, namely SBS7a, SBS7b and SBS7c, were annotated for mutations belonging only to PCAWG's SKCM-US cohort. In the case of SBS7a, I determined that a pSNV matched this signature if the pSNV was a C $>$ T mutation in TpC dinucleotide context (i.e. TC $>$ TT). For SBS7b, I determined that a pSNV contained this signature if there was a C>T mutation at a CpC dinucleotide (i.e. CC>CT or CC>TC) and lastly, a pSNV contained SBS7c if it was either a T>C or T>A mutation, irrespective of the flanking nucleotides. I conducted this analysis through the use of Python 3.11.0.

3.1.11 TF Activity Prediction Using DoRothEA, Collectri and decoupleR

TF activity was predicted using the normalized gene expression (zscore) of each cohort for the both the PCAWG and NCT-MASTER dataset. In particular, four methods of decoupleR (1.6.0) (Badia-i Mompel et al., 2022) were employed: univariate linear model (ULM), multivariate linear model (MLM), weighted sum (WSUM), and VIPER and the consensus. For each algorithm, the default hyperparameters were used.

As part of the decoupleR package, both Collectri (Müller-Dott et al., 2023) and DoRothEA (Garcia-Alonso et al., 2019) could be loaded in directly within Python 3.11.0 as a Pandas (Reback et al., 2020) dataframe. In total, Collectri comprised 43,178 unique TF-gene interactions, corresponding to 1,186 unique TFs and 6,692 unique genes. In comparison, only the three most confident levels (i.e. A, B, and C) of DoRothEA were used, which had 32,275 unique TF-gene interactions, 429 unique TFs and 9,228 unique genes.

3.1.12 Thesis and Figures

This thesis was written entirely in LaTeX using the publicly-available and free version of OverLeaf (https://www.overleaf.com/). Graphs, figures, and tables were generated using either BioRender (https://www.biorender.com/) or Adobe Illustrator 2024. When applicable, figures were initially generated using the Plotly (Inc., 2015a) Python library and then refined in Adobe Illustrator 2024 for improved visualization. Importantly, only cosmetic adjustments (e.g. color scheme, font sizes, location of legend) were made with no alterations to specific data values to misguide the reader. Unless otherwise noted such as within Figure 3, I generated all figures, graphs, and tables independently.

3.2 DREAM Challenge

As discussed in Section 2.2, the objective of the "Predicting Gene Expression Using Millions of Random Promoter Sequence" 2022 DREAM challenge was to predict gene expression within yeast using only a small (∼80 bp) promoter sequence, effectively requiring participants to identify TFBSs without the support of external databases. I individually competed in this challenge against 292 teams, many of whom had prior experience in sequence-to-expression modeling. Specific details of the competition are provided (https://www.synapse.org/Synapse:syn28469146/wiki/617075) whereas a preprint of this challenge has recently been published (Rafi et al., 2023).

To detail the structure of the challenge, there was no limit on the size of a team though some teams consisted of entire research labs. Additionally, each team was given the the training dataset (i.e. CSV file of a sequence and expression) as well as extra computational power (if needed) in the form of Google's tensor processing unit (TPU) access for faster computing.

For the 12 weeks leading up to the final submission date, the project organizers created a pre-submission leaderboard in which teams could benchmark their performance against other teams. Importantly, teams did not necesasrily have to participate in this phase nor did this count towards the final ranking during the actual submission. In total, 292 teams participated in this phase thereby entering the challenge. During this phase, each team could submit up to five models, which would then be used to predict the expression on a heldout validation set originating from the final test set. Teams and their subsequent models were ranked based on two metrics when comparing the actual to predicted expression levels: Pearson correlation r^2 and Spearman ρ correlation.

After 12 weeks, the *final* test set in which teams were truly evaluated on consisted of 71,103 promoter sequence and expression pairs. In addition to being evaluated on the Pearson and Spearman correlation of these sequences, the test set was further divided into small sub-categories such as those sequences with high expression and sequences corresponding with low expression. Final rankings were a weighted sum of these measures with a majority of the scoring weight going to the two main categories of overall Pearson and Spearman correlation.

3.2.1 Data

Both the training and testing datasets contained the gene expression in yeast (Saccharomyces cerevisiae) paired with a promoter sequence. Utilizing a data collection method mirrored in their own previous study (de Boer et al., 2020), the project organizers inserted 6,739,258 random DNA sequences into promoter constructs and cloned these into low-copy-number vectors containing yellow fluorescent protein (YFP) and a red fluorescent protein (RFP) to act as a control. The yeast was grown and the expression was measured by sorting cells into 18 bins based on the logged YFP/RFP fluorescence ratio (YFP / RFP). Yeast from each bin were then grown, vectors isolated, and the promoter sequences were amplified and sequenced. The final data is a table linking each promoter sequence to its expression level, calculated from the sorting bins. Additional sequencing details can be found in Rafi et al. (Rafi et al., 2023).

The promoter expression level for the training data represented the *weighted average* of expression bins in which that promoter was observed (weighted by the fraction of reads in that bin). Therefore, many of the promoters were only seen in a single bin, which led to many promoters with expression levels that exactly correspond to an integer (Figure 16a). However, for the test set, the exact expression, rather than a weighted average, was generated (Figure 16b).

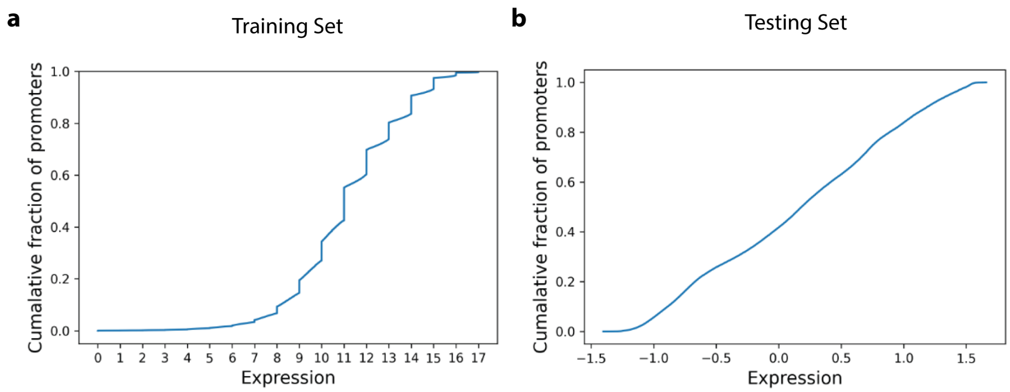

Figure 16: *Cumulative distributions of the expressions of the training (a) and test (b) datasets. These images originated from the official DREAM challenge page:* https://www.synapse.org/Synapse:syn28469146/wiki/617557.

During the protoyping and subsequent training of my model, the original training set, comprising 6,739,258 random DNA sequences and their corresponding expressions, was randomly split according to an 80/20 ratio. Consequently, 5,391,406 sequences were used to train the neural network architecture while the remaining 1,347,852 sequences served as a validation set to assess my model quality.

3.2.2 Pre-processing Strategies

To better simulate the true test set, random noise was introduced to some of the training data. For each promoter sequence in the training dataset with an integer gene expression value of x, a random sample was drawn from a $Normal(x, 0.3)$ distribution. This new value was used to replace the original integer expression. Other variances were also tested throughout the challenge, including sampling from a $Normal(x, 0.2)$ or $Normal(x, 0.4)$ distribution, as well as different distributions such as sampling from a log-normal, Poisson, and Negative Binomial distribution. All random sampling was conducted using Python 3.11.0 using NumPy v.1.26.1 (Harris et al., 2020).

3.2.3 Transformer Neural Network

For this sequence-to-expression task, a neural network was utilized, featuring primarily multi-head attention, convolutional, and basic feed-forward layers, along with additional layers such as batch normalization, dropout, concatenation, padding, and flattening. The final architecture comprised 55 distinct layers with unique hyperparameters, resulting in 1,137,113 trainable parameters (weights and biases).

To briefly describe the basic neural network architecture: the input data was first transformed (one-hot encoded) into a (110, 4) binary matrix, where each nucleotide corresponds to one of the four columns, and a value of 1 indicates the presence of that nucleotide at a specific position. The reverse complement of this input sequence was generated and the network was split into two "tracks," one handling the sequence in the normal context and the other handling the reverse complement. Several multi-headed attention blocks, each comprising a multi-head attention layer followed by an activation function, dropout layer, normalization layer, convolutional layers, and feed-forward layers, were applied to each track separately. These two tracks were then concatenated, followed by additional multi-head attention blocks and several dense layers, culminating in a final dense layer to produce the output.

The loss function used to adjust the learnable parameters was mean squared error (MSE) to penalize large errors and Nadam was used as the final optimizer. The network was trained for 1,000 epochs using computational resources provided by the challenge (Google Research's TPUs) and the epoch corresponding to the lowest validation MSE was used as the best model weight to account for potential overfitting during training.

The neural network was constructed and trained in Python 3.11 using Keras 3.3.0 (Chollet, 2015) and Tensorflow 2.16.1 (Abadi et al., 2015).

3.2.4 Hyperparameter Optimization

To select the best hyperparameters, Bayesian Optimization (BO) was employed with expected improvement (EI) as the acquisition function. The following hyperparameters were optimized utilizing the following prior distribution, minimum values and maximum values:

- Learning Rate with a prior of $Uniform(0.000001, 0.1)$; Starting value of 0.001

- Dropout Rate with a prior of $Uniform(0, 0.5)$; Starting value of 0.03
- L1 Regularizer weight with a prior of $Uniform(0, 1)$; Starting value of $1e^{-5}$
- L2 Regularizer weight with a prior of $Uniform(0, 1)$; Starting value of $1e^{-4}$
- Attention Block Dropout Rate with a prior of $Uniform(0, 0.5)$; Starting value of 0
- Number of Convolutional Filters of 256, 512, 1024 or 2048; Starting value of 256
- Convolutional Kernel Size with a prior of $Discrete\ Uniform(4, 256)$; Starting value of 30

After 20 iterations or calls of BO, the best hyperparameters were determined to be a learning rate of 0.002, dropout rate of 0.023, L1 regularizer weight of 0, L2 Regularizer weight of 0, attention block dropout rate of 0.034, 256 convolutional filters, and a convolutional kernel size of 64. This algorithm utilized the Scikit-Optimize 0.8.1 (Head, 2016) package in Python 3.11.0.

4 Calibrating the REMIND-Cancer Workflow

Within this section, details regarding the rationale and development of the three REMIND-Cancer filtering steps, namely the (1) promoter, (2) gene expression, and (3) TFBS motif and TF expression filters, will be provided in Sections 4.1, 4.2, and 4.3, respectively. Although I provide the rationale for the default threshold values for each filter, users of the open-sourced REMIND-Cancer computational pipeline (Abad, Körner, & Feuerbach, 2024d) are able to adjust these values tailored to their own analysis if needed.

Furthermore, the details regarding the inclusion criteria, particularly for the *retrospective* and *prospective* NCT-MASTER datasets, will be explained in Section 4.4.

4.1 Promoter Filter

As my study's main focus is on SNVs within the promoter region of protein coding genes, the promoter filter was implemented first to significantly reduce the amount of SNVs considered for further analysis. However, determining the promoter region was a key question that needed to be considered.

In the literature, several promoter databases, such as the Tissue-Specific Promoter Database (TiProd) (X. Chen et al., 2006) and the Eukaryotic Promoter Database (EPD) (Périer et al., 2000), have been developed for general use. However, utilizing databases such as these introduce unique challenges such as needing to make gene-specific (e.g. using Ensembl or GENCODE TSSs) and tissue-specific assumptions. Even when these assumptions are met, these databases can have drastically different predictions of the promoter region for the same gene within the same tissue.

Previous studies (Rheinbay et al., 2020) (Fu et al., 2014) (Mularoni et al., 2016) have addressed this issue by using more general promoter regions, thereby trading specificity (potentially including regions that are not true promoters) for increased sensitivity (identifying more potential promoter regions). As my study builds on a previous study (Rheinbay et al., 2020) that used a promoter region of 500 bp upstream to 500 bp downstream of the TSS, I chose to be slightly more inclusive. Thus, I defined the promoter region as 1,000 bp upstream to 500 bp downstream of the TSS to include more SNVs in the analysis.

Furthermore, as GENCODE v19 (Frankish et al., 2019) was previously used for TSS annotation (Rheinbay et al., 2020), I also used this convention. This therefore resulted in 47,970 promoter regions, each of which corresponded to a specific gene and are defined as being 1,500 base pairs long (1,000 bp upstream to 500 bp downstream of the TSS).

4.2 Gene Expression Filter

To detect the aberrant upregulation of a pSNV's corresponding gene, the normalization of genes' expression relative to both the cohort and the gene itself was deemed necessary due to different tissue types having different expression profiles.

As an example, consider the previously-implicated gene *RALY* and its expression profile within our PCAWG dataset in Figure 17. Here, each cohort displays a distinct distribution of expression levels, which renders the direct comparison of expression values between cohorts unreliable.

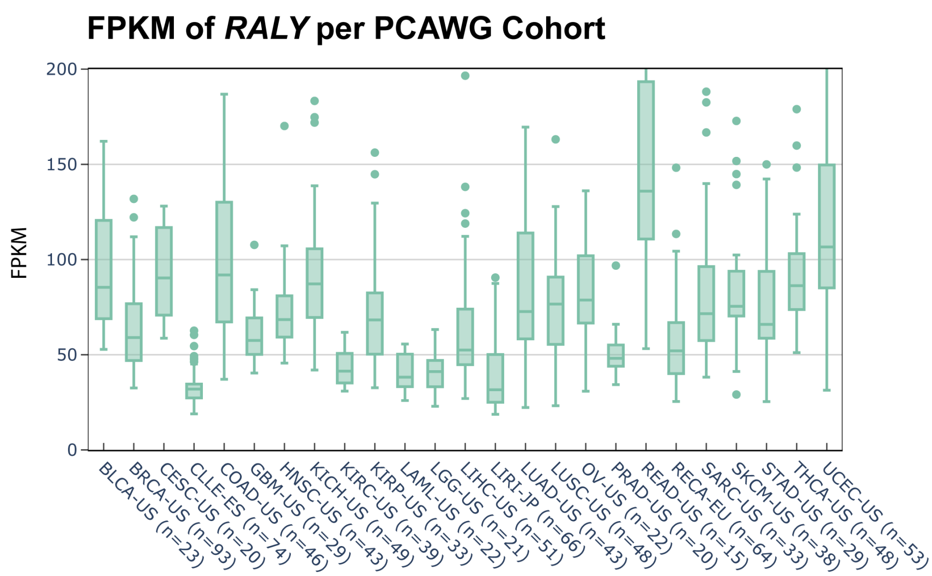

Figure 17: *An example of the differing FPKM levels of RALY within the different PCAWG cohorts. With different distributions of expression levels, this renders the comparison of expressions between cohorts unreliable.*

After zscore normalization, this theoretically implies that the scores will follow a standard normal distribution (i.e. $X \sim \text{Normal}(0, 1)$). Those pSNVs corresponding to genes with zscores above 0 implies that the gene is upregulated *relative to other samples within the cohort* whereas a negative zscore implies the opposite (i.e. downregulated relative to other samples within the cohort). By only keeping those pSNVs with zscores above 0, a semi-stringent filter is put in place where, theoretically, 50% of the remaining pSNVs with RNA-Seq data will remain.

4.3 TFBS Motif and TF Expression Filter

To further detect activating pSNVs, the TFBS motif and TF expression filter was utilized, which follows a multi-step process. As described in Section 3.1.2.2, pSNVs (1) leading to the creation and/or destruction of TFBS(s) and (2) having that corresponding TF being expressed (i.e. FPKM > 0) leads to keeping that pSNV for further analysis.

As detailed in Section 3.1.2.2, determining the creation or destruction of a TFBS involves inputting the WT and MUT sequences into FIMO (Grant et al., 2011) to identify matching TFs in the JASPAR2020 (Fornes et al., 2020) database. If a TF is found in either sequence, a binding affinity score is assigned to that sequence. For example, if a TF is present in the WT sequence but not in the MUT sequence, it receives a binding affinity score for the WT sequence but not for the MUT sequence. Thus, a ratio $\text{Ratio}_{\text{MUT/WT}}$ between a TF's MUT and WT is computed, which is then used to determine whether a TF is created or destroyed.

The default ratio threshold for the creation of a TFBS is 11. This threshold was chosen because of the two $TERT_{C228T}$ and $TERT_{C250T}$ hotspot pSNVs, which are known to create binding sites for ETS-family TFs such as GABPA and ELK4. As such, the $TERT_{C228T}$ $\text{Ratio}_{\text{MUT/WT}}$ for GABPA is 11.3 whereas the $\text{Ratio}_{\text{MUT/WT}}$ for ELK4 is 13.2. Similarly, for $TERT_{C250T}$, the $\text{Ratio}_{\text{MUT/WT}}$ for GABPA and ELK4 are 11.3 and 13.2. Due to wanting to be more inclusive of potentially created TFBSs, a slightly lower threshold of 11 was decided upon as a creation threshold. Using this creation threshold of 11, the reciprocal of this (i.e. $\frac{1}{11} \approx 0.9$) was used as the TFBS destruction threshold.

4.4 Data Inclusion

The PCAWG and NCT-MASTER datasets contain different types of samples. PCAWG samples, for instance, are exclusively from primary tumors, indicated by study abbreviations (e.g., breast cancer = BRCA-US), whereas NCT-MASTER samples were taken either from the primary tumors or metastatic sites. This distinction, along with the lack of clinical annotations in the NCT-MASTER data, was crucial to consider, particularly for normalizing expression values throughout the REMIND-Cancer pipeline.

Particularly for the NCT-MASTER dataset, all cohorts were considered for analysis with the exception of the "Anderes / Other" and "Sarkom / Sarcoma" cohorts. The "Anderes / Other" cohort included unclassified samples assigned via an annotation file, potentially encompassing samples from other cohorts and thereby greatly affecting normalized gene expression. The "Sarkom / Sarcoma" cohort was excluded due to its broad categorization and heterogeneity, exemplified by having over 50 subtypes in soft tissue sarcoma alone (Katz, Palmerini, & Pollack, 2018) and each subtype exhibiting different expression profiles (Sarver, Sarver, Thayanithy, & Subramanian, 2015). This heterogeneity is reflected in the number of samples, with Sarcoma (n=463) being the most represented cohort, far exceeding Neuroendocrine (n=98) and Colorectal (n=94), which are the next most populous cohorts. Further providing evidence of this expression hetereogeneity is the observation of multi-modal expression distributions within a majority of genes in sarcoma samples therefore leading to the exclusion of this cohort.

For tumor-type omission, only samples derived from the primary tumor were included in downstream analysis by the REMIND-Cancer pipeline. Various studies have investigated and reported transcriptomic differences between metastatic tumors and their paired primary tumors from the same patient across various cancer types, typically citing that expression differences are due to cancer stage and the location where the metastatic sample was taken (Y. Zhang, Chen, Balic, & Creighton, 2024) (Aftimos et al., 2021) (Garcia-Recio et al., 2023) (Y. Zhang, Chen, & Creighton, 2023) (Iwamoto et al., 2019) (Cosgrove et al., 2022). Particularly in NCT-MASTER cohorts with small sample sizes, such as brain (n=17) or kidney (n=18), having even one metastatic sample with significantly different expression for a particular gene will drastically alter the zscores and potentially affect the detectability of pSNVs corresponding to upregulated genes. Consequently, only samples from the primary tumor were considered for downstream analysis.

5 Results

5.1 Overview of the REMIND-Cancer Workflow

To summarize the primary aim of my PhD project, the REMIND-Cancer Workflow was created to identify and prioritize activating pSNVs using a recurrence-agnostic approach. In addition to detecting highly-recurrent mutations, I wanted to design a pipeline that also identified often-overlooked singletons and lowly-recurrent mutations.

In order to do so, sample-specific SNVs from the PCAWG and retrospective NCT-MASTER datasets were initially subjugated to three filtering steps: (1) a promoter filter (see 3.1.2.1), (2) a gene expression filter (see 3.1.2.3) and (3) a TFBS motif and TF expression filter (see Section 3.1.2.2). pSNVs were thus annotated with features such as recurrence rate, tumor purity, variant allele frequency, presence within the CGC list, and open chromatin (see Section 3.1.3). Using these features combined with each pSNV's genomic and transcriptomic information, a prioritization score for their subsequent ranking was then computed (see Section 3.1.4). To gain a holistic view of individual pSNVs and to assess their validity, *pSNV Hunter* (see Section 3.1.7) was used to assist in the manual selection of mutations for subsequent *in vitro* validation.

Consequently, this section will present the results of applying this workflow to both the PCAWG and *retrospective* NCT-MASTER dataset separately (see Sections 5.2 and 5.3). Through the additional analysis of these results with *pSNV Hunter* (see Section 5.4), 22 pSNVs were selected and validated *in vitro* (see Section 5.5). Additionally, analyses on UV-induced mutational signatures (see Section 5.6) and TF activity (see Section 5.7) were conducted retrospectively. Finally, the viability of this workflow in a precision oncology setting was assessed through the use of the *prospective* NCT-MASTER dataset as well as the COGNITION dataset (See Section 5.8).

5.2 The REMIND-Cancer Workflow on the PCAWG Dataset

The REMIND-Cancer workflow was first applied to the PCAWG dataset, which consisted of 2,413 unique samples comprising 19,401,901 total mutations. 927 (38.4%) of these samples were accompanied by RNA-seq data, allowing for the analysis of their corresponding 10,924,597 mutations (56.3% of all mutations). These 927 samples belonged to 24 of the 43 original cohorts as 19 cohorts did not have corresponding RNA-Seq data.

366,373 mutations (3.4% of all SNVs) were classified as pSNVs through the promoter filter. Furthermore, 19,250 (0.18% of all SNVs; 5% of all pSNVs) met the remaining filtering criteria, which included the gene expression filter as well as the TFBS motif and TF expression filter. These remaining 19,250 pSNVs were then annotated with additional features, scored and subsequently ranked (Figure 18).

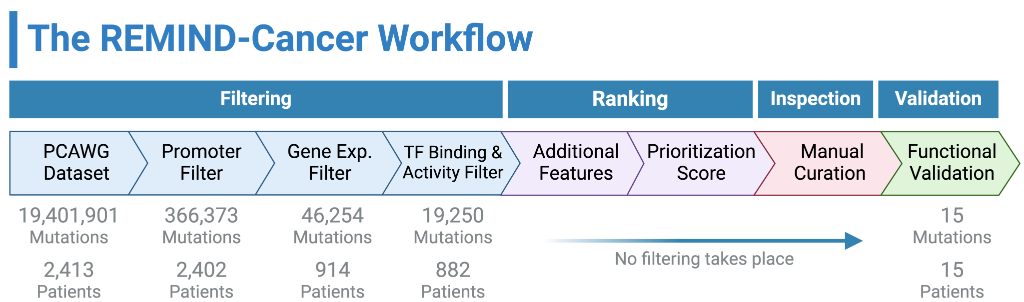

Figure 18: *Results of the REMIND-Cancer workflow on the PCAWG dataset. Created with* `BioRender.com`. *Adapted from Abad et al. (Abad, Glas, et al., 2024).*

5.2.1 Prevalence of Recurrent and Known Functional pSNVs

Of the 19,250 pSNVs passing the pipeline, only 966 (5%) were recurrent, many of which were identified in previous studies such as $TERT_{C228T}$, $TERT_{C250T}$, $RALY_{C927T}$ and $CDC20_{G529A}$.

28 out of the 97 total $TERT_{C228T}$ instances and 7 out of the 36 total $TERT_{C250T}$ instances remainined after the conclusion of the pipeline, making them the two most frequent genomic positions after the conclusion of the pipeline (Figure 19). These 28 $TERT_{C228T}$ mutations were distributed across 18 cohorts whereas the $TERT_{C250T}$ pSNVs were distributed throughout 6 cohorts. Conversely, a majority of both *TERT* mutations that did not pass the pipeline were initially excluded due to the lack of available gene expression data.

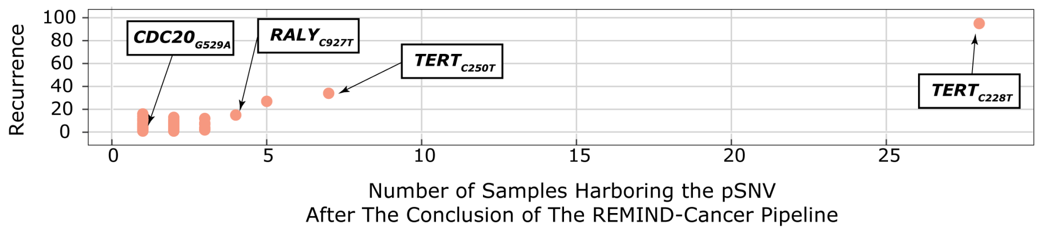

Figure 19: *Scatter plot showing the most frequent genomic positions of pSNVs that passed the REMIND-Cancer pipeline. Four previously-implicated pSNVs, namely* $TERT_{C228T}$, $TERT_{C250T}$, $RALY_{C927T}$ *and* $CDC20_{G529A}$, *can be seen.*

The previously-implicated $RALY_{C927T}$ pSNV also stood out as the fourth most common pSNV after the conclusion of the pipeline (Figure 19). Of the 15 total $RALY_{C927T}$ occurrences, four samples passed the pipeline, all from the skin cutaneous melanoma (SKCM-US) cohort. Interestingly, the 11 other $RALY_{C927T}$ pSNVs that did not pass the pipeline were observed either within the same SKCM-US cohort or within the broader Australian melanoma cohort (MELA-AU), which lacked corresponding RNA-Seq data. By definition, the MELA-AU cohort included acral and mucosal melanoma subtypes, in addition to the cutaneous melanoma subtype present in the SKCM-US cohort.

Despite having a relatively high recurrence rate of 11, only one $CDC20_{G529A}$ pSNV passed the pipeline (Figure 19). This pSNV was detected in one of the six samples with matching RNA-Seq data, suggesting that *CDC20* expression was not above the cohort average in the other five samples, thereby failing the gene expression filter. Similar to $RALY_{C927T}$, this mutation was also found within the SKCM-US cohort.

Even when considering only the top 100 highest-ranking sample-specific pSNVs, 35 of the 54 recurrent mutations belonged to one of these previously known functional pSNVs. In particular, 26 instances of $TERT_{C228T}$, 7 instances of $TERT_{C250T}$, and 2 instances of $RALY_{C927T}$ (Figure 19) were found within the top 100. While a $CDC20_{G529A}$ pSNV did not rank within the top 100, it still achieved a rank of 284, placing it within the top 1.5% of all remaining pSNVs.

A detailed breakdown of the highest-ranking sample-specific scores for each of these four previously-implicated pSNVs is presented in Figure 20a.

In these four sunburst charts, the genomic (green), transcriptomic (orange) and annotation (blue) categories are further divided into the individual components that contribute to their prioritization scores.

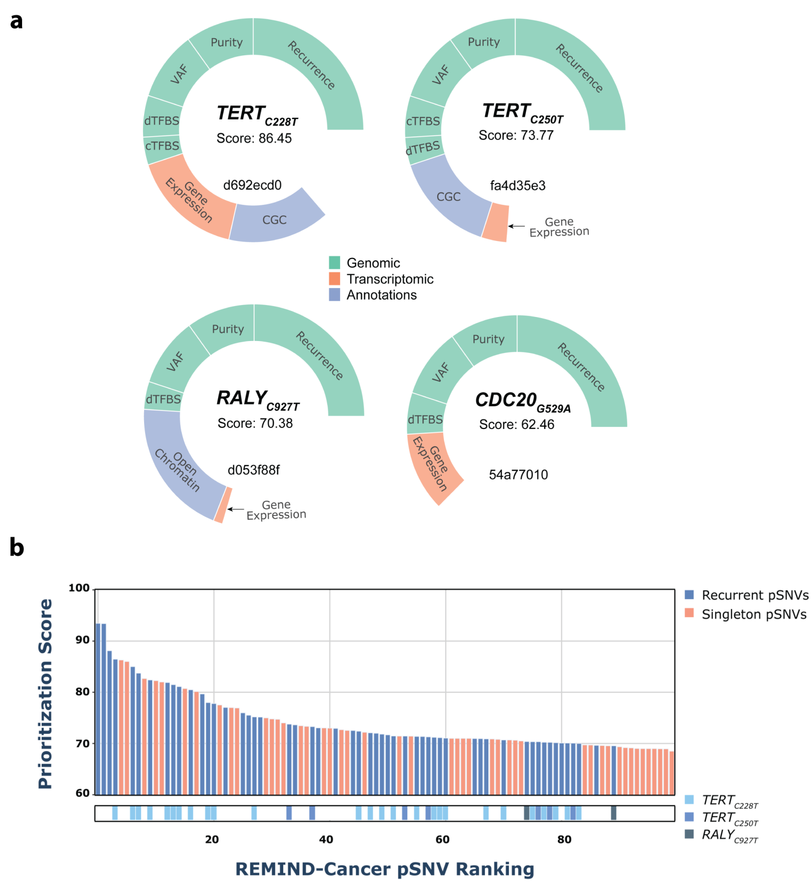

Figure 20: *Overview of four previously-identified pSNVs and the top 100 ranking pSNVs. (a) Sunburst charts displaying the highest-scoring samples for* $TERT_{C228T}$*,* $TERT_{C250T}$*,* $RALY_{C927T}$*, and* $CDC20_{G529A}$*. The general categories (genomic in green, transcriptomic in orange, and annotations in blue) are further detailed by their specific components. cTFBS = number of created TFBSs, dTFBS = number of destroyed TFBSs, CGC = known cancer gene according to the CGC database, VAF = variant allele frequency. (b) Top 100 ranking pSNVs separated by recurrence status (top) and a heatmap corresponding to the* $TERT_{C228T}$*,* $TERT_{C250T}$*, and* $RALY_{C927T}$ *(bottom). Created with* `BioRender.com`. Adapted from Abad et al. (Abad, Glas, et al., 2024).

Furthermore, of the 19,250 sample-specific pSNVs passing the pipeline, 18,284 (95%) were singletons with 46 being within the top 100 ranking mutations (Figure 20b). These singletons were primarily within the SKCM-US cohort with 15 instances within the top 100 followed by bladder urothelial cancer (BLCA-US) with 5 and 4 within both liver cancer (LIRI-JP) and lung squamous cell carcinoma (LUSC-US).

5.2.2 86% of Recurrent pSNVs Passing the Pipeline Still Lack Statistical Power

As introduced in Section 1.4.2, Rheinbay et al. (Rheinbay et al., 2020) investigated the number of pSNVs needed in each cohort to achieve 90% statistical power for its detection, given each cohort's observed mutation rate and overall sample size. Among all cohorts, the fewest instances required to reach sufficient power was four, which was observed in the myeloproliferative neoplasms cohort, the general myeloid cohort, and the central nervous system pilocytic astrocytomas cohort, as shown in Figure 3.

Using four as a lenient threshold for achieving statistical power, 829 of the 966 (86%) remaining recurrent pSNVs did not exceed this threshold. In particular, 612 (73.8%) were recurrent with only one other sample, 138 (16.6%) with two other samples and 79 (9.5%) with three other samples. Though technically recurrent, these pSNVs could be considered lowly-recurrent, implying that they would not have the statistical power needed to be detectable by pre-existing recurrence-based methods.

Conversely, of the 137 sample-specific pSNVs passing this threshold of four, 40 (29%) corresponded to previously-known pSNVs such as $TERT_{C228T}$ (n=28), $TERT_{C250T}$ (n=7), $RALY_{C927T}$ (n=4) and $CDC20_{G529A}$ (n=1). The remaining 97 highly-recurrent pSNVs corresponded to 65 genomic positions, thereby leaving a substantial blindspot in the detection of functional pSNVs.

5.2.3 $ANKRD53_{G529A}$ is the highest and third highest ranking pSNV

The recurrent $ANKRD53_{G529A}$ pSNV was identified as achieving the highest and third highest ranking pSNV of all PCAWG mutations. Irrespective of RNA-Seq data availability, this mutation was present within six total samples, two of which passed the pipeline and the other four of which did not have RNA-Seq data available. In particular, the highest-ranking pSNV

$ANKRD53_{G529A}$ was found in the BLCA-US cohort whereas the third highest ranking pSNV $ANKRD53_{G529A}$ was found in lung adenocarcinoma (LUAD-US) (Figure 21a).

At the transcriptomic level, both samples harboring this mutation exhibited exceptionally high zscores relative to their cohort with the top ranking mutation displaying a zscore of 4.49 (Figure 21b) whereas the third ranking mutation had a zscore of 3.42. Additionally, this G/C>A/T pSNV was predicted to create two new highly conserved binding sites: one for RELA (Figure 21c) and the other for the STAT1:STAT2 heterodimer. This pSNV was also located within a region of open chromatin according to ChromHMM (Ernst & Kellis, 2012) but this gene is not yet a known cancer gene according to the CGC (Sondka et al., 2018) database.

Furthermore, a *GenomeTornadoPlot* was generated in order to visually assess the focal amplifications and deletions of this particular gene relative to the PCAWG dataset (Figure 21d), revealing a high number of focal amplifications throughout the entire cohort. Given that amplification could contribute to increased expression levels of individual samples and considering that the highest and third highest ranking $ANKRD53_{G529A}$ pSNVs displayed relatively normal copy number levels of 3, this may suggest a case of convergent tumor evolution. This pSNV could potentially upregulate *ANKRD53* in the abscence of an increased copy number level thereby reinforcing my interest in validating this pSNV *in vitro*.

Additionally, due to being highly prioritized in two samples, I wanted to ensure this position's biological and technical validity by employing *DeepPileup* (Figure 21e and 21f). These plots revealed that six PCAWG cohorts had tumor samples with a minor allele frequency (MAF) greater than 25% while no normal samples showed noise (Figure 21e). All other PCAWG cohorts not displayed have undifferentiated samples between the normal and tumor samples by showing an MAF of 0%. Notably, both of the cohorts (BLCA-US and LUAD-US) in which $ANKRD53_{G529A}$ was identified in, showed a positive signal, reinforcing the confidence in this position. Using a more sensitive threshold of at least two variant alleles, additional signal from 7 more cohorts were detected, which may indicate subclonal mutations (Figure 21f) as this mutation is only present in a subset of all cohorts.

Though this specific $ANKRD53_{G529A}$ pSNV was identified as the highest ranked and third highest ranked mutation after applying the REMIND-Cancer workflow to the PCAWG dataset, Irina Glas (Abad, Glas, et al., 2024) also discovered this mutation prior to the start of my PhD. To as-

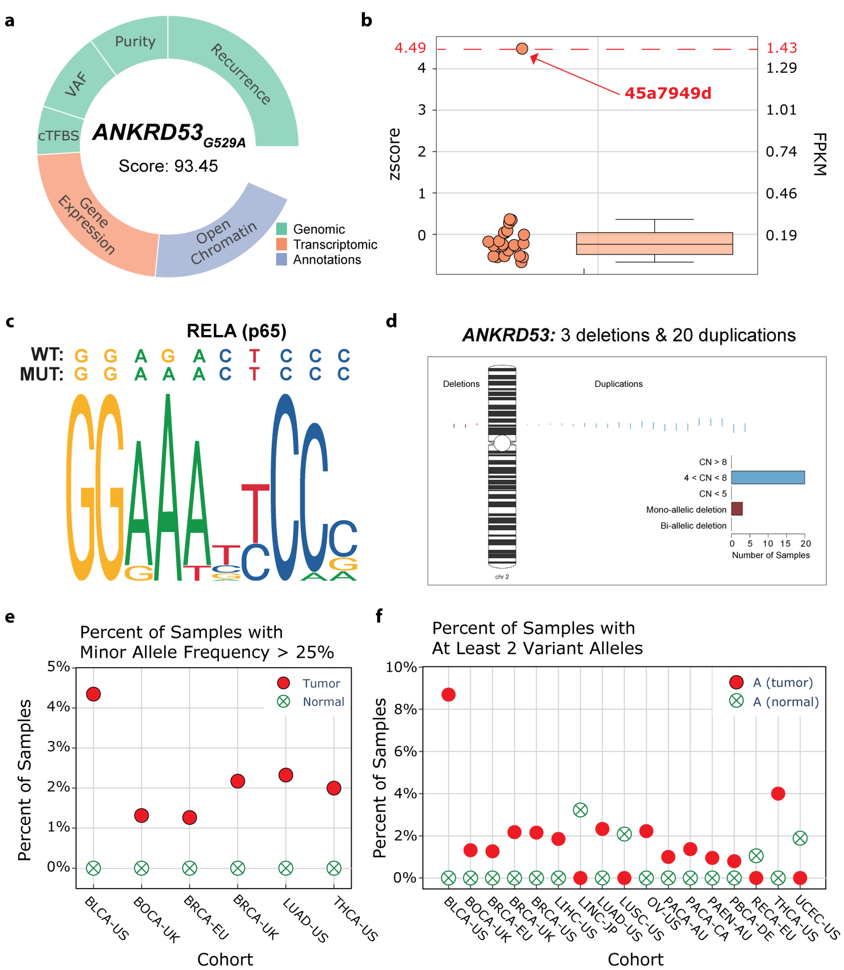

Figure 21: *An overview of ANKRD53$_{G529A}$, which corresponds to the highest and third highest ranking pSNV. (a) Scoring breakdown of the highest scoring pSNV, which is found within the BLCA-US cohort (b) The zscore (left y-axis) and FPKM (right y-axis) of the highest scoring sample harboring this mutation. The red dashed line represents the sample's (relative) expression value. (c) Sequence logo plot of the TFBS that this pSNV is predicted to create through the G¿A mutation. (d) GenomeTornadoPlot of ANKRD53 (e) A DeepPileup plot displaying the percentage of samples with a minor allele frequency above 25% for each cohort. Here, six cohorts, including that of BLCA-US, are shown to have no signal within control samples but signal coming from their tumor samples. (f) Using a less stringent criteria, a secondary DeepPileup plot displaying the percentage of samples with at least two variant alleles. Adapted from Abad et al. (Abad, Glas, et al., 2024).*

sess its functionality *in vitro*, this mutation was introduced into a luciferase reporter assay using the kidney cell line HEK293FT in which a slight 20% increase in promoter activity was observed when comparing the mutant to the wildtype construct.

Given that *RELA* encodes the NF-κB subunit p65, which requires post-translational activation, Irina conducted further experiments to investigate the role of p65. Upon activating this TF with TNF-α, promoter activity increased by 80% compared to the original WT construct, demonstrating the functionality of this TF, specifically within this context. These findings collectively suggest that the REMIND-Cancer workflow is effective in prioritizing pSNVs that significantly contribute to activation, although their full impact may depend on additional factors.

5.2.4 Applying the REMIND-Cancer Workflow to the EOPC-DE Cohort Identifies and Prioritizes $\boldsymbol{MYB_{C964A}}$

As described in Section 3.1.1.2, the integration of the publicly-available early onset prostate cancer (EOPC-DE) WGS data with an internal version of each sample's corresponding RNA-Seq data resulted in a cohort comprising 56,608 SNVs from 23 unique sample. To assess its effectiveness on a singular cohort, the REMIND-Cancer pipeline was applied while querying recurrence against the aforementioned PCAWG dataset as well as other EOPC-DE samples.

The REMIND-Cancer pipeline identified 554 (0.98%) SNVs as being within promoter regions with 54 pSNVs (0.01% of all SNVs; 9.7% of all pSNVs) from 17 samples successfully passing the entire pipeline (Figure 22a).

Among these 54 pSNVs, 53 were identified as singletons when compared across the entire PCAWG dataset, including the highest-ranking pSNV MYB_{C964A}, where *MYB* has been recognized as a proto-oncogene. Notably, the sample harboring MYB_{C964A} displayed exceptionally high *MYB* transcription (z-score of 3.75) (Figure 22b) and was located in a region of open chromatin. Additionally, this pSNV was predicted to create binding sites for two forkhead box (FOX) transcription factors, FOXD1 and FOXO3, both of which showed exceptionally high FPKM values (Figure 22c).

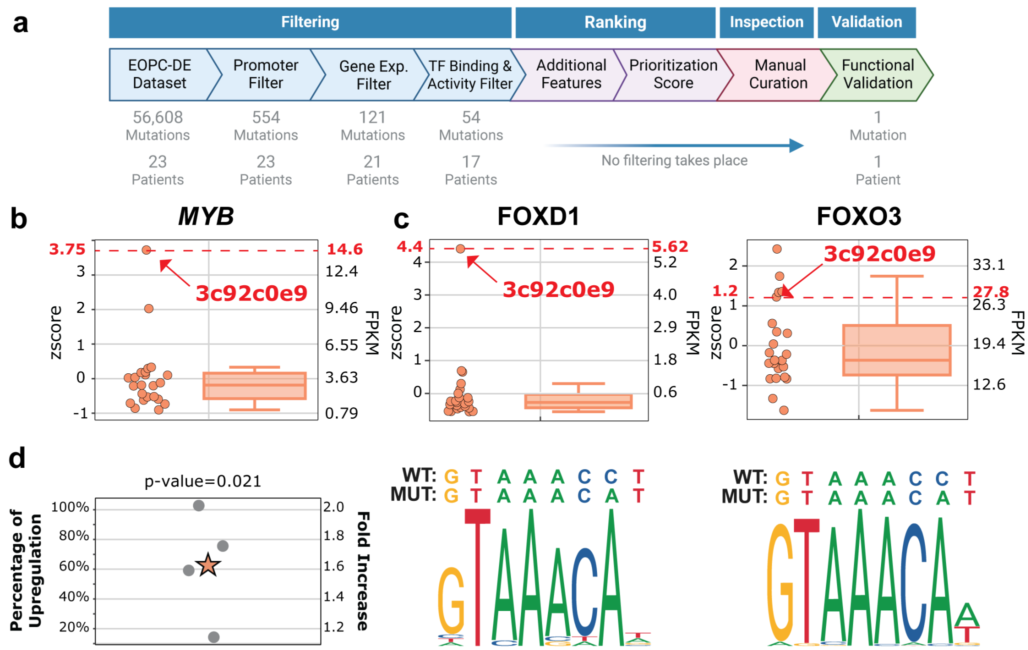

Figure 22: *Overview of the results from applying the REMIND-Cancer workflow to the EOPC-DE cohort. (a) Number of mutations and samples per filtering and annotation step. Created with* `BioRender.com`*. (b) The zscore (left y-axis) and FPKM (right y-axis) of MYB. (c) The zscore (left y-axis) and FPKM (right y-axis) of FOXD1 (left) and FOXO3 (right). Their corresponding logo plots can be seen below the boxplots. (d) The results of the in vitro validation showing a 63% increase in activity (p-value = 0.021; one-sided t-test). Adapted from Abad et al. (Abad, Glas, et al., 2024).*

Similar to $ANKRD53_{G529A}$, MYB_{C964A} was validated *in vitro* via luciferase assays prior to the start of my PhD by Dr. Dieter Wiechenhan (Abad, Glas, et al., 2024). In these experiments, using the kidney cell line HEK293FT, a statistically significant 63% increase (Figure 22d; one-sided t-test; p-value = 0.021) in luciferase activity was observed between the WT and MUT constructs, further supporting the effectiveness of my approach in identifying and characterizing putative activating pSNVs.

5.3 The REMIND-Cancer Workflow on the Retrospective NCT-MASTER Dataset

In addition to applying the REMIND-Cancer workflow to the PCAWG dataset, this was also applied separately to a subset of the NCT-MASTER dataset. As this program is ongoing, I created a data freeze, which includes data from November 2016 to October 2020, corresponding to 2,378 samples comprising 40,660,325 SNVs. As detailed in Section 4.4, both the "Anderes / Other" and "Sarkom / Sarcoma" cohort were omitted as well as samples originating from the metastatic site (Figure 23a).

In doing so, a total of 10,722,278 mutations from 440 samples were used as input into the REMIND-Cancer pipeline. Through the three filtering steps, namely the (1) promoter filter, (2) TFBS motif and TF expression filter and the (3) gene expression filter, this resulted in the scoring and ranking of 6,274 putative pSNVs within 328 patients (Figure 23b), which were spread throughout 15 distinct cohorts.

5.3.1 Not All Previously-Implicated pSNVs Could Be Analyzed

Within Section 1.5.1, 5 previously-implicated pSNVs were introduced, namely $TERT_{C228T}$, $TERT_{C250T}$, $CDC20_{G529A}$, $LEPROTL1_{C921T}$, and $RALY_{C927T}$. However, only $TERT_{C228T}$ and $LEPROTL1_{C921T}$ were among the prioritized pSNVs when applying the REMIND-Cancer workflow to the *retrospective* NCT-MASTER dataset.

For the $TERT_{C228T}$ hotspot mutation, only 4 of the 151 instances of this pSNV remained after the conclusion of the pipeline. However, this was primarily due to these pSNVs belonging to one of the excluded cohorts (n = 107; 70.9%) or not having corresponding RNA-Seq data (n = 21; 13.9%), which automatically excluded 128 (85%) of all $TERT_{C228T}$ pSNVs within the dataset. Surprisingly, of those 23 $TERT_{C228T}$ pSNVs that were analyzable, 19 of those samples displayed a relatively negative *TERT* expression (zscore ≤ 0), which resulted in only 4 instances passing the pipeline. However, these four pSNVs were all highly ranked (position 6, 42, 79 and 106), placing them within the top 2% of all remaining pSNVs.

Secondly, no instance of $TERT_{C250T}$ remained after the completion of the pipeline despite this pSNV being observed in 34 samples. However, similar to that of $TERT_{C228T}$, many samples harboring $TERT_{C250T}$ were excluded due to not having an analyzable cohort (n = 22; 65%) or not having RNA-Seq

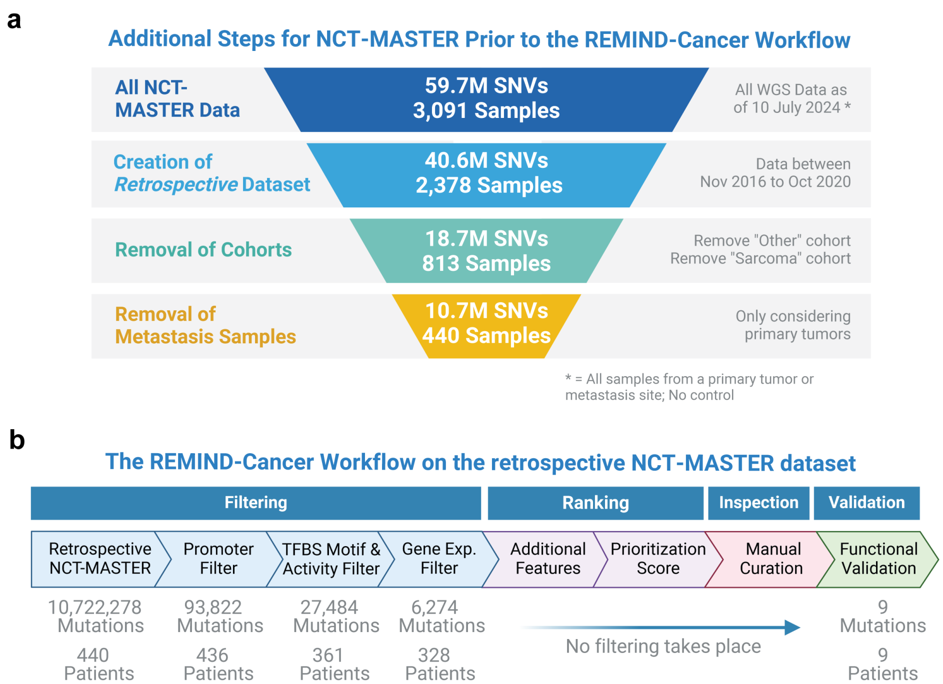

Figure 23: *Overview of the retrospective NCT-MASTER dataset: (a) In comparison to the PCAWG dataset, the NCT-MASTER dataset required additional data preprocessing steps such as the creation of the retrospective dataset, the removal of the "Anderes / Other" and "Sarkom / Sarcoma" cohorts, and the removal of metastasis samples. (b) Results of applying the REMIND-Cancer Workflow to the NCT-MASTER dataset. Created with* `BioRender.com`.

data available (n = 6; 18%). The remaining 6 analyzable $TERT_{C250T}$ pSNVs had slightly negative normalized expression values and therefore did not pass the pipeline.

Furthermore, when considering the other three known pSNVs, only one instance of $LEPROTL1_{C921T}$ was prioritized, which was found within the Head and Neck cohort. Although only a slight upregulation in its associated gene (zscore of 0.6) was observed, this pSNV was observed within four other samples (in addition to being recurrent with one sample within the PCAWG dataset), located within a region of open chromatin and was annotated as being a known cancer gene according to the CGC list. Additionally, $LEPROTL1_{C921T}$ was predicted to create a single binding site for the EWSR1::FLI1 TF fusion, which has been identified as being a possible therapeutic target within sarcoma (Mo et al., 2023).

5.3.2 81 of the Top 100 Ranking pSNVs Did Not Achieve Sufficient Statistical Power

Of all 6,274 pSNVs passing the REMIND-Cancer pipeline, 5,921 (94.3%) were singletons and 270 (4.3%) were lowly-recurrent with a recurrence statistic less than 4. Conversely, this small subset of 83 (1.3%) highly-recurrent pSNVs only occurred within 36 unique genomic positions, further adding to the ineffectiveness of recurrence-based approaches.

Moreover, when focusing on only the top 100 ranking mutations, 73 were singletons and 8 were lowly-recurrent. Of these singletons, however, was the highly-ranked pSNV $SCN1B_{C113T}$, which was found within the skin cohort and ranked at the 93rd position placing it within the top 1.5% of all remaining pSNVs. Interestingly, though prior studies have associated *high SCN1B* expression in normal tissues rather than cancer tissues (Bon et al., 2016), this particular skin sample displayed high expression with a zscore of 1.4.

5.3.3 *pSNV Hunter* Assists in Revealing Other Non-Coding Elements

After the conclusion of the REMIND-Cancer pipeline, the highest ranking singleton, which was observed at the fourth highest ranking position, was identified within the promoter of *RP11-672L10.3*. Furthermore, other high ranking mutations were found within the promoters of *RP11-672L10.3* (singleton; rank 10), *RP11-532F12.5* (recurrence of 4; rank 17), *RP11-296L22.8* (singleton; rank 22), *AC097381.1* (recurrence of 7; rank 30), *RP11-566K11.4* (singleton; rank 42), *RP11-230C9.2* (singleton; rank 47), and *AC005625.1* (singleton; rank 48).

However, upon the manual inspection of these mutations with *pSNV Hunter*, which displays gene information originally provided by GeneCards (Stelzer et al., 2016) and the National Center for Biotechnology Information (NCBI) (Brown et al., 2015), it was determined that these mutations occurred in promoter regions of genes that produce long non-coding RNAs (lncRNAs) rather than protein-coding genes. Although lncRNAs have been shown to play a causal role in cancer progression (Carlevaro-Fita et al., 2020), the primary focus of my thesis was to identify pSNVs that may have a direct and downstream impact on protein production, particularly for clinical applications. Because lncRNAs cannot be translated into proteins, these identified pSNVs were thus not considered for further *in vitro* validation, which underscores the necessity of using *pSNV Hunter* as a diagnostic tool before conducting *in vitro* validation.

5.4 Selecting pSNV Candidates for *in vitro* Validation with *pSNV Hunter*

The REMIND-Cancer computational pipeline significantly reduces the number of putative pSNVs through its filtering process, yet a substantial number still remain. In particular, the PCAWG dataset was reduced from 19.4 million mutations to 19,250, a 99.9% decrease, and the NCT-MASTER dataset was similarly reduced by 99.94% from approximately 10.7 million to 6,274 mutations. Despite these reductions, manually inspecting 19,250 + 6,274 = 25,524 pSNVs to ensure their validity prior to *in vitro* validation is time-consuming and impractical, even given the assistance of our prioritization score. To address this challenge, I developed *pSNV Hunter* (Abad, Körner, & Feuerbach, 2024c), a multi-purpose data aggregation and visualization tool, designed to streamline the analysis of pSNVs from the REMIND-Cancer pipeline and other VCF-formatted files.

While all features of the *pSNV Hunter* software are detailed in Section 3.1.7, four key functionalities expedited the *in vitro* selection process: (1) displaying comprehensive genomic and transcriptomic information for each pSNV in a tabular format, (2) visualizing gene expression plots for pSNV-affected genes alongside the expression in recurrent samples and known gene functions, (3) identifying predicted creation or destruction of TFBSs, and (4) viewing quality control plots such as *DeepPileup*. An example of how this was applied to $ANKRD53_{G529A}$ can be seen in Figure 24.

5.5 *In vitro* Validation

Utilizing *pSNV Hunter*, 22 pSNVs (13 from PCAWG; 9 from retrospective NCT-MASTER) were identified, prioritized, and selected for *in vitro* validation (Supplementary Table S1). Of these mutations, 13 were singletons and 9 were recurrent. Since singletons make up approximately 95% of the remaining pSNVs within both the PCAWG and NCT-MASTER datasets,

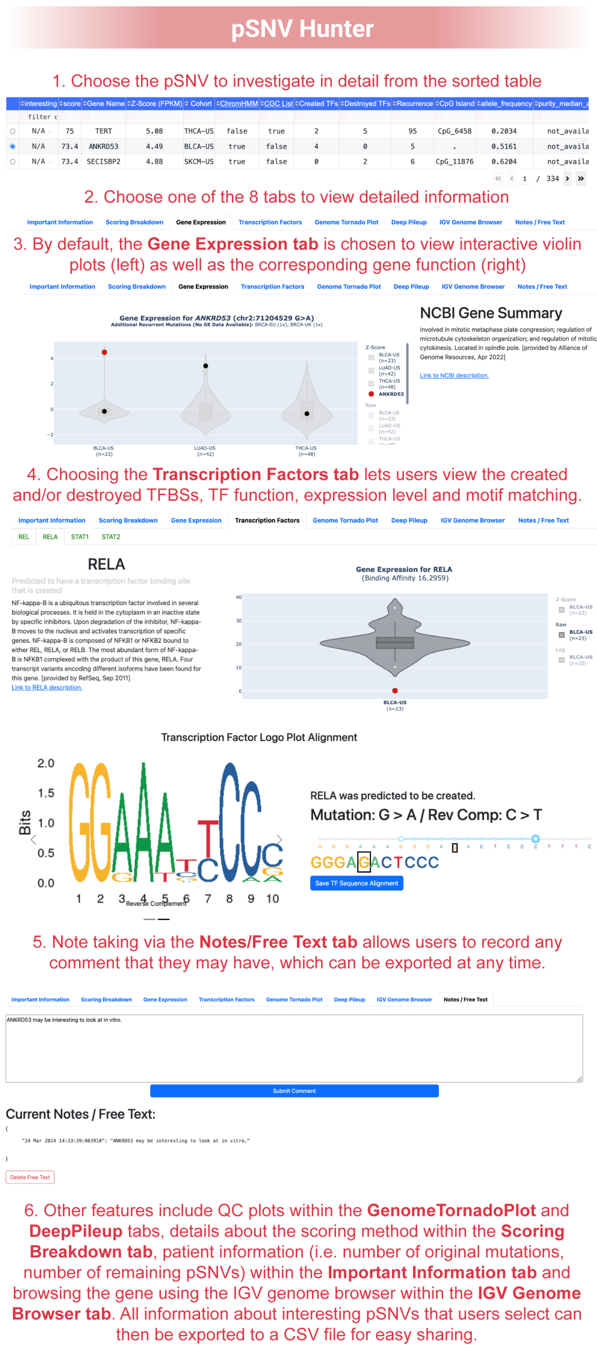

Figure 24: *Screenshots of the pSNV Hunter workflow along with several features using* $ANKRD53_{G529A}$ *as an example.*

there was a deliberate effort to select and validate these often-overlooked pSNVs to test the ability of my pipeline to not only detect singletons but rather detect *activating* singletons.

Additionally, the validation efforts also included the previously-implicated $CDC20_{G529A}$ (He et al., 2021) (Godoy et al., 2023), $RALY_{C927T}$ (Hayward et al., 2017) and $LEPROTL1_{C921T}$ (Rheinbay et al., 2017). Although the selected candidates originated from 8 different cancer types (4 from PCAWG and 4 from NCT-MASTER), the most common was the skin cancer cutaneous melanoma (SKCM-US), which 10 pSNVs belonged to, followed by Skin (n=3).

While no absolute ranking or molecular criteria was used in the selection process, pSNVs generally shared the following characteristics: high relative expression, presence of active TFs, location within regions of open chromatin, and positioning within high-confidence regions without noisy quality control plots. However, these mutations were not exclusively associated with known cancer genes according to the CGC list and did not necessarily have extensive research on their cancer associations.

Although I identified, prioritized and selected these candidate pSNVs, all *in vitro* validation efforts were gratefully conducted by Sabine Karolus and Dr. Cindy Körner of DKFZ's Division of Molecular Genome Analysis led by Prof. Dr. Stefan Wiemann.

5.5.1 10 pSNVs (Including $ANKRD53_{G529A}$ and MYB_{C964A}) Lead to An Increase In Promoter Activity

Of the 22 pSNVs selected using *pSNV Hunter*, eight showed a significant increase (p-value ≤ 0.05; one-sided t-test) in promoter activity when comparing the wild type (WT) to the mutant (MUT) construct (Figure 25a and 25b). Through the additional inclusion of both the positively-validated recurrent $ANKRD53_{G529A}$ pSNV (see Section 5.2.3) as well as the MYB_{C964A} singleton (see Section 5.2.4), 10 (10 of 24; 42%) pSNVs were found to lead to an increase in promoter activity in total.

7 positively validated pSNVs were from the PCAWG dataset whereas the remaining three were found within the NCT-MASTER dataset. Notably, the validation rate of pSNVs selected from the PCAWG dataset (47%; 7 of 15) is higher than that of the retrospective NCT-MASTER dataset (33%; 3 of 9). These findings suggest that the REMIND-Cancer workflow, which includes the manual inspection phase via *pSNV Hunter*, is effective in identifying

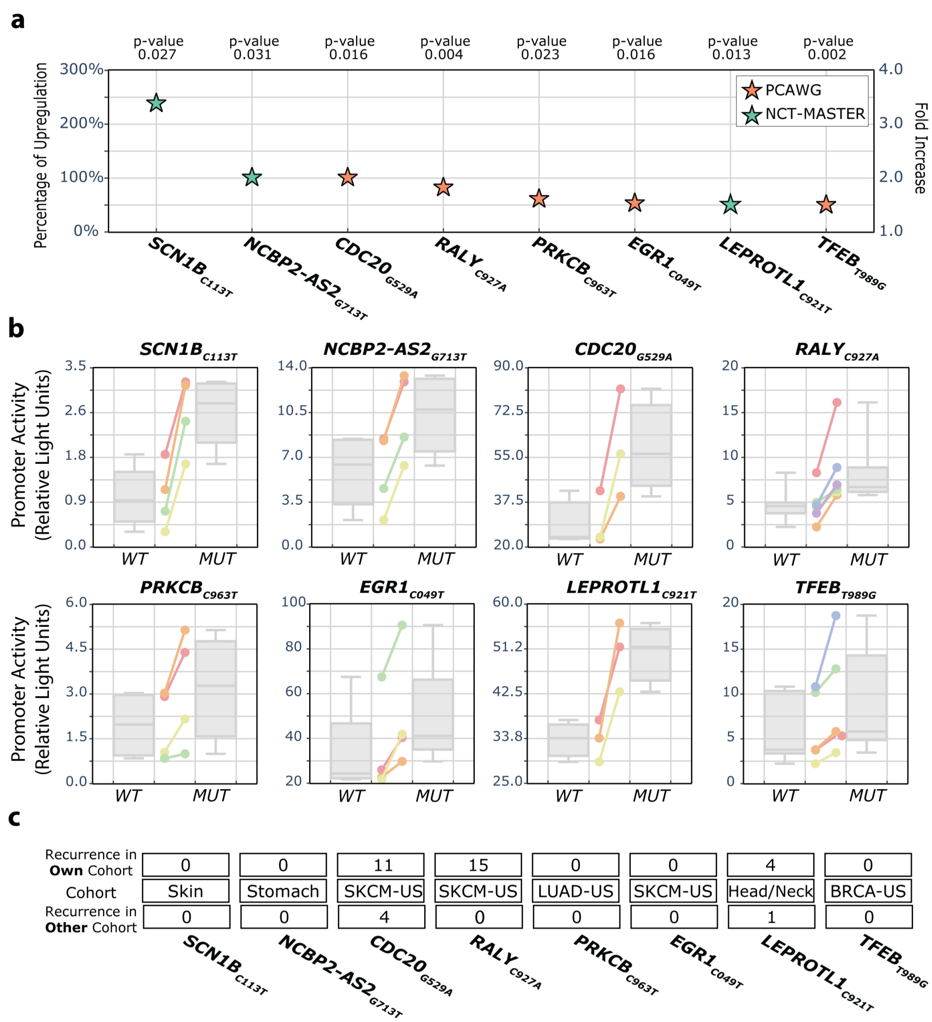

Figure 25: *Overview of the eight pSNVs that lead to an increase in promoter activity. (a) Scatter plot of the average percentage of upregulation (left y-axis) and fold increase (right y-axis) for each positively validated pSNV. pSNVs found within the PCAWG dataset are denoted in green whereas those found within the NCT-MASTER dataset are seen in orange. Adapted from Abad et al. (Abad, Glas, et al., 2024). (b) The promoter activity, measured in relative light units, of each positively validated pSNV. Each WT-MUT replicate pair can be seen in the different colors within each individual plot. (c) The cohort and recurrence rate relative to their own dataset as well as the other dataset. For example, $CDC20_{G529A}$, which was originally found within PCAWG, has a recurrence rate of 11 in the PCAWG dataset, though it was still found within four samples within the NCT-MASTER dataset.*

activating pSNVs within clinical datasets as well as highly-curated datasets such as PCAWG.

Furthermore, among the 10 positively validated pSNVs were the previously-implicated $CDC20_{G529A}$, $RALY_{C927T}$, and $LEPROTL1_{C921T}$ pSNVs. The activating result of $CDC20_{G529A}$ result aligns with He et al. (He et al., 2021), who observed an increase in promoter activity for $CDC20_{G529A}$, but contradicts Godoy et al. (Godoy et al., 2023), who found a repressive effect. Furthermore, while $RALY_{C927T}$ was previously identified and implicated within the cancer type (Hayward et al., 2017), my study extends this finding by functionally validating this pSNV *in vitro* and demonstrating its activating functionality. Lastly, although $LEPROTL1_{C921T}$ was previously validated in the same HEK293FT cell line by Rheinbay et al. (Rheinbay et al., 2017), these results indicate a statistically significant *increase* in promoter activity, contradicting their prior findings in which a *negative* effect was found. Notably, these three pSNVs were also recurrent in both datasets (Figure 25c).

Furthermore, the positive validation rate of recurrent mutations (40%; 4 of 10) was similar to that of singletons (43%; 6 of 14), suggesting that the REMIND-Cancer workflow is equally effective at identifying pSNVs with activating functional significance regardless of their recurrence status.

5.6 Mutational Signatures

5.6.1 The SKCM-US PCAWG Cohort Shows an Abundance of SBS7 Mutations

In prior studies, the previously-identified $CDC20_{G529A}$ (He et al., 2021)(Godoy et al., 2023), $RALY_{C927T}$ (Hayward et al., 2017), $TERT_{C228T}$ (Horn et al., 2013) (F. W. Huang et al., 2013) and $TERT_{C250T}$ (Horn et al., 2013) (F. W. Huang et al., 2013) were all identified within a melanoma sample. Furthermore, in addition to $RALY_{C927T}$ and $CDC20_{G529A}$, 9 additional pSNVs were identified, prioritized and thus validated - all of which also were found within a melanoma sample, particularly that of skin cutaneous melanoma (SKCM-US).

Upon analysis of this cohort, I observed that two mutational signatures provided by COSMIC (Alexandrov et al., 2013), particularly SBS7a and SBS7b, were distinctly overrepresented when considering both SNVs and pSNVs in comparison to all non-melanoma cohorts, matching previous studies (Hayward et al., 2017) (Figure 26a).

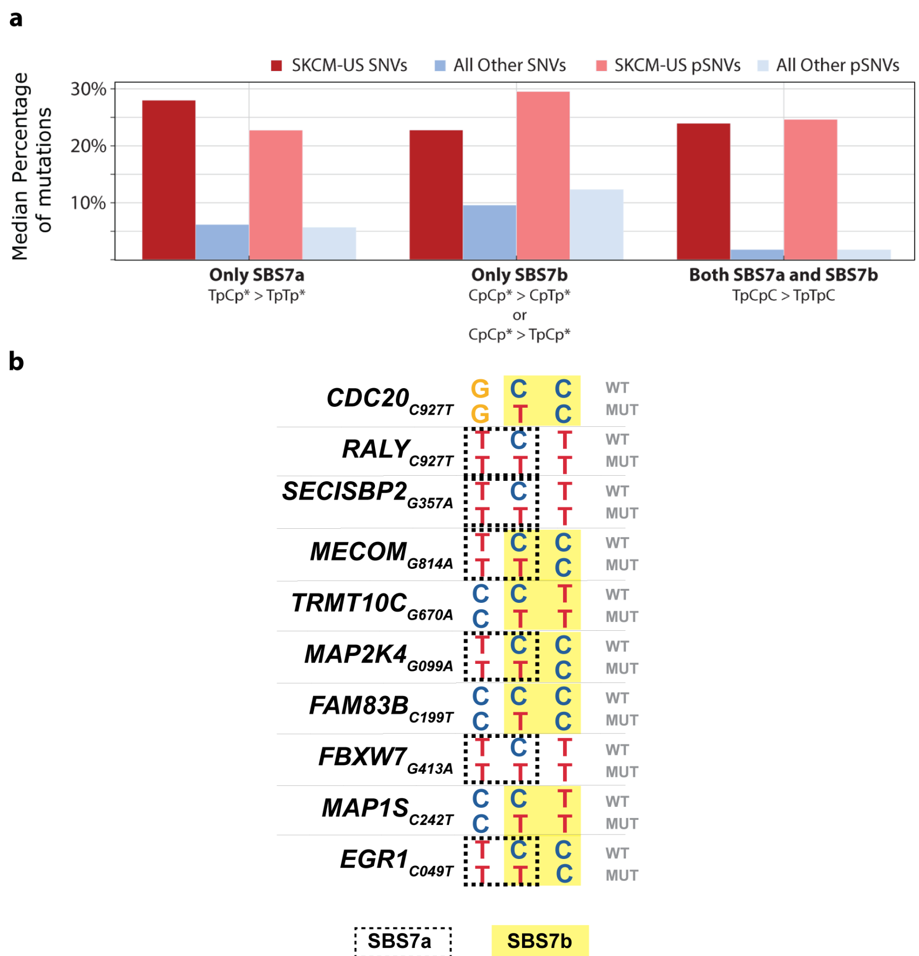

Figure 26: *Overview of the presence of SBS7a and SBS7b (a) A comparison of the median percentage of mutations within all samples (y-axis) belonging to the SBS7a and SBS7b mutational signatures (x-axis) between melanoma and non-melanoma SNVs and pSNVs. Below each mutational signature, its respective trinucleotide context (i.e. TpCp* ¿ TpTp*) is displayed; the star (*) represents any of the four nucleotides (i.e. TpCp* implies that this can be TpCpA, TpCpC, TpCpT, or TpCpG). (b) The presence of SBS7a (highlighted in yellow) and SBS7b (outlined with a black dashed line) in the 11 validated SKCM-US pSNVs. For each pSNV, the top sequence represents the WT, while the bottom sequence represents the MUT.*

As described in Section 3.1.10, SBS7a is considered to be any TpC to TpT (i.e. TC > TT) mutation whereas SBS7b is considered to be any CpC to CpT (i.e. CC > CT) or CpC to TpC (i.e. CC > TC) mutation. Considering that mutational signatures are typically displayed in a trinucleotide context, a mutation is considered to have this mutational signature regardless of the flanking nucleotide using a star (*) notation (e.g. TpCp* represents TpCpA, TpCpC, TpCpG and TpCpT; *CpT represents ApCpT, CpCpT, GpCpT, TpCpT). Moreover, of the 10 pSNVs belonging to the SKCM-US cohort and were validated within this study, all contained at least one of SBS7a and SBS7b (Figure 26b).

5.7 Transcription Factor Activity Prediction

5.7.1 DoRothEA, Collectri and decoupleR Are Unable To Predict Activating TFs Of *TERT* and *ANKRD53*

To retrospectively assess the effectiveness of using solely RNA-Seq data to predict TF activity as described in Section 3.1.11, I individually applied four decoupleR (Badia-i Mompel et al., 2022) methods, namely ULM, WSUM, VIPER and consensus, to two gene-TF interaction databases, namely DoRothEA and Collectri on pSNVs from the PCAWG dataset. More specifically, these eight method-database combinations (e.g. ULM and DoRothEA, WSUM and Collectri, etc.) were applied to all $TERT_{C228T}$ pSNVs and the two high-ranking $ANKRD_{G529A}$ pSNVs from the PCAWG dataset.

To recapitulate prior findings, it has been demonstrated that both $TERT_{C228T}$ and $TERT_{C250T}$ result in the creation of a binding site for the ETS-family transcription factor GABPA. Similarly, previous experiments (Abad, Glas, et al., 2024) found that $ANKRD_{G529A}$ creates a binding site for RELA (p65) (see Section 5.2.3). In both pSNVs, the TF binding and subsequent TF activity lead to the upregulation of the pSNV's associated gene.

I first applied these eight method-database combinations separately to the cohort-based RNA-Seq data of the highest ranking and third highest ranking PCAWG pSNVs, which both correspond to $ANKRD53_{G529A}$. These two were found within the BLCA-US cohort and the LUAD-US cohorts, respectively. In both cases, RELA was predicted to have an activity score less than 0, implying that this TF is inactive and therefore does not affect the transcription of *ANKRD53*. When considering the possibility that RELA's activity may be captured within the TF REL instead, which has a nearly identical binding motif, a negative score also predicted .

Furthermore, the samples containing the three-highest ranking $TERT_{C228T}$ pSNVs were also tested, which is known to create the ETS-family TF GABPA among others. Two of these samples belong to the thyroid carcinoma (THCA-US) cohort whereas the other belongs to the glioblastoma multiforme (GMB-US) cohort. Surprisingly, all eight method-database pairs predicted that GABPA is also inactive with estimates below 0. With this being the case, these estimates' corresponding p-values were also greater than 0.05 thereby implying inconclusive results, potentially attributable to the small cohort sizes.

Overall, when examining high-ranking samples containing $ANKRD53_{G529A}$ and $TERT_{C228T}$ pSNVs, both of which have been experimentally shown to create TFBSs for TFs that upregulate their corresponding genes, these TFs were generally predicted to be inactive. Because of these unexpected results, this raised doubts about the reliability of these tools when prospectively predicting the activity of a TF introduced by a pSNV *de novo* in future iterations of the REMIND-Cancer workflow.

5.8 REMIND-Cancer Workflow in A Precision Oncology Setting

In order to assess the applicability of the REMIND-Cancer workflow in different contexts, my pipeline was applied to individual patients originating from the NCT-MASTER program as well as a focused study on breast cancer from the COGNITION program. Here, computational efficiency and usability were of the utmost importance as both the REMIND-Cancer pipeline as well as *pSNV Hunter* were designed to have a translational impact, particularly in an MTB setting. Though several interesting candidate pSNVs were identified through these analyses, no *in vitro* validation efforts were planned and conducted.

Consequently, this section details the pilot study of samples from the *prospective* NCT-MASTER dataset in Section 5.8.1, the pilot study of samples from the singular breast cancer cohort from the COGNITION program in Section 5.8.2, and, lastly, evaluates the computational efficiency and general usability of this workflow in Section 5.8.3.

5.8.1 Pilot Study of the *Prospective* NCT-MASTER Program

The REMIND-Cancer workflow was first applied to new NCT-MASTER samples received weekly over a nine week period from 9 January to 13 March 2023. This encompassed 1,160,518 mutations across 105 unique samples. However, only 14 (13%) of these samples were analyzable by the REMIND-Cancer pipeline due to several limitations: lack of patient metadata, samples originating from metastasis sites rather than primary tumors and the absence of expression data.

For all primary tumor samples with available RNA-Seq data, I manually assigned the cohorts based on comments provided by clinicians. However, these comments were either left empty or lacked sufficient context to determine a cohort (e.g. a comment for a sample was "tumor grew in size"). Furthermore, no metadata information was saved within NCT-MASTER data folders as well as R databases generated by members of the MTB. Since cohorts are essential in the normalization of gene expression within the pipeline, samples without clear cohort assignments were excluded and were only used to query recurrence. Ultimately, 14 samples from 5 cohorts comprising 226,684 mutations remained for further analysis.

Of these remaining mutations, 1,699 (0.75%) were identified as being pSNVs and 103 (0.06% of all SNVs; 6% of all analyzable pSNVs) successfully passed the pipeline. Notably, none of the previously-implicated pSNVs, including both of the two *TERT* hotspot mutations, passed the pipeline. This result, however, was expected particularly since there were zero instances of $TERT_{C250T}$, $CDC20_{G529A}$, $RALY_{C927T}$, and $LEPROTL1_{C921T}$ across all 105 samples. Furthermore, although $TERT_{C228T}$ was observed in four samples, these samples could not be assigned an analyzable cohort and, therefore, were unable to be analyzed.

5.8.1.1 Nearly All Remaining *Prospective* NCT-MASTER pSNVs Are Singletons

Of the 103 *prospective* NCT-MASTER pSNVs passing the REMIND-Cancer pipeline, 101 were singletons when considering recurrence with the *retrospective* NCT-MASTER cohort, which was the basis for calculating the prioritization score. The two recurrent pSNVs found within the promoters of *LCMT1* and *RP11-358H9.1*, were still, however, lowly-recurrent with recurrence levels of one and four, respectively. Notably, *RP11-358H9.1* was also identified as a lncRNA by *pSNV Hunter*, similar to several high-scoring mutations in the *retrospective* NCT-MASTER dataset, owing to the importance of individually analyzing mutations and their specific functions.

Of the 101 singletons, the highest ranking pSNV was $NDUFA12_{T954A}$ was observed to have an exceptionally high zscore of 5.2 when compared to all *retrospective* samples from the neuroendocrine cohort. Furthermore, this pSNV was found to create a single binding site for the TF CREB3L4, which has been linked to high transcriptional activity associated with cancer cell proliferation in both breast cancer (Pu et al., 2020) and hepatocellular carcinoma (Inagaki et al., 2008).

Conversely, when querying recurrence only against the PCAWG dataset, 100 of the 103 remaining pSNVs were singletons. These three recurrent mutations were found in the promoters of *RP11-358H9.1*, *AC096649.3*, and *WFIKKN2*. In addition to the aforementioned *RP11-358H9.1* pSNV, *AC096649.3* was also found to be a lncRNA mutation through *pSNV Hunter*.

Regardless of annotating the *prospective* NCT-MASTER pSNVs with samples from the *retrospective* NCT-MASTER and/or the PCAWG datasets, nearly all pSNVs identified by the REMIND-Cancer pipeline were singletons and *pSNV Hunter* proved to be essential in determining which pSNVs effect protein-coding genes.

5.8.2 Pilot Study of the COGNITION Dataset

Similar to that of the *prospective* NCT-MASTER dataset, the REMIND-Cancer pipeline was also applied to the COGNITION dataset. The COGNITION dataset consisted of 89 breast-cancer WGS samples, all of which had corresponding RNA-Seq data, though no information regarding biopsy location (i.e. metastatic site or from the primary tumor) was indicated. Prior to data transfer by Dr. Marc Zapatka to myself, each WGS sample was filtered to only include pSNVs according to a given GENCODE v19 BED file.

In total, the COGNITION dataset consisted of 15,561 total pSNVs (mean of 175 pSNVs per sample), which was similar to that of the retrospective NCT-MASTER dataset (mean of 215 pSNVs per sample) and the PCAWG dataset (mean of 153 pSNVs per sample). All COGNITION samples were considered to be within the same cohort. Through the application of the REMIND-Cancer pipeline, 70 pSNVs from 36 samples passed the entirety of the pipeline and were subsequently prioritized.

5.8.2.1 67% of Remaining COGNITION pSNVs Are Recurrent

Of the 70 COGNITION pSNVs prioritized by the pipeline, a surprising 47 pSNVs (67%) were recurrent with at least one other COGNITION sample - an extremely high percentage relative to that of the PCAWG and NCT-MASTER datasets, which found only roughly 5% of remaining pSNVs to be recurrent.

In particular, the pSNVs with the highest recurrence level was found within the promoter of $CCDC173_{C490G}$, which was recurrent within 8 other COGNITION samples, followed by $COL6A5_{G665C}$, which had a recurrence statistic of 4. However, only one instance of $CCDC173_{C490G}$ successfully passed the REMIND-Cancer pipeline whereas two instances of $COL6A5_{G665C}$ remained.

Interestingly, when evaluating the recurrence of these 70 pSNVs in relation to other datasets, only one COGNITION pSNV (1.4%) was recurrent in PCAWG samples and 16 pSNVs (23%) were recurrent in *retrospective* NCT-MASTER samples.

Furthermore, the highest-ranking $CDK2AP2_{A840C}$ pSNV was recurrent when considering all datasets. This pSNV was recurrent with one other sample in the COGNITION dataset, had a recurrence rate of one within PCAWG, and, lastly, had a recurrence rate of 14 within the *retrospective* NCT-MASTER dataset. Though this pSNV was identified within COGNITION, thereby originating from a breast cancer sample, this mutation was found within 7 different NCT-MASTER cohorts, which include breast cancer, as well as esophageal adenocarcinoma (ESAD-UK) within PCAWG.

Though this pSNV looks promising as a future *in vitro* validation candidate when solely looking at recurrence, its high transcriptional activity (zscore = 1.86) may correspond to having a tumor-suppressive functionality. By analyzing this particular gene's function via *pSNV Hunter*, *CDK2AP2* is a paralog of *CDK2AP1*, whose high transcriptional activity within breast cancer has been shown to effectively suppress cell growth (Gera, Mokbel, Jiang, & Mokbel, 2018). Though this is not conclusive evidence of the oncogenic potential of *CDK2AP2*, this adds to the necessity of manual inspection and having tools, such as *pSNV Hunter*, to investigate pSNVs and their corresponding genes further.

5.8.3 Precision Oncology Applicability

To recapitulate the purpose of applying the REMIND-Cancer workflow to the single-sample *prospective* NCT-MASTER dataset and the single-cohort COGNITION program, the two primary focuses were to ensure the computational efficiency and general usability in a precision oncology setting.

Consequently, the open-source REMIND-Cancer computational pipeline supports three modes of analysis: pan-cancer analysis in the case of PCAWG and the *retrospective* NCT-MASTER datasets, single-cohort analysis in the case of the EOPC-DE PCAWG and COGNITION datasets, and single-patient analysis in the case of the weekly *prospective* NCT-MASTER patients. In all three settings, the ability to identify, prioritize and investigate highly-recurrent mutations, as well as the often-overlooked singletons and lowly-recurrent mutations, was possible. With customizable pipeline parameters (i.e. different filtering threshold values), researchers as well as clinicians are able to conduct their pSNV analysis despite potentially having different use-cases.

In terms of computational efficiency, all samples can be analyzed *in parallel* using the DKFZ cluster system (or any cluster system) meaning that multiple samples can be analyzed at the same time. As the time taken to complete the each individual sample scales linearly with the amount of mutations (Figure 27) regardless of dataset (e.g. *prospective* NCT-MASTER or PCAWG), more than 99% of samples can be analyzed in under one hour, thus ensuring quick and reliable results in translational settings.

5.9 DREAM Challenge

As a complementary project during my PhD, I participated in the "Predicting Gene Expression Using Millions of Random Promoter Sequence" DREAM challenge in which the primary objective was to create a neural network to predict gene expression based solely on promoter sequences (see Section 3.2 for details). Consequently, this section will detail the results of this challenge.

5.9.1 A Transformer-based Architecture Leads To An Improvement Over the Benchmark

Prior to the start of the challenge, the project organizers created a neural network, particularly that of a transformer (see Section 2.1.3), to address a

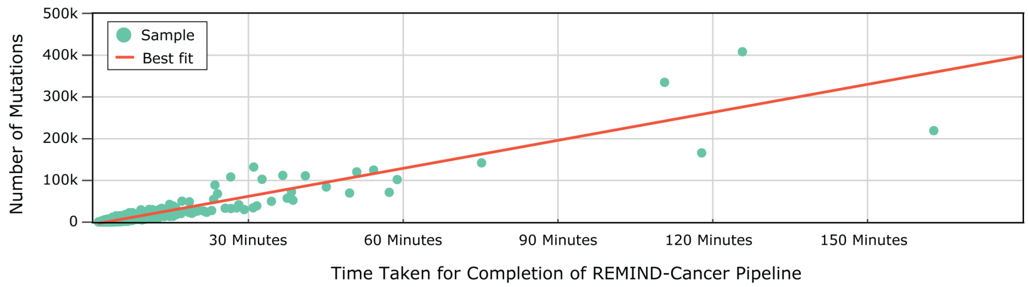

Figure 27: *A scatter plot of the amount of time needed to complete the REMIND-Cancer Pipeline (x-axis) and the number of mutations (y-axis) for each sample (green). The best fit line (red) is also plotted, which has an* R^2 *value of 0.930. For visualization purposes, four outlier samples that took longer than 180 minutes were omitted from the plot, although these values were still included in the calculation of the linear regression / best fit line.*

similar sequence-to-expression task (Vaishnav et al., 2022). As this model performed best in previous benchmarks, the organizers used this exact model as a baseline for teams throughout the competition. Specifically, this model was used for training and prediction similar to all other teams, which implied that models surpassing this baseline would be an advancement in this subfield of bioinformatics.

Within this competition, I leveraged three distinct ideas that led to an improvement over the aforementioned baseline: a transformer-based neural network with split branches to represent both DNA strands (see Section 3.2.3), a preprocessing strategy to generalize the training data's expression (see Section 3.2.2) and bayesian optimization as a hyperparameter optimization method to obtain the optimal hyperparameters in a computationally efficient manner (see Section 3.2.4).

The *final* test set in which teams were truly evaluated on consisted of 71,103 promoter sequence and expression pairs. In addition to being evaluated on the Pearson and Spearman correlation of these sequences, the test set was further divided into small sub-categories such as those sequences with high expression and sequences corresponding with low expression. Final rankings were a weighted sum of these measures with a majority of the scoring weight going to the two main categories of overall Pearson and Spearman correlation.

Of the 292 teams that submitted a model during the pre-submission phase, not all submitted a final model. However, of all teams registering for the competition, I placed 18th, which is within the top 10% of all teams submitting at least a pre-submission model. Considering all sequences, my model corresponded to a Pearson score of 0.88 as well as a Spearman score of 0.94, both of which were exactly equal to the baseline model. When considering the Spearman and Pearson score of sequences with high expressions, I obtained a 0.29 (improvement of 0.03 over the baseline) and 0.5 (improvement of 0.19 over the baseline), respectively. Conversely, my network performed worse on sequences with low expression with scores of 0.24 (decrease of 0.03 from the baseline) and 0.33 (decrease of 0.09 from the baseline). A subset of the results of nine teams (top 3 teams, teams ranked 18th through 20th, and the bottom 3 teams) can be seen in Figure 28.

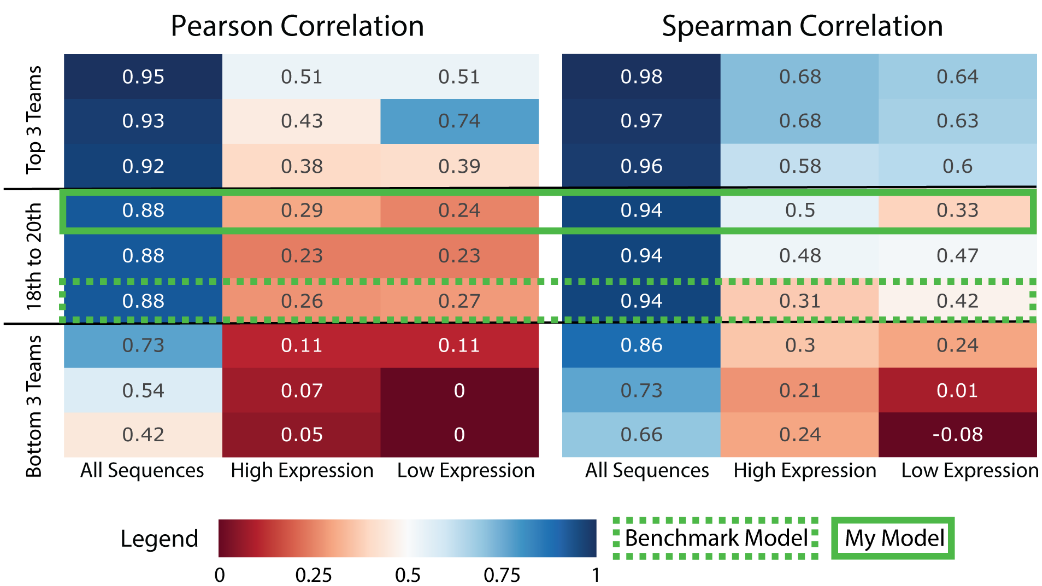

Figure 28: *Overview of Pearson (left) and Spearman (right) correlations of 9 selected teams. The top three teams, teams ranked 18th to 20th, which include my own network (solid green line) and the benchmark (dotted green line), and the bottom 3 teams can be seen.*

6 Discussion

6.1 The REMIND-Cancer Workflow Approach

Within this thesis, I primarily presented the Regulatory Mutation Identification N' Descriptions in Cancer (REMIND-Cancer) workflow, which is an integrative approach to identify and characterize activating functional promoter mutations. This workflow follows a filtering-ranking-inspection-validation paradigm in which a series of filtering steps are first applied to SNVs. Subsequently, biologically-relevant features, such as open chromatin and association with a known cancer gene, are then added to the remaining sample-specific pSNVs. These features are then used to compute a prioritization score for subsequent ranking, which serves as a basis for the inspection phase.

To assist in this inspection phase, *pSNV Hunter*, which is a comprehensive data aggregation and interactive visualization tool, was created to efficiently assess the reliability of high-scoring candidates. After the manual selection of mutations, the functionality of these pSNVs are then measured *in vitro* using a luciferase reporter assay in which the wild type activity is compared to that of its corresponding mutant, thereby further enhancing our understanding of the pSNV's functional implications.

As of today, a manuscript detailing this study is in preparation although a pre-print is currently available and readily accessible (Abad, Glas, et al., 2024). Additionally, four tools have been made publicly-available to facilitate future research in the field:

1. The REMIND-Cancer computational pipeline (Abad, Körner, & Feuerbach, 2024d)

2. *pSNV Hunter* (Abad, Körner, & Feuerbach, 2024c)

3. *DeepPileup* (Abad, Körner, & Feuerbach, 2024a)

4. A wrapper for *GenomeTornadoPlot* (Abad, Körner, & Feuerbach, 2024b)

The following section will discuss these results, their implications and their potential limitations in detail.

6.1.1 Identifying Ten Functional pSNVs and Assessing Their Clinical Potential

Upon application of the REMIND-Cancer workflow to the PCAWG and *retrospective* NCT-MASTER datasets, 19,250 and 6,274 pSNVs, were identified and prioritized, respectively. Utilizing *pSNV Hunter* led to the selection of 22 total pSNVs for *in vitro* validation, 15 of which came from PCAWG whereas the remaining 7 were from the retrospective NCT-MASTER. Of these selected mutations, eight displayed a statistically-significant (p-value ≤ 0.05; one-sided t-test) upregulation of promoter activity. With the addition of the previously positively-validated $ANKRD53_{G529A}$ and MYB_{C964A} detailed in Sections 5.2.3 and 5.2.4, respectively, this thesis thus implicates 10 total pSNVs as having an activating effect on their corresponding gene.

However, the importance of identifying these 10 functional pSNVs lies within the potential to influence therapeutic decision making as well as provide better prognostics for individual patients. By identifying potential biomarkers, such as these 10 functional pSNVs, this enables the detection of specific targets (i.e. genes or proteins that play a critical role in a cancer cell's growth, progression and/or survival) within a singular patient's cancer. These targets could then be assessed for their drugability, allowing for the selection and application of specific drugs that inhibit downstream pathways and/or modulate the activity of these genes directly. Examples of known targets include that of *KRAS* (Bekaii-Saab et al., 2023) (Jänne et al., 2022) (Yaeger et al., 2023) and *BRAF* (da Rocha Dias et al., 2013), both of which are listed within the list of 282 targettable genes provided by COSMIC (Tate et al., 2019).

Because catalogs of targettable genes, such as that provided by COSMIC, do not yet contain any of the genes associated with the 10 functional pSNVs identified in this study, no effective treatments have yet been created to target these genes. By elucidating the effects of these 10 pSNVs, this research could potentially guide future drug development. It could help in first assessing the true oncogenic potential of these genes and their corresponding proteins and then in developing drugs to counteract these effects. More broadly, this thesis could help elucidate the importance of further investigating the effects of *non-coding* mutations on genes rather than just focusing on *coding* mutations as current cancer research typically does.

In addition to their therapeutic potential, these 10 pSNVs can also add prognostic value to clinicians. By understanding the specific mutations that drive cancer progression, clinicians can better predict disease outcomes and tailor monitoring strategies accordingly. For example, certain mutations may lead to a resistance to conventional therapies or may be associated with more aggressive diseases.

For instance, since prostate cancer is the second leading cause of cancer-related deaths in men (Rawla, 2019), extensive research has focused on understanding how androgens stimulate prostate cancer growth and how inhibiting the androgen receptor (AR) can counteract this process. Though 70-80% of men with prostate cancer treated with this inhibition respond positively, most patients develop androgen-resistance (J. Edwards et al., 2001). This development of androgen-resistance has been linked to the overexpression of *MYB*, typically achieved through an increase in copy number level. However, as this study has shown that this overexpression may also be achieved by MYB_{C964A}, this may point to an alternative mechanism of developing androgen-resistance in the absence of amplification. In the identification of this mechanism, this may lead to clinicians being able to diagnose androgen-resistance in an alternative way.

Taken together, the identification, prioritization, and positive *in vitro* validation of these pSNVs highlight the potential for improving personalized cancer treatment by enhancing therapeutic decision-making and providing critical prognostic value.

6.1.2 Singletons Must Also Be Considered

Due to their low frequency, singletons, such as the aforementioned MYB_{C964A}, have been considered to be insignificant (Desai et al., 2024) though this category of mutations comprise 90% of some publicly-available WGS datasets (Chakraborty, Arora, Begg, & Shen, 2019). However, recent studies have highlighted the oncogenic potential of these rare variants (Zhao et al., 2022) (E. Kim et al., 2016) (Ostroverkhova et al., 2023), which are hidden in the "long tail of infrequent molecular alterations" (Scholl & Fröhling, 2019). Consequentially, I specifically targeted these understudied types of mutations using the REMIND-Cancer workflow.

Among the 19,250 (PCAWG) and 6,274 (retrospective NCT-MASTER) pSNVs remaining after the conclusion of REMIND-Cancer pipeline, approximately 95% of these mutations were singletons. As such, 14 singleton pSNVs as well as two lowly-recurrent pSNVs were prioritized and selected for *in vitro* validation in order to measure their potential activating effects on the promoter. Of these mutations, six singletons were positively validated within this study, namely $SCN1B_{C113T}$, $NCBP2\text{-}AS2_{G713T}$, $PRKCB_{C963T}$, $EGR1_{C049T}$, $TFEB_{T989G}$ and MYB_{C964A}.

These findings demonstrate that this underrepresented class of mutations can indeed be identified and prioritized using methods that do not inherently require recurrence. Unlike prior approaches such as FunSeq2 (Fu et al., 2014), OncodriveFML (Mularoni et al., 2016), and ncDriver (Hornshøj et al., 2018), which rely on comparing a statistical background model to an expected mutation rate, these methods miss singletons and lowly-recurrent mutations due to a lack of statistical power. Consequently, truly functional mutations, such as these six, are not able to be distinguished from their background noise, therefore leaving a significant blindspot in the identification of many functional pSNVs.

As one major goal of cancer research is to add to the list of known oncogenic mutations, these six activating singleton pSNVs demonstrate that singletons can lead to a functional effect and must also be considered in any driver identification analysis.

6.1.3 $ANKRD53_{G529A}$ and MYB_{C964A} Reveal a Two-Hit Mechanism That May Be Required For Promoter Activation

Both $ANKRD53_{G529A}$ and MYB_{C964A} have been shown to increase promoter activity, albeit under unique circumstances. As detailed in Section 5.2.3, $ANKRD53_{G529A}$ alone leads to a slight 20% increase in promoter activity. However, activation of the TF RELA (p65), which corresponds to the TFBS that is predicted to be created, further increases promoter activity to 80% over the WT. This significant effect is reversed by the knockdown of p65, thereby indicating that this second hit is necessary for a substantial growth advantage.

Furthermore, through the integration and analysis of the EOPC-DE dataset detailed in Section 5.2.4, MYB_{C964A} was identified and prioritized as being the top ranking mutation, particularly due to the sample's high *MYB* expression (zscore of 3.75) and the creation of binding sites for FOXD1 and FOXO3, both of which exhibited high transcription. Consequently, the functional validation of this pSNV revealed a significant 63% increase in promoter activity. Similar to $ANKRD53_{G529A}$, the transcriptomic upregulation of FOXD1 and FOXO3, along with the MYB_{C964A} pSNV, may be necessary in order to substantially increase promoter activity in this specific context.

Taken together, these secondary activation mechanisms suggest that these pSNVs significantly activate their corresponding promoters only under specific conditions. This indicates a potential selection bias where a secondary mechanism is required to enhance the selective advantage conferred by the pSNV.

6.1.4 Efficiency is Key in Precision Oncology

The time frame from biopsy to therapeutic decision by a molecular tumor board (MTB) is extremely short. For instance, the NCT-MASTER program reports a turnaround time of approximately 6 weeks while other programs and studies report a clinically actionable time frame of 2-4 weeks (Acanda De La Rocha et al., 2024) (Meggendorfer et al., 2022). Within this time, oncologists, clinicians, bioinformaticians and other health care professionals interpret the clinical and molecular information of individual cancer patients on a case-by-case basis in order to recommend personalized treatment and/or clinical trial options (Tsimberidou et al., 2023). As personalized oncology programs become more common (Tsimberidou et al., 2023), having efficient yet reliable analyses is a necessity, particularly since quick and accurate decisions can significantly impact patient outcomes.

Consequently, the REMIND-Cancer computational pipeline and *pSNV Hunter* were specifically designed to address these concerns. Since efficiency is key, the REMIND-Cancer pipeline can identify and prioritize functional pSNVs in pan-cancer, single-cohort, and single-sample datasets typically in under an hour. Furthermore, the pipeline itself is easily modifiable meaning that annotations, thresholds, and parameters can be adjusted within a single configuration file to fit specific research and program needs.

Additionally, *pSNV Hunter* saves time by automatically aggregating patient information, along with their genomic, transcriptomic, and annotation-based data, and displays this information in the form of an interactive and visual dashboard. As multiple individuals typically assess patients simultaneously, *pSNV Hunter* also allows users to leave comments about individual mutations, characterize whether a mutation may require further investigation, and export results to send to their colleagues. Altogether, this tool saves time on time-consuming tasks (i.e. file integration, individual mutation graph generation) while giving users a holistic view of the mutation at hand.

6.1.5 Studying Other Non-Coding Elements is a Natural Extension of the REMIND-Cancer Workflow

Within this study, characterizing singular point mutations within the core and proximal promoter region of protein-coding genes were the primary focus, ultimately leading to observing the activating functionality of 10 pSNVs *in vitro*. However, a natural extension to this project would be to also incorporate a promoter's interaction with other non-coding elements, particularly enhancers, as well as consider the idea of a promoter only being activating in combination with other mutations (i.e. latent drivers).

In the case of incorporating enhancer-promoter interaction into the REMIND-Cancer workflow, it is well known that enhancers are able to interact directly with multiple promoters. Through mechanisms such as DNA looping, enhancers can increase promoter activity, even when located thousands of base pairs away, either upstream or downstream of the promoter. This enhances the transcription of target genes, which is the phenomenon that I was directly investigating through the REMIND-Cancer workflow. However, this interaction is not accounted for in the original pipeline implementation. The presence of an active enhancer could serve as an additional annotation, contributing to the selection of future candidate pSNVs. This integration, however, presents challenges such as the inability of relatively short *in vitro* luciferase assays to capture long-range enhancer-promoter interactions and also the limited knowledge of known enhancer-promoter interactions.

In addition to a promoter's interaction with enhancers, another potential extension of the REMIND-Cancer workflow would be to incorporate the possibility of latent drivers. Recent studies have shown that functionally weak individual mutations can still confer a growth advantage or lead to increased drug resistance when coupled with other mutations, particularly when these mutations effect the same gene (Saito et al., 2020) (Yavuz et al., 2023) (Vasan et al., 2019). These latent drivers do not inherently fit the classical definition of a driver or a passenger mutation as their functionality depends on the presence of another specific mutation and are thus typically characterized as passengers (Yavuz et al., 2023). Consequently, they have not yet been extensively studied (Nussinov & Tsai, 2015).

However, of those studies that have investigated this third class of mutations, they have predominantly focused on synergistic, recurrent protein-coding mutations in well-known oncogenes, such as *PIK3CA* and *EGFR* (Saito et al., 2020). Given that pSNVs, similar to that of coding mutations, can indeed *individually* be functional, it is plausible that a pSNV can also work synergestically with other mutations, known drivers or not, to also alter gene functionality. This could be achieved by modifying the original REMIND-Cancer pipeline to identify and prioritize singular pSNVs in combination with other mutations, thereby providing deeper insights into the role of pSNVs in cancer progression.

6.1.6 Special Care Must Be Taken When Dealing With Clinical Data

The REMIND-Cancer workflow was applied to two sets of WGS and RNA-Seq data: publicly-available data (i.e. PCAWG) and clinical data (i.e. NCT-MASTER and COGNITION). Although PCAWG samples were originally from multiple clinics, the PCAWG consortium only ensured that samples were of high quality with as few artifacts as possible. Consequently, this dataset has been the basis of many cancer studies.

However, the raw clinical data of the NCT-MASTER and COGNITION programs, understandably, did not have these same benefits. In particular, I observed distinct differences between these two categories of data that must be taken into account when, not only applying the REMIND-Cancer workflow, but also when applying any bioinformatics tool to clinical NGS data in general. Here, I discuss three significant issues: (1) data quality, (2) lack of metadata and (3) lack of data annotations.

6.1.6.1 Data Quality

The first and arguably most important observation is the noticeable differences within data quality. As noted previously in Section 1.10.1, each sample used in generating the PCAWG study was deemed to be of high quality through rigorous quality assurance measures. This resulted in my analysis of 2,413 approved donors corresponding to 19,401,901 mutations. However, though quality control measures are cited as being applied to the WGS data from the NCT-MASTER project, several noticeable problems emerged: heavy sample outliers (e.g. samples with up to 254 times the median number of mutations) though this may be attributable to late-stage cancer patients, duplicated and/or incomplete data points (e.g. SNV VCF files denoting an SNV but with no ‘ALT‘ nucleotide to represent the actual nucleotide change, SNV VCF files denoting an indel, multiple rows with the same exact information), and SNVs sporadically assigned to genes of interest (i.e. two of the same genomic positions being mapped to different genes on the same DNA strand).

6.1.6.2 Lack of Sample Metadata: Sample Cohorts

As detailed in Section 3.1.2.3, the REMIND-Cancer workflow detects pSNVs leading to an increase in transcription of its corresponding gene by normalizing FPKM values based on the gene name and other samples within the sample's cohort. This, therefore, requires cohort information for each sample to be readily-available and accurate. Though several studies, including that of the NCT-MASTER's initial seminal study (Horak et al., 2017), cite specific cohort details, this sample-to-cohort information was not available within any metadata file or NCT-MASTER database (e.g. the internal NCT-MASTER Data Object used by Barbara Hutter and Malgorzata Oles of the NCT-MASTER MTB or the One-Touch Pipeline). Similarly, unlike the emails sent to me on a weekly-basis for new NCT-MASTER patients, clinician comments were not documented and saved either, rendering me unable to manually assign cohorts to specific samples for a two-year period between 2021 and 2023.

6.1.6.3 Lack of Sample Metadata: Biopsy Location

In addition to the lack of cohort information, metadata regarding the location of a biopsy was missing, which should ideally be readily-available for any researcher particularly when conducting any transcriptomic analysis. When analyzing samples originating from the NCT-MASTER program, the REMIND-Cancer workflow only considered RNA-Seq samples from the primary tumor (See Section 4.4), which ultimately led to the exclusion of approximately 10 million SNVs from 373 metastatic samples from the *retrospective* NCT-MASTER dataset. However, as the NCT-MASTER program specifically enrolls patients with advanced-stage cancer who have may have been pre-treated, metastatic samples are to be expected.

This, however, presents a challenge as the expression of genes within metastatic samples depend on characteristics such as cancer grade, tumor purity, the tumor of origin as well as where the sample was taken (Y. Zhang et al., 2024) (Aftimos et al., 2021) (Garcia-Recio et al., 2023) (Cosgrove et al., 2022). Therefore, including metastatic samples with primary tumor samples during the expression normalization process could significantly effect the detection of putative pSNVs stemming from primary tumors. In future studies that utilize transcriptomic data from both sets of samples, any bioinformatic analysis must take into account the differences in expression in order to fully utilize and take advantage of entire NGS datasets. However, to do so, having metadata regarding biopsy location would be extremely beneficial in this normalization process.

As normal cells from one tissue type are mixed with tumor cells from a different tissue of origin within a metastatic sample, bulk RNA-Seq provides an average gene expression profile of the combination of these cells. If certain genes are typically highly expressed in the normal cells but have low expression in the tumor tissue of origin, the resulting gene expression value in the bulk RNA-Seq data could be misleadingly high, especially in samples with low tumor purity. Therefore, also having metadata on the location of the biopsy could help distinguish between gene expression changes due to the tumor cells and those due to the surrounding normal tissue. This information is crucial for accurate interpretation and effective normalization in transcriptomic analyses, ultimately improving the detection of clinically relevant mutations.

6.1.6.4 Lack of Data Annotations

In addition to having data quality and metadata issues, a third major problem I encountered was the lack of file annotations. In particular, there were no details regarding which files were used for internal MTB analysis. Although nearly all patients had multiple WGS files enabling a potential longitudinal analysis, many patients had WGS files with extremely similar time stamps though having drastically different mutational information (e.g. one patient had two WGS files with the same date though one file had approximately 20,000 SNVS whereas the other had 6,000). Though one or more of these files could be attributed to a potential sequencing mistake, bioinformaticians outside of the MTB, such as myself, must make assumptions as to which file to use in analysis.

6.1.7 Luciferase Assays Could Guide Subsequent Endogenous Testing

In this thesis, I implicated 10 pSNVs, including $ANKRD53_{G529A}$ and MYB_{C964A}, that lead to a statistically-significant increase in promoter activity when measured *in vitro* using a luciferase reporter assay. Although luciferase assays are commonly used for this purpose (Horn et al., 2013) (He et al., 2021) (Godoy et al., 2023), these assays only include a small fragment of DNA that lacks chromatin context and other regulatory components (e.g. enhancers, silencers), therefore being unable to capture the true endogenous context of a pSNV (Rojas-Fernandez et al., 2015). Ideally, an experimental system should take into account *all* components of gene regulation to capture a pSNVs true effect, though doing so may have financial implications.

To account for the inherent limitations of luciferase reporter assays, several studies have advocated for using precise genome editing technologies such as CRISPR-Cas9 (Cong et al., 2013) (Cho, Kim, Kim, & Kim, 2013) (P. D. Hsu et al., 2013). CRISPR-Cas9 allows for targeted modifications directly in the genome, preserving the natural chromatin environment and endogenous regulatory elements. This technology can create specific mutations at a desired location, providing a more accurate representation of the functional impact of pSNVs within their native genomic context (i.e. inclusive of long-range interactions and overall chromatin structure).

With this in mind, luciferase assays, such as the one within this study, can serve as an essential preliminary step in identifying which genomic positions to target with CRISPR. By using luciferase reporter assays, researchers can quickly and cost-effectively screen for pSNVs that result in significant changes in gene expression, providing an initial indication of which mutations have a potential regulatory impact. Once specific pSNVs are identified, such as the 10 pSNVs within this study, these genomic positions can then be prioritized for further investigation using CRISPR-Cas9. Specifically in pSNV validation, this two-step approach (i.e. initial screening with luciferase assays followed by precise genome editing) would enable a more efficient and focused exploration of putative functional pSNVs, thereby enhancing our understanding of the oncogenic roles of pSNVs and the non-coding genome as a whole.

6.2 DREAM Challenge: Sequence-to-Expression

6.2.1 Do Transformer Architectures *Always* Perform Better?

Since their creation in 2017 (Vaswani et al., 2017), transformers have been at the forefront of NLP tasks due to their ability to capture both short-term and long-term relationships and dependencies in an efficient fashion, particularly through the use of a self-attention mechanism. In many tasks, transformer-based architectures outperform its predecessors (e.g. CNNs and LSTMs) (Vaishnav et al., 2022) using both bioinformatics and non-bioinformatics datasets and have even garnered public attention particularly through OpenAI's suite of tools (e.g. ChatGPT, Dall-E).

Consequently, my reasoning for choosing this type of model was that if this architecture has shown great promise in both research and in a commercial setting, it may be logical to use this approach within this sequence-to-expression task as well. However, as this type of neural network has been attempted previously in this setting, my aim was to improve upon this using better preprocessing techniques and hyperparameter optimization methods. This reasoning and approach was slightly validated as this led me to placing 18th out of the 292 participating teams, placing me within the top 10%.

However, the question remained: what techniques did the top-ranking methods employ and did they also utilize transformers? Surprisingly, none of the top three teams employed a transformer architecture. Instead, these teams relied on the direct predecessors of transformers, namely CNNs and LSTMs, but used creative preprocessing techniques in order to augment their data prior to training. Conversely, however, one of the bottom three ranking teams also employed a CNN-based architecture although no explicit data preprocessing techniques were conducted other than the one-hot encoding of the input sequence. The exact details of these methods can be found within Rafi et al. (Rafi et al., 2023).

All together, these results imply that in this particular task, especially when the given training dataset is structurally different from the true test dataset, effective preprocessing is necessary for achieving good predictive performance and that relying solely on the choice of neural network architecture is not sufficient.

6.2.2 Future Integrations Within The REMIND-Cancer Workflow

As presented throughout this thesis, the goal of the REMIND-Cancer workflow was to identify, prioritize and validate activating pSNVs using a recurence-agnostic approach with a focus on detecting singletons and lowly-recurrent mutations. This was largely accomplished through the use of three filtering steps, one of which was the TFBS motif and TF expression filter, which initially detects the presence of known TFBSs in both the WT and MUT sequences. Given that this challenge implicitly requires the detection of TFBSs, future iterations of the REMIND-Cancer workflow may benefit from incorporating neural network architectures, such as CNNs and transformers, and effective preprocessing techniques into the methodology when attempting to detect TFBSs.

Recent studies suggest that new TFBS motifs are continuously being detected (Fornes et al., 2020) (Inukai, Kock, & Bulyk, 2017). However, because the current filtering step relies on comparing sequences to a database of *known* TFs, there is a risk of overlooking consequential TFBSs that have not yet been identified. By leveraging approaches similar to those used in the DREAM challenge, future iterations of the REMIND-Cancer workflow may benefit by being able to learn TFBSs *de novo*, regardless of their prior discovery, thereby improving the overall ability of detecting activating pSNVs.

Acknowledgements

I would first like to thank Prof. Dr. Benedikt Brors for giving me the opportunity to do my PhD within his group. Throughout the entirety of my PhD, he has provided excellent support, both scientifically and personally, guiding me in my career and life in general for which I will always be thankful for.

Secondly, I want to express my deepest gratitude to both Dr. Lars Feuerbach and Dr. Cindy Körner for providing me with perfect supervision and unwavering support throughout these past four years. From our weekly scientific meetings to our personal conversations about life in general, I could not have asked for two more insightful, understanding and enjoyable mentors to learn from each and every day. My PhD journey would not have been possible without you two.

Thirdly, I would like to thank my thesis advisory committee, particularly Prof. Dr. Stefan Wiemann, Dr. Jakob Skou Pederson, Prof. Dr. Carl Herrmann and the aforementioned Prof. Dr. Benedikt Brors, Dr. Cindy Körner and Dr. Lars Feuerbach. Throughout our numerous meetings, each one of you provided me with valuable and constructive feedback that shaped my PhD project into what it is today. Thank you for all of your insights and support throughout these past years.

Additionally, from a scientific perspective, I would like to thank Sabine Karolus, Irina Glas and Dr. Dieter Wiechenhan for their amazing work in validating my results *in vitro*, Dr. Barbara Hutter and Dr. Malgorzata Oles for their continuous feedback and assistance with the NCT-MASTER program, Dr. Chen Hong and Yoann Pageaud for their prior work in regards to the PCAWG dataset, and Dr. Mark Zapatka for providing me with the COGNITION dataset.

From a group perspective, I would also like to thank all of the ABIs, both past and present, and, in particular my deskmates Lisa Häfele and Niklas Beumer, for supporting me and helping me both scientifically and personally along the way. To Lisa and Niklas, after these past years, I still can't believe that all three of us are at this point in our PhD - it still seems surreal to me. Also, a special thanks to Corinna Sprengart for being the most joyful, positive, and helpful person I could imagine. Without your constant help throughout my PhD and also helping me navigate the ins-and-outs of German bureaucracy, none of this would have been possible.

From a personal perspective, I would like to thank my mom, dad and brother for supporting me throughout these long couple of years and always being by my side. Additionally, I can't thank enough my amazing girlfriend Luisa Schwarzmüller (in addition to Elfie) for always providing a voice of reason and being there for me throughout everything. Last but not least, I'd like to thank my grandpa for being such a positive influence on my life with his never-ending positivity and playful attitude towards life. I know for a fact he's up there somewhere smiling down at me while eating his balut.

Supplement

pSNV	Location	Cohort	FPKM zscore	Recurrence	Created TFBS	Destroyed TFBS	Dataset
CDC20$_{G529A}$	chr1:43,824,529	SKCM-US	2.3	11	None	REL, ETV5, RELA, TFAP2C	PCAWG
EGR1$_{C049T}$	chr5:137,801,049	SKCM-US	2.7	0	ELK4, ZBTB7A, ELK3, ETB5	ZNF263	PCAWG
TRMT10C$_{G670}$	chr3:101,280,670	SKCM-US	1.2	9	None	ZBTB7A	PCAWG
SECISBP2$_{G357A}$	chr9:91,933,357	SKCM-US	4.9	6	None	ELK4, ZBTB7A	PCAWG
RALY$_{C927T}$	chr20:32,580,927	SKCM-US	0.5	14	None	ELF5, ZBTB7A	PCAWG
ZC3H10$_{G026C}$	chr12:56,512,026	HNSC-US	3.3	0	ZNF341	SPIB, ELF5, ETV6	PCAWG
PRKCB$_{C993T}$	chr16:23,846,993	LUAD-US	1.1	0	None	NR2C2	PCAWG
MECOM$_{G814A}$	chr3:169,381,814	SKCM-US	4.6	0	ZNF384	None	PCAWG
MAP2K4$_{G099A}$	chr17:11,924,099	SKCM-US	1.1	5	None	KLF16, KLF10 , ZNF460	PCAWG
FAM83B$_{C199T}$	chr6:54,711,199	SKCM-US	5.9	0	None	ZEB1, RBPJ	PCAWG
FBXW7$_{G413A}$	chr4:153,457,413	SKCM-US	1.0	0	None	ELK4, ETV6, SPIC, ZBTB7A	PCAWG
MAP1S$_{C242T}$	chr19:17,830,242	SKCM-US	1.7	5	None	ELK1, ZBTB7A, ELK3, ETV3, ETV5, ETS2	PCAWG
TFEB$_{T989G}$	chr6:41,703,989	BRCA-US	2.1	0	None	PRDM1	PCAWG
SMUG1$_{G889A}$	chr12:54,582,889	Skin	1.8	2	None	ELK4, ELF4, ZBTB7A, ETV5	NCT-MASTER
SCN1B$_{C113T}$	chr19:35,521,113	Skin	1.5	0	RORA	None	NCT-MASTER
NCBP2-AS2$_{G713T}$	chr3:196,669,713	Stomach	2.3	0	None	NHLH1, MYOG, BHLHE22, MYF5	NCT-MASTER
POU4F2$_{G763T}$	chr4:147,559,763	Head and Neck	8.6	0	None	NHLH1, MYOG, TCF12, BHLHE22, MYF5	NCT-MASTER
MAP3K13$_{C687G}$	chr3:185,000,687	Colorectal	4.3	0	EGR1, TCF3	None	NCT-MASTER
SDF4$_{G437A}$	chr1:1,167,437	Skin	1.8	4	None	ELK4, ETV6, ZBTB7A, ETV5	NCT-MASTER
XRRA1$_{C837T}$	chr11:74,659,837	Colorectal	1.0	0	MYB	MYOD1, MYOG, MSC, MEIS2, *	NCT-MASTER
FOXQ1$_{A824G}$	chr6:1,311,824	Stomach	1.3	0	NRF1, MXI1	None	NCT-MASTER
LEPROTL1$_{C921T}$	chr8:29,952,921	Head and Neck	0.6	6	EWSR1-FLI1	None	NCT-MASTER

* = MEIS3, ASCL1, ATOH1, BHLHA15, BHLHE22, MYF5

Table S1: *Table of the 22 pSNVs that were validated in vitro using a luciferase reporter assay.*

References

Abad, N., Glas, I., Hong, C., Small, A., Pageaud, Y., Maia, A., ... others (2024). The promoter mutation paucity as part of the dark matter of the cancer genome. *bioRxiv*, 2024–06.

Abad, N., Körner, C., & Feuerbach, L. (2024a, April). *DeepPileup.* https://github.com/nicholas-abad/deep-pileup.

Abad, N., Körner, C., & Feuerbach, L. (2024b, April). *Genome-TornadoPlot Wrapper.* https://github.com/nicholas-abad/genome-tornado-plot-wrapper.

Abad, N., Körner, C., & Feuerbach, L. (2024c, April). *pSNV Hunter: A Comprehensive Visualization Tool for Promoter SNVs.* https://github.com/nicholas-abad/pSNV-hunter.

Abad, N., Körner, C., & Feuerbach, L. (2024d, April). *REMIND-Cancer: Identifying and Characterizing Functional Promoter SNVs.* https://github.com/nicholas-abad/REMIND-Cancer.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from https://www.tensorflow.org/ (Software available from tensorflow.org)

Acanda De La Rocha, A. M., Berlow, N. E., Fader, M., Coats, E. R., Saghira, C., Espinal, P. S., ... others (2024). Feasibility of functional precision medicine for guiding treatment of relapsed or refractory pediatric cancers. *Nature Medicine*, 1–11.

Adler, D., & Wernert, N. (2012). Ets transcription factors and prostate cancer: the role of the family prototype ets-1. *International journal of oncology*, *40*(6), 1748–1754.

Aftimos, P., Oliveira, M., Irrthum, A., Fumagalli, D., Sotiriou, C., Gal-Yam, E., ... others (2021). Genomic and transcriptomic analyses of breast cancer primaries and matched metastases in aurora, the breast international group (big) molecular screening initiative. cancer discov. 2021; 11: 2796–2811. doi: 10.1158/2159-8290. *DOI: https://doi.org/10.1158/2159-8290. CD-20-1647. PMID: https://www.ncbi.nlm.nih.gov/pubmed/34183353*, 2796–2811.

Agarwal, V., & Shendure, J. (2020). Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell reports*, *31*(7).

AI, F. (n.d.). *Dash bootstrap components.* https://dash-bootstrap-components.opensource.faculty.ai/. London, England. (Accessed: 24 March 2024)

Akhtar, W., & Veenstra, G. J. C. (2011). Tbp-related factors: a paradigm of diversity in transcription initiation. *Cell & bioscience*, *1*, 1–12.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., ... others (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... others (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, *33*(8), 831–838.

Allen, B. L., & Taatjes, D. J. (2015). The mediator complex: a central integrator of transcription. *Nature reviews Molecular cell biology*, *16*(3), 155–166.

Alonso-Alconada, L., Eritja, N., Muinelo-Romay, L., Barbazan, J., Lopez-Lopez, R., Matias-Guiu, X., ... Abal, M. (2014). Etv5 transcription program links bdnf and promotion of emt at invasive front of endometrial carcinomas. *Carcinogenesis*, *35*(12), 2679–2686.

Alvarez, M. J., Giorgi, F., & Califano, A. (2014). Using viper, a package for virtual inference of protein-activity by enriched regulon analysis. *Bioconductor*, 1–14.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, *8*, 1–74.

Amodio, N., Raimondi, L., Juli, G., Stamato, M. A., Caracciolo, D., Tagliaferri, P., & Tassone, P. (2018). Malat1: a druggable long non-coding rna for targeted anti-cancer approaches. *Journal of hematology & oncology*, *11*, 1–19.

Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, *21*(2), 71–87.

Arriojas, A., Patalano, S., Macoska, J., & Zarringhalam, K. (2023, December). A Bayesian noisy logic model for inference of transcription factor activity from single cell and bulk transcriptomic data. *NAR Ge-*

nomics and Bioinformatics, *5*(4), lqad106. Retrieved 2024-03-27, from https://doi.org/10.1093/nargab/lqad106 doi: 10.1093/nargab/lqad106

Ascierto, P. A., Kirkwood, J. M., Grob, J.-J., Simeone, E., Grimaldi, A. M., Maio, M., ... Mozzillo, N. (2012). The role of braf v600 mutation in melanoma. *Journal of translational medicine*, *10*, 1–9.

Badia-i Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., ... others (2022). decoupler: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, *2*(1), vbac016.

Baker, M. (2016). Reproducibility crisis. *nature*, *533*(26), 353–66.

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, *27*(2), 299–308.

Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013). Collaborative hyperparameter tuning. In *International conference on machine learning* (pp. 199–207).

Baydilli, Y., & Atila, U. (2018). Understanding effects of hyper-parameters on learning: A comparative analysis. In *Proceedings of the international conference on advanced technologies, computer engineering and science, safranbolu, turkey* (pp. 11–13).

Bekaii-Saab, T. S., Yaeger, R., Spira, A. I., Pelster, M. S., Sabari, J. K., Hafez, N., ... others (2023). Adagrasib in advanced solid tumors harboring a kras g12c mutation. *Journal of Clinical Oncology*, *41*(25), 4097–4106.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bon, E., Driffort, V., Gradek, F., Martinez-Caceres, C., Anchelin, M., Pelegrin, P., ... others (2016). Scn4b acts as a metastasis-suppressor gene preventing hyperactivation of cell migration in breast cancer. *Nature communications*, *7*(1), 13648.

Bondy-Chorney, E., Baldwin, R. M., Didillon, A., Chabot, B., Jasmin, B. J., & Côté, J. (2017). Rna binding protein raly promotes protein arginine methyltransferase 1 alternatively spliced isoform v2 relative expression and metastatic potential in breast cancer cells. *The International Journal of Biochemistry & Cell Biology*, *91*, 124–135.

Boroń, D., Nowakowski, R., Grabarek, B. O., Zmarzły, N., & Opławski, M. (2021). Expression pattern of leptin and its receptors in endometrioid endometrial cancer. *Journal of Clinical Medicine*, *10*(13), 2787.

Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. (1985). Characterization of a "silencer" in yeast: a dna sequence with properties opposite to those of a transcriptional enhancer. *Cell*, *41*(1), 41–48.

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., ... others (2015). Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, *43*(D1), D36–D42.

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). Atac-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, *109*(1), 21–29.

Bullock, M., Lim, G., Zhu, Y., Åberg, H., Kurdyukov, S., & Clifton-Bligh, R. (2019). Ets factor etv5 activates the mutant telomerase reverse transcriptase promoter in thyroid cancer. *Thyroid*, *29*(11), 1623–1633.

Burley, S. K. (1996). The tata box binding protein. *Current opinion in structural biology*, *6*(1), 69–75.

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular cell*, *49*(5), 825–837.

CancerResearchUK. (2023). *Types of cancer.* Retrieved from https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/types-of-cancer#:~:text=For%20example%2C%20nerves%20and%20muscles,of%20cell%20they%20start%20in.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., ... others (2012). Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. *The ISME journal*, *6*(8), 1621–1624.

Carlevaro-Fita, J., Lanzos, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J. S., & Johnson, R. (2020). Cancer lncrna census reveals evidence for deep functional conservation of long noncoding rnas in tumorigenesis. *Communications biology*, *3*(1), 56.

Carrasco Pro, S., Bulekova, K., Gregor, B., Labadorf, A., & Fuxman Bass, J. I. (2020). Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. *Scientific Reports*, *10*(1), 17632.

Chakraborty, S., Arora, A., Begg, C. B., & Shen, R. (2019). Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nature communications*, *10*(1), 5506.

Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., ... others (2013). Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj*, *346*.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., ... others (2011). Improved survival with vemurafenib in melanoma with braf v600e mutation. *New England Journal of Medicine*, *364*(26), 2507–2516.

Chen, J., Nelson, C., Wong, M., Tee, A. E., Liu, P. Y., La, T., ... others (2021). Targeted therapy of tert-rearranged neuroblastoma with bet bromodomain inhibitor and proteasome inhibitor combination therapy. *Clinical Cancer Research*, *27*(5), 1438–1451.

Chen, M.-J. M., Chou, L.-C., Hsieh, T.-T., Lee, D.-D., Liu, K.-W., Yu, C.-Y., ... Chen, C.-Y. (2012). De novo motif discovery facilitates identification of interactions between transcription factors in saccharomyces cerevisiae. *Bioinformatics*, *28*(5), 701–708.

Chen, X., Wu, J.-m., Hornischer, K., Kel, A., & Wingender, E. (2006). Tiprod: the tissue-specific promoter database. *Nucleic acids research*, *34*(suppl_1), D104–D107.

Cheng, S., Castillo, V., & Sliva, D. (2019). Cdc20 associated with cancer metastasis and novel mushroom-derived cdc20 inhibitors with antimetastatic activity. *International journal of oncology*, *54*(6), 2250–2256.

Cheng, X., Jin, Z., Ji, X., Shen, X., Feng, H., Morgenlander, W., ... others (2019). Ets variant 5 promotes colorectal cancer angiogenesis by targeting platelet-derived growth factor bb. *International journal of cancer*, *145*(1), 179–191.

Cho, S. W., Kim, S., Kim, J. M., & Kim, J.-S. (2013). Targeted genome engineering in human cells with the cas9 rna-guided endonuclease. *Nature biotechnology*, *31*(3), 230–232.

Chollet, F. (2015). *Keras.* https://keras.io/.

Chopra, V. S., Kong, N., & Levine, M. (2012). Transcriptional repression via antilooping in the drosophila embryo. *Proceedings of the National Academy of Sciences*, *109*(24), 9460–9464.

Cler, E., Papai, G., Schultz, P., & Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor tfiid. *Cellular and Molecular Life Sciences*, *66*(13), 2123–2134.

Cline, M. S., & Karchin, R. (2011). Using bioinformatics to predict the functional impact of snvs. *Bioinformatics*, *27*(4), 441–448.

Cockerill, P. N. (2011). Structure and function of active chromatin and dnase i hypersensitive sites. *The FEBS journal*, *278*(13), 2182–2210.

Collins, F. S., & Fink, L. (1995). The human genome project. *Alcohol health* *and research world*, *19*(3), 190.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., … others (2013). Multiplex genome engineering using crispr/cas systems. *Science*, *339*(6121), 819–823.

Consortium, E. P., et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, *489*(7414), 57.

Cosgrove, N., Varešlija, D., Keelan, S., Elangovan, A., Atkinson, J. M., Cocchiglia, S., … others (2022). Mapping molecular subtype specific alterations in breast cancer brain metastases identifies clinically relevant vulnerabilities. *Nature communications*, *13*(1), 514.

Couckuyt, A., Seurinck, R., Emmaneel, A., Quintelier, K., Novak, D., Van Gassen, S., & Saeys, Y. (2022). Challenges in translational machine learning. *Human Genetics*, *141*(9), 1451–1466.

Cox, A. D., Fesik, S. W., Kimmelman, A. C., Luo, J., & Der, C. J. (2014). Drugging the undruggable ras: Mission possible? *Nature reviews Drug discovery*, *13*(11), 828–851.

D'Acquisto, F., May, M. J., & Ghosh, S. (2002). Inhibition of nuclear factor kappa b (nf-b). *Molecular interventions*, *2*(1), 22.

da Rocha Dias, S., Salmonson, T., van Zwieten-Boot, B., Jonsson, B., Marchetti, S., Schellens, J. H., … Pignatti, F. (2013). The european medicines agency review of vemurafenib (zelboraf®) for the treatment of adult patients with braf v600 mutation-positive unresectable or metastatic melanoma: summary of the scientific assessment of the committee for medicinal products for human use. *European journal of cancer*, *49*(7), 1654–1661.

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature biotechnology*, *38*(1), 56–65.

de Paula, R., Holanda, M., Gomes, L. S., Lifschitz, S., & Walter, M. E. M. (2013). Provenance in bioinformatics workflows. *BMC bioinformatics*, *14*(Suppl 11), S6.

Desai, S., Ahmad, S., Bawaskar, B., Rashmi, S., Mishra, R., Lakhwani, D., & Dutt, A. (2024). Singleton mutations in large-scale cancer genome studies: uncovering the tail of cancer genome. *NAR cancer*, *6*(1), zcae010.

Ding, K., Dixit, G., Parker, B. J., & Wen, J. (2023). Crmnet: A deep learning model for predicting gene expression from large regulatory sequence datasets. *Frontiers in Big Data*, *6*, 1113402.

Dolcet, X., Llobet, D., Pallares, J., & Matias-Guiu, X. (2005). Nf-kb in development and progression of human cancer. *Virchows archiv*, *446*, 475–482.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Edwards, J., Krishna, N., Mukherjee, R., Watters, A., Underwood, M., & Bartlett, J. (2001). Amplification of the androgen receptor may not explain the development of androgen-independent prostate cancer. *BJU international*, *88*(6), 633–637.

Edwards, S., Campbell, C., Flohr, P., Shipley, J., Giddings, I., Te-Poele, R., ... others (2005). Expression analysis onto microarrays of randomly selected cdna clones highlights hoxb13 as a marker of human prostate cancer. *British journal of cancer*, *92*(2), 376–381.

Eggermont, A. M., Spatz, A., & Robert, C. (2014). Cutaneous melanoma. *The Lancet*, *383*(9919), 816–827.

Elliott, K., & Larsson, E. (2021). Non-coding driver mutations in human cancer. *Nature Reviews Cancer*, *21*(8), 500–509.

Er-Lukowiak, M., Hänzelmann, S., Rothe, M., Moamenpour, D. T., Hausmann, F., Khatri, R., ... Lotter, H. (2023, December). Testosterone affects type I/type II interferon response of neutrophils during hepatic amebiasis. *Frontiers in Immunology*, *14*. Retrieved 2024-03-27, from https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1279245/full (Publisher: Frontiers) doi: 10.3389/fimmu.2023.1279245

Ernst, J., & Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, *9*(3), 215–216.

Ferlay, J., Ervik, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., ... Bray, F. (2024). *Global cancer observatory: Cancer today.* Retrieved from https://gco.iarc.who.int/today/en

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, *27*(13), 1741–1748.

Flaherty, K. T., Gray, R., Chen, A., Li, S., Patton, D., Hamilton, S. R., ... others (2020). The molecular analysis for therapy choice (nci-match) trial: lessons for genomic trial design. *JNCI: Journal of the National Cancer Institute*, *112*(10), 1021–1029.

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., ... others (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic* *acids research*, *48*(D1), D87–D92.

Fornes, O., Gheorghe, M., Richmond, P. A., Arenillas, D. J., Wasserman, W. W., & Mathelier, A. (2018). Manta2, update of the mongo database for the analysis of transcription factor binding site alterations. *Scientific Data*, *5*(1), 1–7.

Fountzilas, E., Tsimberidou, A. M., Vo, H. H., & Kurzrock, R. (2022). Clinical trial design in the era of precision medicine. *Genome medicine*, *14*(1), 101.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., … others (2019). Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, *47*(D1), D766–D773.

Freiman, R. N. (2009). Specific variants of general transcription factors regulate germ cell development in diverse organisms. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1789*(3), 161–166.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., … Gerstein, M. (2014). Funseq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology*, *15*, 1–15.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, *29*(8), 1363–1375.

Garcia-Recio, S., Hinoue, T., Wheeler, G. L., Kelly, B. J., Garrido-Castro, A. C., Pascual, T., … others (2023). Multiomics in primary and metastatic breast tumors from the aurora us network finds microenvironment and epigenetic drivers of metastasis. *Nature Cancer*, *4*(1), 128–147.

Garraway, L. A., & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, *153*(1), 17–37.

Gera, R., Mokbel, L., Jiang, W. G., & Mokbel, K. (2018). mrna expression of cdk2ap1 in human breast cancer: correlation with clinical and pathological parameters. *Cancer genomics & proteomics*, *15*(6), 447–452.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., … others (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128.

Giuliani, C., Bucci, I., & Napolitano, G. (2018). The role of the transcription factor nuclear factor-kappa b in thyroid autoimmunity and cancer. *Frontiers in endocrinology*, *9*, 471.

Gluck, C., Glathar, A., Tsompana, M., Nowak, N., Garrett-Sinha, L. A., Buck, M. J., & Sinha, S. (2019). Molecular dissection of the oncogenic role of ets1 in the mesenchymal subtypes of head and neck squamous cell carcinoma. *PLoS genetics*, *15*(7), e1008250.

Godoy, P. M., Oyedeji, A., Mudd, J. L., Morikis, V. A., Zarov, A. P., Longmore, G. D., ... Kaufman, C. K. (2023). Functional analysis of recurrent cdc20 promoter variants in human melanoma. *Communications Biology*, *6*(1), 1216.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., ... Lopez-Bigas, N. (2013). Intogen-mutations identifies cancer drivers across tumor types. *Nature methods*, *10*(11), 1081–1082.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, *17*(6), 333–351.

Göös, H., Kinnunen, M., Salokas, K., Tan, Z., Liu, X., Yadav, L., ... Varjosalo, M. (2022). Human transcription factor protein interaction networks. *Nature communications*, *13*(1), 766.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018.

Gurevich, I., Zhang, C., & Aneskievich, B. J. (2010). Scanning for transcription factor binding by a variant emsa. *Epidermal Cells: Methods and Protocols*, 147–158.

Hanker, A. B., Brown, B. P., Meiler, J., Marín, A., Jayanthan, H. S., Ye, D., ... others (2021). Co-occurring gain-of-function mutations in her2 and her3 modulate her2/her3 activation, oncogenesis, and her2 inhibitor sensitivity. *Cancer cell*, *39*(8), 1099–1114.

Hanussek, M., Bartusch, F., & Krüger, J. (2021). Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources. *PLoS Computational Biology*, *17*(7), e1009244.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... others (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hayward, N. K., Wilmott, J. S., Waddell, N., Johansson, P. A., Field, M. A., Nones, K., ... others (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, *545*(7653), 175–180.

He, Z., Wu, T., Wang, S., Zhang, J., Sun, X., Tao, Z., ... Liu, X.-S. (2021).

Pan-cancer noncoding genomic analysis identifies functional cdc20 promoter mutation hotspots. *Iscience*, *24*(4).

Head, T. (2016). *Scikit-optimize.* https://scikit-optimize.github.io/stable/index.html.

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (emsa) for detecting protein–nucleic acid interactions. *Nature protocols*, *2*(8), 1849–1861.

Hess, J. M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N. J., ... Getz, G. (2019). Passenger hotspot mutations in cancer. *Cancer Cell*, *36*(3), 288–301.

Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, *14*(8), 2.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., ... others (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, *173*(2), 291–304.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J.-P., ... others (2012). A landscape of driver mutations in melanoma. *Cell*, *150*(2), 251–263.

Hombach, D., Schwarz, J. M., Robinson, P. N., Schuelke, M., & Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. *BMC genomics*, *17*, 1–10.

Hong, C., Thiele, R., & Feuerbach, L. (2022). Genometornadoplot: a novel r package for cnv visualization and focality analysis. *Bioinformatics*, *38*(7), 2036–2038.

Horak, P., Heining, C., Kreutzfeldt, S., Hutter, B., Mock, A., Hüllein, J., ... others (2021). Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer discovery*, *11*(11), 2780–2795.

Horak, P., Klink, B., Heining, C., Gröschel, S., Hutter, B., Fröhlich, M., ... others (2017). Precision oncology based on omics data: the nct heidelberg experience. *International journal of cancer*, *141*(5), 877–886.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., ... others (2013). Tert promoter mutations in familial and sporadic melanoma. *Science*, *339*(6122), 959–961.

Hornshøj, H., Nielsen, M. M., Sinnott-Armstrong, N. A., Świtnicki, M. P., Juul, M., Madsen, T., ... others (2018). Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. *NPJ genomic medicine*, *3*(1), 1.

Hosseini-Gerami, L., Higgins, I. A., Collier, D. A., Laing, E., Evans, D., Broughton, H., & Bender, A. (2023, April). Benchmarking causal reasoning algorithms for gene expression-based compound mechanism of action analysis. *BMC Bioinformatics*, *24*(1), 154. Retrieved 2024-03-27, from https://doi.org/10.1186/s12859-023-05277-1 doi: 10.1186/s12859-023-05277-1

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., ... others (2013). Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, *31*(9), 827–832.

Hsu, T., Trojanowska, M., & Watson, D. K. (2004). Ets proteins in biological control and cancer. *Journal of cellular biochemistry*, *91*(5), 896–903.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013). Highly recurrent tert promoter mutations in human melanoma. *Science*, *339*(6122), 957–959.

Huang, L., Guo, Z., Wang, F., & Fu, L. (2021). Kras mutation: from undruggable to druggable in cancer. *Signal transduction and targeted therapy*, *6*(1), 386.

ICGC Consortium, I. C. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998.

Ignatieva, E. V., Levitsky, V. G., & Kolchanov, N. A. (2015). Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: A study of 1000 genomes project data. *International Journal of Genomics*, *2015*(1), 260159.

Inagaki, Y., Yasui, K., Endo, M., Nakajima, T., Zen, K., Tsuji, K., ... others (2008). Creb3l4, ints3, and snapap are targets for the 1q21 amplicon frequently detected in hepatocellular carcinoma. *Cancer genetics and cytogenetics*, *180*(1), 30–36.

Inc., P. T. (2015a). *Collaborative data science.* https://plot.ly. Montreal, QC. (Accessed: 24 March 2024)

Inc., P. T. (2015b). *Dash bio.* https://dash.plotly.com/dash-bio and https://github.com/plotly/dash-bio. Montreal, QC. (Accessed: 24 March 2024)

Inc., P. T. (2015c). *Dash python user guide.* https://dash.plotly.com/.

Montreal, QC. (Accessed: 24 March 2024)

Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor–dna binding: beyond binding site motifs. *Current opinion in genetics & development*, *43*, 110–119.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... others (2009). Repeatability of published microarray gene expression analyses. *Nature genetics*, *41*(2), 149–155.

Iwamoto, T., Niikura, N., Ogiya, R., Yasojima, H., Watanabe, K.-i., Kanbayashi, C., ... others (2019). Distinct gene expression profiles between primary breast cancers and brain metastases from pair-matched samples. *Scientific reports*, *9*(1), 13343.

Jagannathan, V., Roulet, E., Delorenzi, M., & Bucher, P. (2006). Htpselex—a database of high-throughput selex libraries for transcription factor binding sites. *Nucleic acids research*, *34*(suppl_1), D90–D94.

Jaisswal, A., & Naik, A. (2021). Effect of hyperparameters on backpropagation. In *2021 ieee pune section international conference (punecon)* (pp. 1–5).

Jänne, P. A., Riely, G. J., Gadgeel, S. M., Heist, R. S., Ou, S.-H. I., Pacheco, J. M., ... others (2022). Adagrasib in non–small-cell lung cancer harboring a krasg12c mutation. *New England Journal of Medicine*, *387*(2), 120–131.

Jeeta, R. R., Gordon, N. S., Baxter, L., Goel, A., Noyvert, B., Ott, S., ... others (2019). Non-coding mutations in urothelial bladder cancer: biological and clinical relevance and potential utility as biomarkers. *Bladder Cancer*, *5*(4), 263–272.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., ... others (2013). Dna-binding specificities of human transcription factors. *Cell*, *152*(1), 327–339.

Jones, D., Raine, K. M., Davies, H., Tarpey, P. S., Butler, A. P., Teague, J. W., ... Campbell, P. J. (2016). cgpcavemanwrapper: simple execution of caveman in order to detect somatic single nucleotide variants in ngs data. *Current protocols in bioinformatics*, *56*(1), 15–10.

Katz, D., Palmerini, E., & Pollack, S. M. (2018). More than 50 subtypes of soft tissue sarcoma: paving the path for histology-driven treatments. *American Society of Clinical Oncology Educational Book*, *38*, 925–938.

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLoS computational biology*, *16*(7), e1008050.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y.,

& Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, *28*(5), 739–750.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, *17*(2), 93–108.

Kim, E., Ilic, N., Shrestha, Y., Zou, L., Kamburov, A., Zhu, C., … others (2016). Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. *Cancer discovery*, *6*(7), 714–726.

Kim, G.-C., Kwon, H.-K., Lee, C.-G., Verma, R., Rudra, D., Kim, T., … Im, S.-H. (2018). Upregulation of ets1 expression by nfatc2 and nfkb1/rela promotes breast cancer cell invasiveness. *Oncogenesis*, *7*(11), 91.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2016). Progress and challenges in bioinformatics approaches for enhancer identification. *Briefings in bioinformatics*, *17*(6), 967–979.

Kordes, U., Krappmann, D., Heissmeyer, V., Ludwig, W., & Scheidereit, C. (2000). Transcription factor nf-κb is constitutively activated in acute lymphoblastic leukemia cells. *Leukemia*, *14*(3), 399–402.

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of cheminformatics*, *9*, 1–13.

Kumar, R., Ichihashi, Y., Kimura, S., Chitwood, D. H., Headland, L. R., Peng, J., … Sinha, N. R. (2012). A high-throughput method for illumina rna-seq library preparation. *Frontiers in plant science*, *3*, 28988.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., … Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650–665.

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, *29*(12), 1305–1312.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., … Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, *505*(7484), 495–501.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., ... others (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218.

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., & Yeh, H.-Y. (2019). Classifying promoters by interpreting the hidden information of dna sequences via deep learning and combination of continuous fasttext n-grams. *Frontiers in bioengineering and biotechnology*, *7*, 305.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Liu, E. M., Martinez-Fundichely, A., Bollapragada, R., Spiewack, M., & Khurana, E. (2021). Cncdatabase: a database of non-coding cancer drivers. *Nucleic Acids Research*, *49*(D1), D1094–D1101.

Liu, L., Hu, N., Wang, B., Chen, M., Wang, J., Tian, Z., ... Lin, D. (2011). A brief utilization report on the illumina hiseq 2000 sequencer. *Mycology*, *2*(3), 169–191.

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., ... others (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, *2*(10), e537–e548.

Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., & Gerstein, M. (2015). Larva: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic acids research*, *43*(17), 8123–8134.

Long, G. V., Swetter, S. M., Menzies, A. M., Gershenwald, J. E., & Scolyer, R. A. (2023). Cutaneous melanoma. *The Lancet*, *402*(10400), 485–502.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome biology*, *20*, 1–14.

Makkonen, H., Jääskeläinen, T., Pitkänen-Arsiola, T., Rytinki, M., Waltering, K., Mättö, M., ... Palvimo, J. (2008). Identification of ets-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells. *Oncogene*, *27*(36), 4865–4876.

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., ... Campbell, P. J. (2017). Universal patterns of selection in cancer and somatic tissues. *Cell*, *171*(5), 1029–1041.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., ... others (2020). A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, *20*(10), 555–572.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., ... others (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, *31*(1), 374–378.

McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., & Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, *110*(8), 2910–2915.

McNabb, D. S., Reed, R., & Marciniak, R. A. (2005). Dual luciferase assay system for rapid assessment of gene expression in saccharomyces cerevisiae. *Eukaryotic cell*, *4*(9), 1539–1549.

Meggendorfer, M., Jobanputra, V., Wrzeszczynski, K. O., Roepman, P., de Bruijn, E., Cuppen, E., ... others (2022). Analytical demands to use whole-genome sequencing in precision oncology. In *Seminars in cancer biology* (Vol. 84, pp. 16–22).

Melton, C., Reuter, J. A., Spacek, D. V., & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature genetics*, *47*(7), 710–716.

Miller, M. L., Reznik, E., Gauthier, N. P., Aksoy, B. A., Korkut, A., Gao, J., ... Sander, C. (2015). Pan-cancer analysis of mutation hotspots in protein domains. *Cell systems*, *1*(3), 197–209.

Minnoye, L., Marinov, G. K., Krausgruber, T., Pan, L., Marand, A. P., Secchia, S., ... others (2021). Chromatin accessibility profiling methods. *Nature Reviews Methods Primers*, *1*(1), 10.

Mo, J., Tan, K., Dong, Y., Lu, W., Liu, F., Mei, Y., ... others (2023). Therapeutic targeting the oncogenic driver ewsr1:: Fli1 in ewing sarcoma through inhibition of the fact complex. *Oncogene*, *42*(1), 11–25.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M., & Chambon, P. (1981). The sv40 72 base repair repeat has a striking effect on gene expression both in sv40 and other chimeric recombinants. *Nucleic acids research*, *9*(22), 6047–6068.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., & López-Bigas, N. (2016). Oncodrivefml: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, *17*, 1–13.

Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i Mompel, P., Fallegger, R., ... Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, *51*(20), 10934–10949.

Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., & Lu, T. (2014). Role of chip-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, *13*(18), 2847–2852.

Näär, A. M., Lemon, B. D., & Tjian, R. (2001). Transcriptional coactivator complexes. *Annual review of biochemistry*, *70*(1), 475–501.

Naugler, W. E., & Karin, M. (2008). Nf-κb and cancer—identifying targets and mechanisms. *Current opinion in genetics & development*, *18*(1), 19–26.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., ... others (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., ... others (2020). Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nature medicine*, *26*(9), 1320–1324.

Nussinov, R., & Tsai, C.-J. (2015). 'latent drivers' expand the cancer mutational landscape. *Current Opinion in Structural Biology*, *32*, 25–32.

Nussinov, R., Tsai, C.-J., & Jang, H. (2019). Why are some driver mutations rare? *Trends in pharmacological sciences*, *40*(12), 919–929.

Ohler, U., Harbeck, S., Niemann, H., & Reese, M. G. (1999). Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics (Oxford, England)*, *15*(5), 362–369.

Ostroverkhova, D., Przytycka, T. M., & Panchenko, A. R. (2023). Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine*, *29*(7), 554–566.

O'Dwyer, P. J., Gray, R. J., Flaherty, K. T., Chen, A. P., Li, S., Wang, V., ... others (2023). The nci-match trial: lessons for precision oncology. *Nature medicine*, *29*(6), 1349–1357.

Pang, B., van Weerd, J. H., Hamoen, F. L., & Snyder, M. P. (2023). Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology*, *24*(6), 383–395.

Panno, J. (2005). *Cancer: The role of genes, lifestyle, and environment.* Infobase Publishing.

Patel, M. B., & Wang, J. (2018). The identification and interpretation of cis-regulatory noncoding mutations in cancer. *High-throughput*, *8*(1),

1.
PCAWG Consortium, P. C. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93.
Perez, M. F., & Sarkies, P. (2023, October). Histone methyltransferase activity affects metabolism in human cells independently of transcriptional regulation. *PLOS Biology*, *21*(10), e3002354. Retrieved 2024-03-27, from https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002354 (Publisher: Public Library of Science) doi: 10.1371/journal.pbio.3002354
Périer, R. C., Praz, V., Junier, T., Bonnard, C., & Bucher, P. (2000). The eukaryotic promoter database (epd). *Nucleic acids research*, *28*(1), 302–303.
Perissi, V., Jepsen, K., Glass, C. K., & Rosenfeld, M. G. (2010). Deconstructing repression: evolving models of co-repressor action. *Nature Reviews Genetics*, *11*(2), 109–123.
Peterson, C. L., & Laniel, M.-A. (2004). Histones and histone modifications. *Current Biology*, *14*(14), R546–R551.
Pickering, M., & O'Connor, J. J. (2007). Pro-inflammatory cytokines and their effects in the dentate gyrus. *Progress in brain research*, *163*, 339–354.
Pixberg, C., Zapatka, M., Hlevnjak, M., Benedetto, S., Suppelna, J., Heil, J., ... others (2022). Cognition: a prospective precision oncology trial for patients with early breast cancer at high risk following neoadjuvant chemotherapy. *ESMO open*, *7*(6), 100637.
Pop, R. T., Pisante, A., Nagy, D., Martin, P. C., Mikheeva, L. A., Hayat, A., ... Zabet, N. R. (2023). Identification of mammalian transcription factors that bind to inaccessible chromatin. *Nucleic Acids Research*, *51*(16), 8480–8495.
Pradeepa, M. M. (2017). Causal role of histone acetylations in enhancer function. *Transcription*, *8*(1), 40–47.
Prior, I. A., Lewis, P. D., & Mattos, C. (2012). A comprehensive survey of ras mutations in cancer. *Cancer research*, *72*(10), 2457–2467.
Pu, Q., Lu, L., Dong, K., Geng, W.-w., Lv, Y.-r., & Gao, H.-d. (2020). The novel transcription factor creb3l4 contributes to the progression of human breast carcinoma. *Journal of Mammary Gland Biology and Neoplasia*, *25*, 37–50.
Puli, O. R., Danysh, B. P., McBeath, E., Sinha, D. K., Hoang, N. M., Powell, R. T., ... Hofmann, M.-C. (2018). The transcription factor etv5 medi-

ates brafv600e-induced proliferation and twist1 expression in papillary thyroid cancer cells. *Neoplasia*, *20*(11), 1121–1134.

Qian, C., Li, D., & Chen, Y. (2022). Ets factors in prostate cancer. *Cancer letters*, *530*, 181–189.

Rabbie, R., Ferguson, P., Molina-Aguilar, C., Adams, D. J., & Robles-Espinoza, C. D. (2019). Melanoma subtypes: genomic profiles, prognostic molecular markers and therapeutic possibilities. *The Journal of pathology*, *247*(5), 539–551.

Rafi, A. M., Nogina, D., Penzar, D., Lee, D., Lee, D., Kim, N., ... others (2023). Evaluation and optimization of sequence-based gene regulatory deep learning models. *bioRxiv*.

Ramasamy, S., Aljahani, A., Karpinska, M. A., Cao, T. N., Velychko, T., Cruz, J. N., ... Oudelaar, A. M. (2023). The mediator complex regulates enhancer-promoter interactions. *Nature Structural & Molecular Biology*, *30*(7), 991–1000.

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., ... Getz, G. (2015). Oncotator: cancer variant annotation tool. *Human mutation*, *36*(4), E2423–E2429.

Rawla, P. (2019). Epidemiology of prostate cancer. *World journal of oncology*, *10*(2), 63.

Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., ... others (2020). pandas-dev/pandas: Pandas 1.0. 5. *Zenodo*.

Reiter, F., Wienerroither, S., & Stark, A. (2017). Combinatorial function of transcription factors and cofactors. *Current opinion in genetics & development*, *43*, 73–81.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., ... others (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102–111.

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J. M., Kim, J., ... others (2017). Recurrent and functional regulatory mutations in breast cancer. *Nature*, *547*(7661), 55–60.

Richmond, T. J., & Davey, C. A. (2003). The structure of dna in the nucleosome core. *Nature*, *423*(6936), 145–150.

Rickman, D. S., Pflueger, D., Moss, B., VanDoren, V. E., Chen, C. X., de la Taille, A., ... others (2009). Slc45a3-elk4 is a novel and frequent erythroblast transformation–specific fusion transcript in prostate cancer. *Cancer research*, *69*(7), 2734–2738.

Riethoven, J.-J. M. (2010). Regulatory regions in dna: promoters, enhancers, silencers, and insulators. *Computational biology of transcription factor binding*, 33–42.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., Consortium, W., ... Lunter, G. (2014). Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, *46*(8), 912–918.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rojas-Fernandez, A., Herhaus, L., Macartney, T., Lachaud, C., Hay, R. T., & Sapkota, G. P. (2015). Rapid generation of endogenously driven transcriptional reporters in cells through crispr/cas9. *Scientific Reports*, *5*(1), 9811.

Roy, A. L., & Singer, D. S. (2015). Core promoters in transcription: old problem, new insights. *Trends in biochemical sciences*, *40*(3), 165–171.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, *71*, 599–607.

Saito, Y., Koya, J., Araki, M., Kogure, Y., Shingaki, S., Tabata, M., ... others (2020). Landscape and function of multiple mutations within individual oncogenes. *Nature*, *582*(7810), 95–99.

Samuel, S., & Mietchen, D. (2024). Computational reproducibility of jupyter notebooks from biomedical publications. *GigaScience*, *13*, giad113.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., ... others (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, *173*(2), 321–337.

Sarver, A. E., Sarver, A. L., Thayanithy, V., & Subramanian, S. (2015). Identification, by systematic rna sequencing, of novel candidate biomarkers and therapeutic targets in human soft tissue tumors. *Laboratory Investigation*, *95*(9), 1077–1088.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, *18*(20), 6097–6100.

Scholl, C., & Fröhling, S. (2019). Exploiting rare driver mutations for precision cancer medicine. *Current Opinion in Genetics & Development*, *54*, 1–6.

Segert, J. A., Gisselbrecht, S. S., & Bulyk, M. L. (2021). Transcriptional silencers: driving gene expression with the brakes on. *Trends in Genetics*, *37*(6), 514–527.

Sementchenko, V. I., & Watson, D. K. (2000). Ets target genes: past, present and future. *Oncogene*, *19*(55), 6533–6548.

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R. J., ... others (2017). Dna methylation regulates discrimination of enhancers from promoters through a h3k4me1-h3k4me3 seesaw mechanism. *BMC genomics*, *18*, 1–21.

Shaw, A. T., Kim, D.-W., Nakagawa, K., Seto, T., Crinó, L., Ahn, M.-J., ... others (2013). Crizotinib versus chemotherapy in advanced alk-positive lung cancer. *New England Journal of Medicine*, *368*(25), 2385–2394.

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, *15*(4), 272–286.

Shuai, S., Gallinger, S., & Stein, L. D. (2020). Combined burden and functional impact tests for cancer driver discovery using driverpower. *Nature communications*, *11*(1), 734.

Sicklick, J. K., Kato, S., Okamura, R., Schwaederle, M., Hahn, M. E., Williams, C. B., ... others (2019). Molecular profiling of cancer patients enables personalized combination therapy: the i-predict study. *Nature medicine*, *25*(5), 744–750.

Sikorski, T. W., & Buratowski, S. (2009). The basal initiation machinery: beyond the general transcription factors. *Current opinion in cell biology*, *21*(3), 344–351.

Sloutskin, A., Shir-Shapira, H., Freiman, R. N., & Juven-Gershon, T. (2021). The core promoter is a regulatory hub for developmental gene expression. *Frontiers in Cell and Developmental Biology*, *9*, 666508.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

Solt, L. A., & May, M. J. (2008). The iκb kinase complex: master regulator of nf-κb signaling. *Immunologic research*, *42*, 3–18.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, *18*(11), 696–705.

Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window z-curve method. *Nucleic acids research*, *40*(3), 963–971.

Song, L., & Crawford, G. E. (2010). Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb–prot5384.

Spencer, D. H., Zhang, B., & Pfeifer, J. (2015). Single nucleotide variant detection using next generation sequencing. In *Clinical genomics* (pp. 109–127). Elsevier.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., ... others (2016). The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, *54*(1), 1–30.

Swain, S. M., Baselga, J., Kim, S.-B., Ro, J., Semiglazov, V., Campone, M., ... others (2015). Pertuzumab, trastuzumab, and docetaxel in her2-positive metastatic breast cancer. *New England journal of medicine*, *372*(8), 724–734.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... others (2019). Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, *47*(D1), D941–D947.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., ... Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381–386.

Trescher, S., & Leser, U. (2019). Estimation of transcription factor activity in knockdown studies. *Scientific Reports*, *9*(1), 9593.

Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, *9*(1), 60.

Tsimberidou, A. M., Kahle, M., Vo, H. H., Baysal, M. A., Johnson, A., & Meric-Bernstam, F. (2023). Molecular tumour boards—current and future considerations for precision oncology. *Nature Reviews Clinical Oncology*, *20*(12), 843–863.

Tsofack, S. P., Garand, C., Sereduk, C., Chow, D., Aziz, M., Guay, D., ... Lebel, M. (2011). Nono and raly proteins are required for yb-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines. *Molecular cancer*, *10*, 1–18.

Tudose, C., Bond, J., & Ryan, C. J. (2023, October). Gene essentiality in cancer is better predicted by mRNA abundance than by gene regulatory network-inferred activity. *NAR Cancer*, *5*(4), zcad056. Retrieved 2024-03-27, from https://academic.oup.com/narcancer/article/doi/10.1093/narcan/zcad056/7453243 doi: 10.1093/narcan/zcad056

Umarov, R. K., & Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, *12*(2), e0171410.

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., ... Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory dna. *Nature*, *603*(7901), 455–463.

Valencia, C. A., Pervaiz, M. A., Husami, A., Qian, Y., Zhang, K., Valencia, C. A., ... Zhang, K. (2013). Sanger sequencing principles, history, and landmarks. *Next Generation Sequencing Technologies in Medical Genetics*, 3–11.

Van Tilburg, C. M., Pfaff, E., Pajtler, K. W., Langenberg, K. P., Fiesel, P., Jones, B. C., ... others (2021). The pediatric precision oncology inform registry: clinical outcome and benefit for patients with very high-evidence targets. *Cancer discovery*, *11*(11), 2764–2779.

Vasan, N., Razavi, P., Johnson, J. L., Shao, H., Shah, H., Antoine, A., ... others (2019). Double pik3ca mutations in cis increase oncogenicity and sensitivity to pi3kα inhibitors. *Science*, *366*(6466), 714–723.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *science*, *339*(6127), 1546–1558.

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, *15*(10), 776–784.

Wahida, A., Buschhorn, L., Fröhling, S., Jost, P. J., Schneeweiss, A., Lichter, P., & Kurzrock, R. (2023). The coming decade in precision oncology: six riddles. *Nature Reviews Cancer*, *23*(1), 43–54.

Wang, T., Ruan, S., Zhao, X., Shi, X., Teng, H., Zhong, J., ... Mao, F. (2021). Oncovar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic acids research*, *49*(D1), D1289–D1301.

Wang, Z., Wang, P., Li, Y., Peng, H., Zhu, Y., Mohandas, N., & Liu, J. (2021). Interplay between cofactors and transcription factors in hematopoiesis and hematological malignancies. *Signal Transduction and Targeted Therapy*, *6*(1), 24.

Warner, J. L., Sethi, T. K., Rivera, D. R., Venepalli, N. K., Osterman, T. J., Khaki, A. R., & Rubinstein, S. (2020). *Trends in fda cancer registration trial design over time, 1969-2020.* American Society of Clinical Oncology.

Watanabe, K., Taskesen, E., Van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. *Nature communications*, *8*(1), 1826.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., & Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, *46*(11), 1160–1165.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., ... Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113–1120.

Weniger, M. A., & Küppers, R. (2016). Nf-κb deregulation in hodgkin lymphoma. In *Seminars in cancer biology* (Vol. 39, pp. 32–39).

Wetterstrand, K. A. (2019). *Dna sequencing costs: Data from the nhgri genome sequencing program (gsp).* Available at www.genome.gov/sequencingcostsdata. (Accessed 15 April 2024)

Weymann, D., Pataky, R., & Regier, D. A. (2018). Economic evaluations of next-generation precision oncology: a critical review. *JCO Precision Oncology*, *2*, 1–23.

Whitlock, J. H., Wilk, E. J., Howton, T. C., Clark, A. D., & Lasseigne, B. N. (2024, January). The landscape of SETBP1 gene expression and transcription factor activity across human tissues. *PLOS ONE*, *19*(1), e0296328. Retrieved 2024-03-27, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0296328 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0296328

Wood, K., Hensing, T., Malik, R., & Salgia, R. (2016). Prognostic and predictive value in kras in non–small-cell lung cancer: a review. *JAMA oncology*, *2*(6), 805–812.

WorldHealthOrganization. (2024). *Global cancer burden growing, amidst mounting need for services.* Retrieved from

Globalcancerburdengrowing,amidstmountingneedforservices

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, *18*(10), 1161–1168.

Wu, W.-j., Hu, K.-s., Wang, D.-s., Zeng, Z.-l., Zhang, D.-s., Chen, D.-l., ... Xu, R.-h. (2013). Cdc20 overexpression predicts a poor prognosis for patients with colorectal cancer. *Journal of translational medicine*, *11*, 1–8.

Xi, X., Cao, T., Qian, Y., Wang, H., Ju, S., Chen, Y., ... Hou, S. (2022). Cdc20 is a novel biomarker for improved clinical predictions in epithelial ovarian cancer. *American Journal of Cancer Research*, *12*(7), 3303.

Yaeger, R., Weiss, J., Pelster, M. S., Spira, A. I., Barve, M., Ou, S.-H. I., ... others (2023). Adagrasib with or without cetuximab in colorectal cancer with mutated kras g12c. *New England Journal of Medicine*, *388*(1), 44–54.

Yang, L., Li, N., Wang, M., Zhang, Y.-H., Yan, L.-D., Zhou, W., ... Cai, J. (2021). Tumorigenic effect of tert and its potential therapeutic target in nsclc. *Oncology Reports*, *46*(2), 1–12.

Yavuz, B. R., Tsai, C.-J., Nussinov, R., & Tuncbag, N. (2023). Pan-cancer clinical impact of latent drivers from double mutations. *Communications Biology*, *6*(1), 202.

Yin, W., Xiang, P., & Li, Q. (2005). Investigations of the effect of dna size in transient transfection assay using dual luciferase system. *Analytical biochemistry*, *346*(2), 289–294.

Yiu Chan, C. W., Gu, Z., Bieg, M., Eils, R., & Herrmann, C. (2019). Impact of cancer mutational signatures on transcription factor motifs in the human genome. *BMC medical genomics*, *12*, 1–14.

Yu, H., Lin, L., Zhang, Z., Zhang, H., & Hu, H. (2020). Targeting nf-κb pathway for the therapy of diseases: mechanism and clinical study. *Signal transduction and targeted therapy*, *5*(1), 209.

Yuan, X., Dai, M., & Xu, D. (2020). Tert promoter mutations and gabp transcription factors in carcinogenesis: More foes than friends. *Cancer letters*, *493*, 1–9.

Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., ... Satija, R. (2022). Characterizing cellular heterogeneity in chromatin state with sccut&tag-pro. *Nature biotechnology*, *40*(8), 1220–1230.

Zhang, L., Fu, R., Liu, P., Wang, L., Liang, W., Zou, H., ... Tao, L. (2021). Biological and prognostic value of etv5 in high-grade serous ovarian cancer. *Journal of Ovarian Research*, *14*, 1–9.

Zhang, T., Li, L., Sun, H., Xu, D., & Wang, G. (2023). Deepicsh: a complex deep learning framework for identifying cell-specific silencers and their strength from the human genome. *Briefings in Bioinformatics*, *24*(5), bbad316.

Zhang, Y., Chen, F., Balic, M., & Creighton, C. J. (2024). An essential gene signature of breast cancer metastasis reveals targetable pathways. *Breast Cancer Research*, *26*(1), 98.

Zhang, Y., Chen, F., & Creighton, C. J. (2023). Pan-cancer molecular subtypes of metastasis reveal distinct and evolving transcriptional programs. *Cell Reports Medicine*, *4*(2).

Zhang, Y., Gong, M., Yuan, H., Park, H. G., Frierson, H. F., & Li, H. (2012). Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. *Cancer discovery*, *2*(7), 598–607.

Zhao, J., Martin, V., & Gordân, R. (2022). Transcription factor-centric approach to identify non-recurring putative regulatory drivers in cancer. In *International conference on research in computational molecular biology* (pp. 36–51).

Zhou, C., Wu, Y.-L., Chen, G., Feng, J., Liu, X.-Q., Wang, C., ... others (2011). Erlotinib versus chemotherapy as first-line treatment for patients with advanced egfr mutation-positive non-small-cell lung cancer (optimal, ctong-0802): a multicentre, open-label, randomised, phase 3 study. *The lancet oncology*, *12*(8), 735–742.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, *12*(10), 931–934.

Zhu, Z., Guo, Y., Liu, Y., Ding, R., Huang, Z., Yu, W., ... Liu, C.-Y. (2023). Elk4 promotes colorectal cancer progression by activating the neoangiogenic factor lrg1 in a noncanonical sp1/3-dependent manner. *Advanced Science*, *10*(32), 2303378.

Zinatizadeh, M. R., Schock, B., Chalbatani, G. M., Zarandi, P. K., Jalali, S. A., & Miri, S. R. (2021). The nuclear factor kappa b (nf-kb) signaling in cancer development and immune diseases. *Genes & diseases*, *8*(3), 287–297.

Zrimec, J., Fu, X., Muhammad, A. S., Skrekas, C., Jauniskis, V., Speicher, N. K., ... others (2022). Controlling gene expression with deep generative design of regulatory dna. *Nature communications*, *13*(1), 5099.

Abbreviations and Acronyms

Acronym	Meaning
ANKRD53	Ankyrin Repeat Domain-Containing Protein 53
BAM	Binary Alignment Map
bp	Basepair
CDC20	Cell Division Cycle 20
chr	Chromosome
CGC	Cancer Gene Census
CNN	Convolutional Neural Network
COGNITION	Comprehensive assessment of clinical features, genomics and further molecular markers to identify patients with early breast cancer for enrolment on marker driven trials trials
COSMIC	Catalogue of Somatic Mutations in Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DREAM	Dialogue on Reverse Engineering Assessment and Methods
ELK4	ETS Transcription Factor ELK4
ETS	E-twenty-six
FIMO	Find Individual Motif Occurrences
FOXD1	Forkhead Box D1
FOXO3	Forkhead Box O3
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GABPA	GA Binding Protein Transcription Factor Subunit Alpha
GTF	General Transcription Factor
HEK293FT	Human embryonic kidney 293 cells
ICGC	International Cancer Genome Consortium
Indel	Insertion or Deletion
JSON	JavaScript Object Notation
LEPROTL1	Leptin Receptor Overlapping Transcript Like 1
LSTM	Long short-term memory
MAF	Minor Allele Frequency
ML	Machine learning
MLM	Multivariate Linear Model

Acronym	Meaning
MUT	Mutant
MTB	Molecular Tumor Board
MYB	V-Myb Avian Myeloblastosis Viral Oncogene Homolog
NCT-MASTER	Molecularly Aided Stratification for Tumor Eradication Research
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear Factor kappa-light-chain-enhancer of activated B cells
NGS	Next-Generation Sequencing
PCAWG	Pan-cancer Analysis of Whole Genomes
PIC	Pre-initiation Complex
PID	Patient ID or Patient Identifier
pSNV	Promoter Single Nucleotide Variant
RALY	Heterogeneous Nuclear Ribonucleoprotein
RELA	Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B-Cells 3
REMIND-Cancer	Regulatory Mutation Identification 'N' Description in Cancer
RNA Poly II	RNA Polymerase II
RNA-Seq	RNA Sequencing
SBS	Single base substitution
SNV	Single Nucleotide Variant
TCGA	The Cancer Genome Atlas
TERT	Telomerase reverse transcriptase
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TPM	Transcripts Per Kilobase Million
TSS	Transcription Start Site
ULM	Univariate Linear Model
UV	Ultraviolet
VAF	Variant Allele Frequency
VCF	Variant Call File
VIPER	Virtual Inference of Protein-activity by Enriched Regulon analysis
WGS	Whole Genome Sequencing
WSUM	Weighted Sum
WT	Wild Type