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

M.Sc Chen Hong

born in Maanshan, Anhui, China

Oral examination: 04.09.2023

Analysis of whole-genome sequencing data from ICGC-PanCancer project

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Stefan Wiemann

Abstract

Cancer is one of the greatest health challenges of the 21st century and one of the deadliest diseases in the world. It is a group of different diseases which are caused by abnormal cell growth. In the human body, cell division and apoptosis are well regulated under normal circumstances so that the number of cells is in a dynamic balance. However, normal cells could transform into tumor cells because of genetic mutations. The tumorigenesis can happen in almost any cell of the human body. One of the central tools to address cancer is the profiling of cancer cell genomes and transcriptomes by next generation sequencing (NGS) and subsequent analysis by computational methods.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) project is the core project of the International Cancer Genome Consortium. This project provides massive amounts of cancer biological data for analysis. Include more than 2900 patients and 48 types of cancer samples. As part of this intensive effort, I have conducted a very detailed analysis on the molecular mechanisms of cancers. In particular, I conducted a comprehensive study of the relationship between genomic mutations and cancer development. These series of studies include the exploration of cancer driver genes, analysis of telomere maintenance mechanisms and data visualization at the cohort level.

First, I explored potential cancer genes by performing statistical analysis of genomic point mutations, insertions and deletions, copy number variations and structural variations. Further, I analyzed the distribution of point mutations and structure variations in cancer genomes. Based on Knudson's two-hit hypothesis, I integrated point mutation and copy number variation information to construct a biallelic inactivation map of the cancer genome. With the biallelic inactivation information, I analyzed potential cancer drivers and applied this finding to synthetic lethality assays associated with cancer driver genes to uncover novel genetic targets that could be used to treat cancer patients with certain driver gene defects. In addition, I designed and improved the CaSINo model to score the relative mutation frequency of chromosomal sequences to screen for potential cancer driver mutations, which can be used not only in coding genes but also in non-coding regions. Moreover, I analyzed point mutations on promoters, trying to find those mutation sites that play a key role in the up-regulation of gene expression. Finally, I designed and improved a scoring method for copy number variation focality to explore the association of focal copy number variation with cancer driver genes at the cohort level.

Second, as part of the PCAWG research projects, I analyzed the mechanisms of telomere maintenance in cancer cells. After analyzing the differences between alternative telomere lengthening and telomerase-positive samples, I designed a machine learning model based on repeat sequences, content, and mutation rate to determine whether an unknown cancer sample is an alternative lengthening of telomere (ALT) or telomerase-positive.

Finally, for the massive data of the PCAWG project, I designed and implemented two bioinformatics visualization tools. TumorPrint is software in R and shell, which can be used to visualize genomic mutations and RNA-seq expression levels of a single gene or gene pairs, allowing users to quickly search for genes or gene pairs of interest. GenomeTornadoPlot is a software written in the R language for visualizing focal copy number variants of a single gene or adjacent paired genes, and can automatically calculate its copy number variation aggregation score.

Zusammenfassung

Krebs ist eine der größten gesundheitlichen Herausforderungen des 21. Jahrhunderts und eine der tödlichsten Krankheiten der Welt. Genauer gesagt handelt sich bei Krebs um eine Gruppe verschiedener Krankheiten, die durch abnormales Zellwachstum verursacht werden. Im menschlichen Körper sind Zellteilung und Apoptose unter normalen Umständen gut reguliert, so dass sich die Zahl der Zellen in einem dynamischen Gleichgewicht befindet. Allerdings können sich normale Zellen aufgrund von Genmutationen in Tumorzellen verwandeln. Die Tumorentstehung kann in fast jeder Zelle des menschlichen Körpers stattfinden. Eines der wichtigsten Instrumente zur Bekämpfung von Krebs ist die Erstellung von Profilen der Genome und Transriptome von Krebszellen durch Next-Generation-Sequencing (NGS) und die anschließende Analyse durch computergestützte Methoden.

Das Projekt Pan-Cancer Analysis of Whole Genomes (PCAWG) ist das Kernprojekt des International Cancer Genome Consortium (ICGC). Es stellt enorme Mengen an biologischen Krebsdaten zur Analyse bereit, und umfasst mehr als 2900 Patienten die sich 48 Arten von Krebs zuordnen lassen. Im Rahmen dieser Doktorarbeit habe ich eine detaillierte Analyse der molekularen Mechanismen von Krebserkrankungen durchgeführt. Der Schwerpunkt dieser Untersuchungen bildet dabei der Zusammenhang zwischen genomischen Mutationen und der Krebsentwicklung. Diese Studienreihe umfasst die Erforschung von Krebstreibergenen, die Analyse von Telomererhaltungsmechanismen und die Visualisierung von Daten auf Kohortenebene.

Zunächst untersuchte ich potenzielle Krebsgene, indem ich eine statistische Analyse von genomischen Punktmutationen, Insertionen und Deletionen, Kopienzahlvariationen und Strukturvariationen in Krebsgenomen durchführte. Auf der Grundlage der Two-Hit-Hypothese von Knudson habe ich Informationen über Punktmutationen und Kopienzahlvariationen integriert, um eine Karte der biallelischen Inaktivierung des Krebsgenoms zu erstellen. Diese Informationen

nutzte ich um potenzielle Krebstreiber mit Hilfe von synthetische Letalitätsassays zu suchen, und somit neue genetische Ziele zu entdecken, die zur Behandlung von Krebspatienten mit bestimmten Treibergen-Defekten verwendet werden könnten. Darüber hinaus habe ich das CaSINo-Modell zur Bewertung der relativen Mutationshäufigkeit chromosomaler Sequenzen entwickelt und verbessert, um nach potenziellen Krebstreibermutationen zu suchen, die nicht nur in kodierenden Genen, sondern auch in nicht kodierenden Regionen auftreten können. Außerdem analysierte ich Punktmutationen an Promotoren, um diejenigen Mutationsstellen zu finden, die eine Schlüsselrolle bei der Hochregulierung der Genexpression spielen. Schließlich habe ich eine Scoring-Methode für die Fokalität von Kopienzahlvariationen entwickelt und verbessert, um die Assoziation von fokalen Kopienzahlvariationen mit Krebstreibergenen auf Kohortenebene zu untersuchen.

Zweitens analysierte ich im Rahmen der PCAWG-Forschungsprojekte die Mechanismen der Telomererhaltung in Krebszellen. Basierend auf Telomerelängenvorhersagen, genetischen Mutationen und der Frequenz von Telomererepeatvarianten entwickelte ich ein maschinelles Lernmodell, das die Aktivierung von alternativer Telomerverlängerung (ALT) vorhersagen kann.

Schließlich habe ich für die umfangreichen Daten des PCAWG-Projekts zwei Bioinformatik-Visualisierungstools entwickelt und implementiert. TumourPrint ist eine Software in R und Shell, mit der genomische Mutationen und RNA-seq-Expressionsniveaus eines einzelnen Gens oder von Genpaaren visualisiert werden können, so dass die Benutzer schnell nach Genen oder Genpaaren von Interesse suchen können. GenomeTornadoPlot ist eine in R geschriebene Software zur Visualisierung von fokalen Kopienzahlvarianten eines einzelnen Gens oder benachbarter Genpaare und kann automatisch den Aggregationswert der Kopienzahlvariationen berechnen.

Table of Contents

List of figures

List of tables

List of supplements

Abbreviations and acronyms

ALK	anaplastic lymphoma kinase
ALT	alternative lengthening of telomeres
AML	acute myeloid leukemia
Bladder-TCC	bladder transitional cell carcinoma
BMR	background mutation rate
BRCA	breast cancer
CDS	coding region sequences
CGC	Cancer Gene Census
CLC	Cancer LncRNA Census
CN	copy number
CNV	copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	Deoxyribonucleic acid
ESAD	esophageal adenocarcinoma
FDR	false discovery rate
FPKM	Fragments Per Kilobase per Million
ICGC	the International Cancer Genome Consortium
INDEL	insertion or deletion
IPA	Qiagen Ingenuity Pathway Analysis
IQR	interquartile range
LAML	acute myeloid leukemia
lncRNA	long non-coding RNA
Lung-scc	lung squamous cell adenocarcinoma
LUSC	lung squamous cell carcinomas
ME	mutual exclusive

MutsigCV	Mutation Significance Co-Variates
NGS	next generation sequencing
NSCLC	Non-small cell lung cancer
OG	oncogene
OS	overall survival
OV	ovarian cancer
PCA	Principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PDB	protein database
PML	promyelocytic leukemia
PTEN	Phosphatase and tensin homolog gene
RET	rearranged during transfection
RNA	Ribonucleic acid
ROC curve	receiver operating characteristic curve
SKCM	skin cutaneous melanoma
SNV	single nucleotide variation
SV	structural variation
TCGA	The Cancer Genome Atlas
TERT	telomerase reverse transcriptase
TFBS	transcription factors binding site
TMM	telomere maintenance mechanisms
TSG	tumor suppressor gene
TSS	transcription start site
TVR	telomere variant repeat
VCF	variant call format
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

WHO	World Health Organization

1 Introduction

1.1 Cancer Research and Bioinformatics

Cancer is a group of distinct diseases that are caused by abnormal cell growth. Under normal circumstances, cell division and apoptosis are regulated programmely so that the number of cells is in a dynamic balance. However, normal cells can transform into tumor cells because of genetic variations. tumorigenesis can happen in almost any cell in the human body. Different from benign tumors, malignant cancer cells are able to spread and invade other tissues or parts of the human body. In the late stage of cancer progression, the lethal cancer cells disrupt the functions of important organs and tissues of the human body, damage normal metabolism, and waste plenty of nutrition. Uncontrollable proliferation of tumor cells can be lethal.

Cancer is one of the deadliest diseases in the world. According to a study by the World Health Organization (WHO), 18.1 million new cancer cases were found in 2018 and more than 9.6 million people died of cancers. (“Erratum,” 2020) Experts expected that the prevalence of cancers would increase in the future.

Bioinformatics has been a very powerful weapon in the fight against cancers. Over recent decades, the improving high-throughput technologies provided massive amounts of data, including not only traditional clinical information but also -omics data. Here -omics refers to genomics, transcriptomics, epigenomics, and other information and refers to the complete characterization of a particular layer of cell biology with a single molecular biological assay. Scientists integrated data from different platforms through bioinformatic approaches. Meanwhile, bioinformatics helps them discover insights from these datasets and accelerate biological studies.

Cancer is caused essentially by the accumulation of DNA variants in the genome. Because of this, next-generation sequencing (NGS) plays an important role in cancer diagnosis, treatment choice, and as a research tool. NGS has been steadily improved in recent decades. This progress made sequencing cheaper, faster, and more accurate. Nowadays, NGS technology is widely used in whole-genome sequencing (WGS), whole-exome sequencing (WES), panel sequencing, epigenomics sequencing, and RNA sequencing (RNA-seq). By the end of the year 2020, although the sequencing cost for a single human genome has not yet decreased to 1,000 US dollars as predicted, it is significantly lower than before. (Gordon et al., 2020; Plöthner et al., 2017) It produces NGS data for cancer patients at a steadily increasing rate.

On the other hand, big biomedical data poses a great challenge to analytical methods. Accordingly, a growing relevance is attributed to bioinformatics techniques over the last decades. The main focus of cancer bioinformatics nowadays is storing, analyzing, integrating, accessing, and visualizing large amounts of biological data and related information.(Demir et al., 2004) Higher performance hardware and better-optimized analysis methods, like machine learning, accelerate cancer research efficiency remarkably.

1.2 Pan-Cancer Analysis of Whole Genomes (PCAWG) project

In recent decades, along with technological improvement and the decrease in sequencing costs, several international cancer genome research projects have been launched. The Cancer Genome Atlas (TCGA), which started in 2005, is the pioneer of the Cancer Genomics Initiative. In this massive program, researchers have collected genomic, epigenomic, transcriptomic, and proteomic data from more than 20,000 primary cancers and matched normal samples from 33 different cancer types. (Tomczak et al., 2015)

As a reaction to this realization, the Pan-Cancer Analysis of Whole Genomes (PCAWG) project was called into life. It is combined with ICGC and TCGA working groups’ research results. The project aims to generate catalogs of genome abnormal variations in cancer, which include somatic mutations, epigenetic changes and expression abnormalities. Scientists can rapidly and freely access data from tumors of more than 2,900 patients from 48 different types or subtypes of cancers. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020)

These projects are comprised of WGS data from clinical centers around the world. The WGS data of each patient includes at least one tumor sample and one control sample, which were sequenced by Illumina HiSeq with 100-150bp paired-end sequencing reads and an average coverage of at least 30 reads in tumor samples and 25 in control samples. The control samples were collected from blood or healthy tumor-adjacent tissues in view of cancer types. In some cases, additional samples were collected. For all patients, clinical data, including patient ID, age, sex, survival time, etc., are also gathered. Moreover, the dataset provides RNA sequencing data for approximately two thirds of the patients. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020)

Compared to the previous The Cancer Genome Atlas (TCGA) project, which is mostly focused on coding sequences, the PCAWG project concerns somatic and germline sequences for both coding and non-coding regions of cancer genomes. The PCAWG dataset provides not only whole genome sequencing data (WGS), which can be discovered for somatic mutations like single nucleotide variations (SNVs), structural variations (SVs), copy number alterations (CNVs), and gene fusions, but also RNA sequencing and methylation data. The collection of multi-omics data enhances the chance to discover the complicated functions of combinations of mutations.

Because of the artifacts contained in the raw data and lack of standards for dealing with them, the mutation calls from different pipelines are not in concordance. (Alioto et al., 2015) However, a substantial advantage of PCAWG is that its data is reprocessed by a homogenous computational workflow. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020) The advantage guarantees the reliability and reproducibility of the data so that the pan-cancer genome comparative analysis is no longer burdened by biases introduced by inconsistencies in the computational processing of the data.

In the year 2021, more than 20 papers based on the PCAWG project were published in Nature Research Journals. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020)

1.3 Cancer Biological Problems

In past years, scientists exerted a lot of effort in attempting to discover how cancers are initiated and how they develop. According to our preliminary knowledge and the newest cancer research trend, I listed a few study directions in which we could benefit from PCAWG data analysis.

1.3.1 Identification of Cancer Driver Mutations

Many cancers are the result of somatic mutations in genomes. However, not all the mutations found in tumor tissue genomes are carcinogenic. Only a subset called driver mutations, either gain of function or loss of function, lead to cell tumorigenesis. (Tokheim et al., 2016) The rest of them are harmless and neutral in cancer development and are named “passenger” mutations.

For a long time, due to the diversity of cancer development and the lack of sufficiently large datasets low-frequency variations and complicated driver principles were often neglected or underestimated. Because of the PCAWG project, now the availability of massive data and high-performance algorithms makes it possible to identify new potential cancer drivers.

Since scientists realized the causal relationship between tumor development and genetic alterations, they have set their sights on discovering cancer drivers. In 1982, the first cancer driver mutation, a G to T transversion in the P21 protein-coding region, was found in the T24 human bladder carcinoma oncogene. (Reddy *et al.*, 1982) In the next thirty-five years, almost 600 genes were identified as cancer driver genes. A list of the most relevant cancer drivers includes the genes TP53, MDM2, KRAS, PTEN. (Hamarsheh et al., 2020; Hou et al., 2019; Milella et al., 2015; Olivier et al., 2010)

The identification of cancer drivers is very helpful in cancer diagnosis and treatment. Because whole-genome sequencing technology is now available for patients, driver mutations can be applied as cancer biomarkers in the prediction of cancer progression. Additionally, knowledge of cancer drivers promoted the development of targeted therapy. For example, the drugs, which were designed to inhibit the expression of mutated ERFG and ALK genes, had an excellent effect in the treatment of non-small cell lung cancer. (Mello et al., 2016)

The oral drug Alectinib is an interesting instance. It blocks anaplastic lymphoma kinase (ALK) activity and is used in the treatment of non-small cell lung cancer (NSCLC). The active ingredient in the drug blocks the ALK and is rearranged during transfection (RET) proto-oncogenes. Inhibition of ALK results in blockade of cellular signaling pathways, including STAT3 and PI3K/AKT/mTOR pathways, and induction of tumor cell apoptosis. This drug is highly effective and has low toxicity. (Avrillon and Pérol, 2017)

TERT which plays a key role in cancer formation is another instance. It ensures chromosomal stability by maintaining telomere length and provides the possibility for unlimited cell proliferation. Therefore, telomerase inhibitors may be given to the patient in order to damage tumor cells. Although still in the research stage, cancer treatment targeting TERT has great promise. (Liu et al., 2012)

Validating a “real” cancer driver needs expensive wet-lab experiments. In consequence, many bioinformatics tools were designed to identify potential cancer drivers to shorten the gene candidate list, which will be later sent to molecular biology labs. MutsigCV (Mutation Significance Co-Variates) is one of the advanced tools for analyzing mutations in the genome and screening out the genes which are mutated more frequently than the background mutation rate (BMR). MutsigCV uses patient-specific and gene-specific BMRs individually, which are calculated by non-silent, non-coding, and silent mutations of a “center” gene and genes nearby. For the gene-specific BMR, the genes with similar features, such as expression, replication timing, or HiC data, are defined as a bagel and the local BMR will be calculated for the “center” gene in the bagel. Genes with a non-silent mutation rate higher than BMR and a false discovery rate (FDR) q-value smaller than 0.1 will be identified as a significant mutated genes.(Lawrence et al., 2013) There are also other computational methods, such as OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), OncodriveClust (Tamborero et al., 2013), ActiveDriver(Reimand and Bader, 2013) and MuSIC (Dees et al., 2012), that focus on different aspects such as analysis of recurrence, spatial clustering or predicted impact of somatic mutations to identify new candidates from whole-genome sequencing data. These statistical and machine learning-based approaches accelerated the research and found several interesting genes such as HLA-A, FLNB, GRM1, and POU2F1. (Lawrence et al., 2013; Reimand and Bader, 2013)

However, identifying functional cancer driver genes from thousands of genes in the whole genome is not a simple task. Many potential candidates remain undetected due to the lack of power to discriminate driver mutations from the background mutational load. The most direct way to solve this problem is to increase the sample size. (Hofree et al., 2016) Another problem is that we don't have a gold standard for "real" cancer driver genes. The lack of ground truth makes many statistical methods vulnerable. (Tokheim et al., 2016) Last but not least, these studies mostly focus on mutations in coding regions, and thus the variations in functional non-coding sequences are neglected. Consequently, driver mutations in non-coding, functional elements are like the deep sea that has never been explored before, waiting for us to study.

1.3.2 Non-coding Mutations and Cancer

Non-coding DNA refers to the DNA sequence in the genome that does not encode proteins. It is very common in eukaryotic cells. For example, it is reported that 98% of the human genome is non-coding sequences. Previously, non-coding DNA was thought to be not biologically functional and used to be called "Junk DNA". Now, it is realized that substantial parts of it have strong biological activity and are crucial in the biological process. (Comfort, 2015) For example, non-coding DNA contains regulatory elements which could upregulate or downregulate gene expression. There are special binding sites in these sequences which can be recognized and bound by special transcription-related proteins. There are four classes of non-coding functional elements: promoters, enhancers, silencers, and insulators. (Figure 1)

Promoters are typically located ahead of the coding region of genes on the DNA strand. They can be recognized by RNA polymerase. The assembly of the RNA polymerase complex then launches the process of transcription.

Enhancers provide binding sites for different types of proteins that help activate transcription. In contrast, silencers are able to bind to proteins for transcription repression. These two elements can upregulate and downregulate gene expression. They can be either located near their target genes or far away from them.

Insulators are the regulatory elements of eukaryotic genomes and work as an enhancer inhibitors from a distance. They can also impact promoters' functions by binding to specific proteins to regulate expression.

According to the research on coding gene regions, it is known that the upregulation of oncogenes and downregulation of tumor suppressor genes (TSGs) are carcinogenic. Therefore, it is easy to understand that genomic abnormalities in the non-coding regulatory elements which are impacting cancer-related genes also have the potential to be cancer drivers, similar to the functions of copy number changes or point mutations in coding regions. (Zhang and Meyerson, 2020)

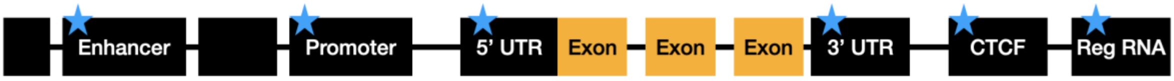


Figure 1: Mutations on non-coding regions related to cancer

The mutations on enhancers, promoters, 5' UTRs, 3' UTRs, CTCF, or regulatory RNA sequences can lead to cancer development.

The Telomerase reverse transcriptase (TERT) gene is the most famous example. Telomerase activity is correlated with the number of times a cell can divide. The overexpression of TERT spurs the level of telomerase and then lengthens telomeres in cells. This abnormality causes immortality of cell lines and leads to carcinogenesis. In past years, scientists observed recurrent mutations in TERT promoters in more than 50 types of cancers. (Bell et al., 2016) Although these mutations do not occur in coding regions, they still have a strong influence on the expression level of TERT. (Bell et al., 2016; "Erratum," 2020)

TERT promoter mutations are famous because of the extremely high recurrence among cancers. There are also other instances of non-coding mutations which drive cancer development. The promoter mutation hotspots of FOXA1 are recurrently found in breast cancer. (Rheinbay et al., 2017) In various types of cancers, like lung cancer and acute myeloid leukemia (AML), point mutations in enhancers of MYC play important roles. (Lancho and Herranz, 2018)

However, if we talk about the recurrence of these mutations, we noticed that they are obviously lower than TERT. It could be because of not only the low coverage of promoter sequencing due to the high GC-content, but also because of the robustness and mutational endurance of transcription regulatory elements. Without a large dataset, it is very difficult to screen the potential non-coding driver mutations by recurrency analysis. (Elliott and Larsson, 2021)

With the development of understanding genomes, scientists realized non-coding DNAs are not limited to the cis-regulatory elements. For instance, long non-coding RNAs (lncRNAs) are a promising hot topic in cancer research. These genes do not encode proteins but the long RNA molecules play crucial roles in the regulation of transcription, post-transcription, and epigenetics levels. MALAT1, as a good example, plays a role as an oncogene in many different cancer types. It is known as a housekeeping gene in splicing progress. Overexpression of MALAT1 is observed in multiple cancers. (Carlevaro-Fita et al., 2020) The experiments in vivo showed that the suppression of MALAT1 can decrease the efficiency of cell proliferation and metastasis. (Gutschner et al., 2013)

Based on the PCAWG data, we can look into interesting cancer-related non-coding mutations even if their recurrences are low. With matched SNV, CNV, and RNA sequencing data, we are able to build a complete landscape of non-coding DNA mutations and their roles in tumors.

1.3.3 Functional Promoter Mutations

A promoter is a DNA region that initiates the transcription of genes and is located typically the near upstream of the transcription start site. RNA polymerase and transcription factors bind to promoters and initiate gene transcription. Although it is difficult to identify promoters in eukaryotic cells due to the diversity, we now know that there are several functional elements, such as TATA box, BRE, and INR box, that occur in many promoters and play an important role in enzyme and transcription factors binding. RNA polymerase II binds to the associated promoter and initiates the transcription. (Roy and Singer, 2015)

Transcription factors are proteins that can bind to a specific nucleotide sequence in upstream of a gene and can regulate its transcription. In the region of gene promoters, these specific nucleotide sequences are called transcription factor binding sites (TFBSs). Generally, transcription factors bind to their target promoters through the recognition of short DNA motifs whose length is usually about 6-8 bp. (Smith and Matthews, 2016) The transcription factors can upregulate- and downregulate the expression of its corresponding gene . These regulations keep the steady state of gene expression, especially the functions of TSGs and oncogenes. (Figure 2) However, if the transcription factors for cancer driver genes do not work well, for example, if the binding of transcription factors are prohibited by other proteins or the binding site structure is damaged, cancer development and progression may occur. (Capasso et al., 2020)

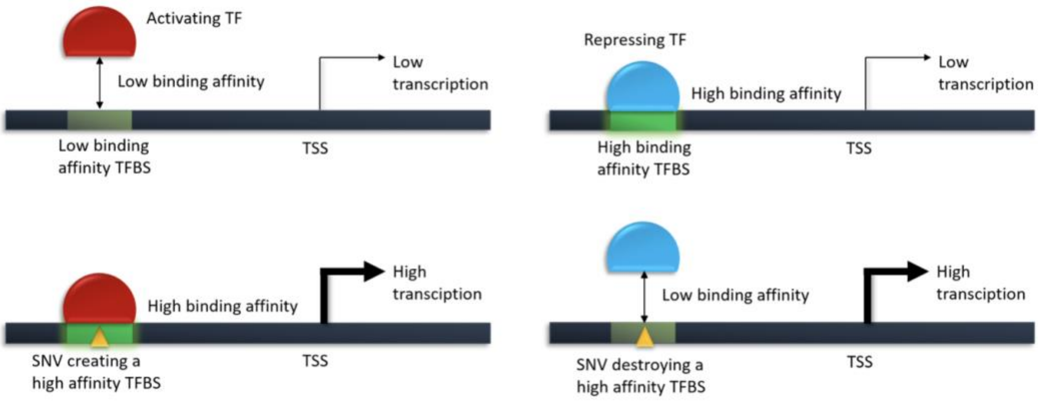


Figure 2: Gene Activation through TFBS altering SNVs

The TFs that bind a promoter region determine to a large extent the expression level of the associated gene. This figure shows the effect of TFBS alterations for a single TF, which in turn affects gene transcription, eventually also by interaction with further TFs. A promoter mutation creates a new TFBS and subsequently an activating TF binds to the promoter causing increased gene expression of the corresponding gene. In the initial state, a repressing TF is tightly bound to its TFBS in the promoter resulting in gene silencing. The introduction of a disruptive SNV leads to the dissociation of the TF and therefore, the respective gene is transcribed. (adapted from Irina Glas,2019)

Genomic variations in TFBSs are able to destroy the TFBS function and inhibit the transcription factor-induced regulation. Unlike large-scaled CNVs, the dysregulation of TFBSs can accurately modify the expression of a single gene.

A study by Vorontsov et al. indicated that mutations in a TFBS are under negative selection in cancer, suggesting that TFBS mutations are generally rare in cancer compared to other genomic regions. (Vorontsov et al., 2016) As we know, promoter regions are conserved more than other non-coding regions, meaning that any variations in these areas are slightly interesting in comparison to their neighbors in cancer patients. TERT is one of the most well-known oncogenes. The overexpression of TERT can disturb the normal function of the telomere and may cause the unlimited replication of a tumor cell. The promoter mutation which upregulates TERT expression is a classic example. Mutations in TFBS regions of this gene may be the root cause of most melanomas. (Horn et al., 2013) In many types of cancers, there are two common mutation sites in the promoter region of TERT that can up-regulate expression, which are located at -124 and -146 base pairs upstream of the transcription initiation site, are known as C228T and C250T mutations. These mutations in TERT promoters are reported to be cancerous and result in a worse prognosis.(Arantes et al., 2020; Horn et al., 2013; Powter et al., 2021) These two mutations create an ETS-1 binding motif and thus increase the expression of TERT. (Liu et al., 2014) Besides the famous gene TERT, cancer-related genes such as PLEKHS1, WDR74, SDHD, and FOXA1, are also impacted by promoter mutations in TFBSs. (Gan et al., 2018)

1.3.4 Biallelic Inactivation and Cancer

The two-hit hypothesis, which was put forward by Knudson in 1971, claimed that most tumor suppressor genes require two alleles to be inactivated to cause phenotypic changes. (Knudson, 1971) At the DNA mutation level, we can look into biallelic inactivations and find evidence of a double-hit hypothesis on potential cancer driver genes.

Biallelic inactivations are caused by different types of variations. Normally, one functional mutation of one allele and then gene deletion of the remaining allele leads to biallelic inactivation. Moreover, the homozygous deletions or homozygous functional mutations can also deactivate both alleles. (Figure 3)

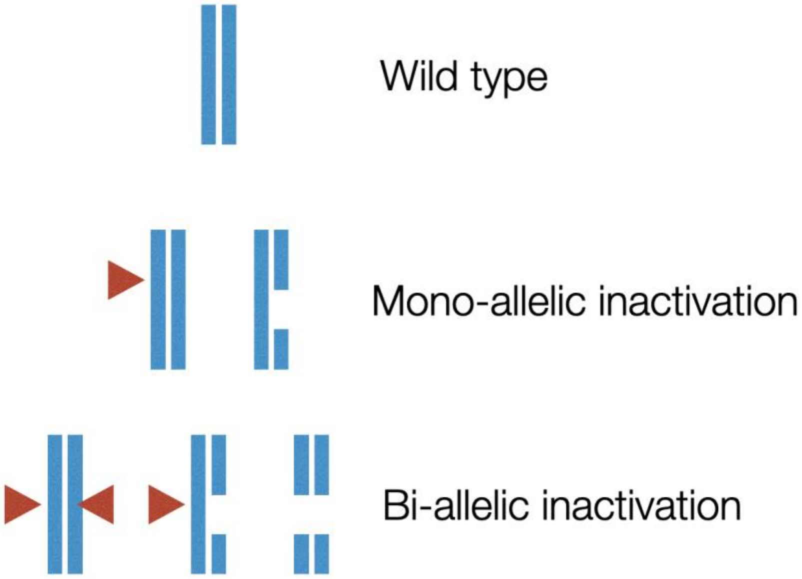


Figure 3: Different types of biallelic inactivations

Gaps stand for deletions and triangles represent functional point mutations, including non-synonymous SNVs.

As early as 1998, Veigl et al. claimed that biallelic inactivations of tumor suppressor genes can result in many cancers. (Veigl et al., 1998) Many well-known tumor suppressor genes were found to be linked with biallelic inactivations. For example, the famous TSG TP53 and PTEN are often biallelically inactivated in several types of cancers. (Kurose et al., 2000; Malcikova et al., 2009; Molinari and Frattini, 2014) FHIT is observed with biallelic deletions in the majority of breast and lung cancers. (Ismail et al., 2011; Yang et al., 2002) In clear cell renal cell carcinoma, melanoma, and neuroblastoma, CDKN2A/B are often biallelically inactivated. (Girgis, 2017; Zeng et al., 2018)

In the past, biallelic inactivations were widely believed to be very rare in cancer genomes. However, Sabarinathan et al. conducted a comprehensive analysis to determine the frequency of these biallelic mutations affecting tumor suppressor genes in both germline and somatic contexts across various tumor types in PCAWG cancer patients. According to their report, over 90% of tumors harbored at least one "double hit" in quite a few cohorts, such as bladder transitional cell carcinomas (Bladder-TCC), lung squamous cell adenocarcinomas (Lung-SCC), and Pancreatic adenocarcinomas (Panc-AdenoCA). (Sabarinathan et al., 2017)

Compared to monoallelic functional mutations or deletions, the biallelic inactivations are more likely to be true hits. They are less impacted by sequencing artifacts and passenger mutations than point mutations or monoallelic deletions. Therefore, the analysis of biallelic inactivation is a powerful instrument for cancer driver identification. In their study, Cheng et al. compiled a comprehensive collection of 2,218 primary tumors spanning 12 different human cancer types. Their primary objective was to systematically investigate homozygous deletions with the aim of discovering infrequent tumor suppressor genes. Through their analysis, they successfully identified 16 well-established tumor suppressors and put forward 27 potential candidate tumor suppressor genes. Notably, among these genes, MGMT, RAD17, and USP44 were already recognized as tumor suppressor genes prior to this study. (Cheng et al., 2017)

These exciting studies make us eager to perform biallelic inactivation analysis on PCAWG data. We will have the opportunity to explore the difference between the effects of gene expression of monoallelic inactivation and biallelic inactivation. It is also possible to understand the frequency of repeated occurrences in different cancer types through simple statistical methods in an attempt to identify new cancer driver genes.

1.3.5 Synthetic Lethality of Cancer-Related Genes

Synthetic lethality is defined as the "simultaneous perturbation of two genes, which results in cellular or organismal death". (Nijman, 2011) In the housekeeping pathways of human cells, a perturbation of inactivation of one gene would inhibit the whole pathway. In this case, one or several alternative pathways would be activated and proceed with normal cell function. However, if the key genes in the alternative pathways are deactivated as well, these housekeeping pathways will become dysfunctional causing the cell to die. (Figure 4)

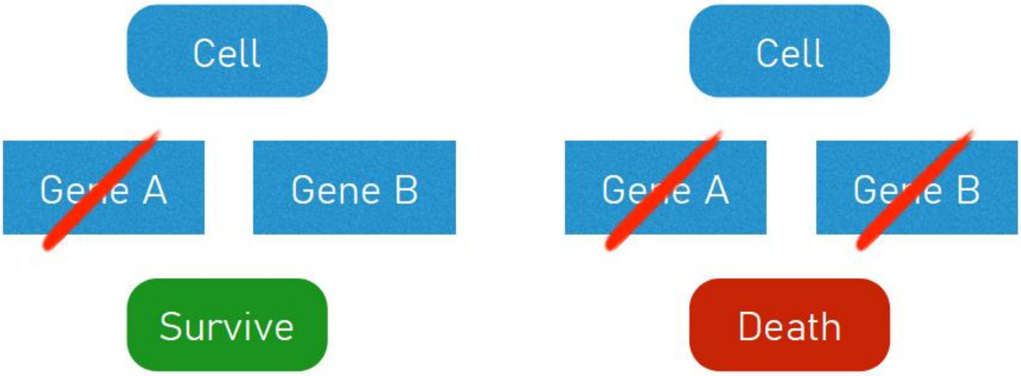


Figure 4: Schematic diagram of synthetic lethality.

If any one of gene A or gene B remained activated, the cell would survive. If both gene A and gene B were deactivated, the cell dies.

The concept of synthetic lethality has garnered considerable interest in recent times due to its potential for a novel class of cancer medications. In the past, the majority of molecularly targeted cancer drugs focused on a single cancer gene or protein. However, not all of these genes are treatable with conventional approaches.

In 1997, Lee Hartwell and Stephen Friend proposed the idea that cancer cells have already suffered from molecular changes that distinguish them from normal body cells. Hence, cancer cells have different genetic vulnerabilities from healthy tissues. Scientists have the opportunity to develop drugs that target these weaknesses of cancer cells so that healthy cells would be free of damage. (Hartwell et al., 1997)

The loss of functionality of TSGs like TP53, RB1, and BRCA1 are crucial causes for many cancer patients. Unlike the overexpression of oncogenes, the loss of TSGs cannot be treated by traditional targeted drugs because the biological molecular functions are already lost. In this case, the idea of inhibition of their synthetic lethality partners seems to be a promising alternative. So, the identification of their SL partner genes becomes a valuable research goal.

PARP inhibitors in BRCA-mutant ovarian cancers was the earliest application of synthetic lethality to target defection of tumor suppressor gene. According to Farmer H. et al., PARP, BRCA1, and BRCA2 are components of two efficient DNA repair pathways. Those tumor cells with functional BRCA1 or BRCA2 mutations are more sensitive to PARP inhibition compared to the healthy ones with at least one copy of BRCA1 or BRCA2. (Farmer et al., 2005) While the precise mechanism behind the synthetic lethality of PARP-BRCA1 and PARP-BRCA2 remains unclear, it did not prevent the FDA from approving four PARP inhibitors for clinical medicine with BRCA-mutant cancers: niraparib, olaparib, rucaparib and talazoparib. (Huang et al., 2020) A similar application was also developed for MTAP deficient patients. PRMT5 inhibitors have proved useful in their treatments. (Kryukov et al., 2016)

In chapter 3 of this thesis, we are interested in Phosphatase and tensin homolog gene (PTEN). Mutations in the PTEN gene are known to play a crucial role in the pathogenesis of various cancer types, including glioblastoma, lung cancer, breast cancer, and prostate cancer. The PCAWG dataset provides a sufficient cohort size of PTEN-deficient samples, which allows us to perform a robust, mutually exclusive statistical model for analysis.

1.3.6 Mutations and Telomere Lengthening Mechanisms

Telomeres are the regions located at the linear chromosome both ends of eukaryotic cells. In the mid-1980s, telomerase was discovered by scientists. (Greider and Blackburn, 1985) When cell DNA replication is terminated by telomerase, DNA can replicate through the telomere-dependent template to compensate for the shortening of the ends caused by the removal of primers. Therefore, telomerase is essential in the maintenance of telomeres. As the number of cell divisions increases, the length of telomeres gradually shortens. When telomere length approaches a critical value, a checkpoint is triggered and the cells will stop dividing or even undergo apoptosis.

Scientists believe that the shortening of telomeres is directly related to the onset of several diseases. Many studies have shown that human telomeres can show deletions, fusions, or shortened sequences when gene mutations and tumors are formed. Telomerase, a critical enzyme for maintaining telomere length, is found to be upregulated in approximate 85% of human cancers. This upregulation is achieved through diverse mechanisms, such as TERT (telomerase reverse transcriptase) amplifications, rearrangements, or mutations occurring in the TERT promoter region.(Horn et al., 2013; Huang et al., 2013)

Meanwhile, in some other tumors, there is an alternative lengthening of telomeres (ALT) pathway, which uses the DNA recombination of telomeric sequences. This mechanism is linked to with loss-of-function variants in the chromatin remodeling genes ATRX (-thalassaemia/mental retardation syndrome X-linked) and DAXX (death-domain associated protein). In ALT cells, telomeres frequently contain a range of telomere variant repeats (TVRs) and are thus in heterogeneous lengths. (Sieverling et al., 2020)

Although both telomere maintenance mechanisms resulted in the equivalent fact that tumor cells have unlimited replicative potential, telomerase and ALT have different impacts to the clinical outcome of cancer patients. (Recagni et al., 2020) For example, the overall survival (OS) of patients is related to telomere maintenance mechanisms (TMMs) in osteosarcoma, breast cancers, CNS, and soft tissue tumors. (Gaspar et al., 2018)

It is not far to seek that the different TMMs require alternative treatments. Telomerase inhibitor molecules have been widely used. In addition to the ongoing discovery of new molecular targets in modern pharmacy industry, drugs for ALT-targeted therapy will hopefully be developed in the near future. (Guterres and Villanueva, 2020)

Therefore, predicting the TMM of cancer patients is beneficial. Precise identification of tumors with active TMM is a prerequisite for the selection of treatment, such as the use of telomerase inhibitors, anti-telomerase immunotherapies, and anti-telomerase viral therapies. (Buseman et al., 2012) ALT has several different hallmarks, such as ALT-associated promyelocytic leukemia (PML) bodies, heterogeneous telomere length, and abundant extrachromosomal telomere repeat.

Nevertheless, the technology limitation of short-read whole-genome sequencing workflow cannot detect these sequences. However, if we use ATRX/DAXXtrunc as indicators of ALT and smartly select a few genomic variations as features, a robust machine learning model based on WGS data might be an ideal approach to solve this problem. (Conomos et al., 2012; Heaphy et al., 2011; Lee et al., 2014; Varley et al., 2002)

1.3.7 Focal CNVs and Drivers

Copy number variations (CNVs), which can result in amplification, loss of heterozygosity, or complete loss of gene functions are known to have a significant impact on the development of various cancer types. The CNVs can upregulate gain-of-expression in oncogenes and launch a deficiency or dysfunction in tumor suppressors. (Zhang et al., 2016) Meanwhile, CNVs can lead to other complicated genomic abnormalities such as gene fusions or regulation elements abbreviation.

CNVs can be classified into two major types according to their lengths. The broad CNVs are long and extend to a large fraction of a chromosome arm. On the other hand, focal CNVs are comparatively short and focus on a small region. It is also believed that these two modes arise from different mechanisms. Broad CNVs are caused by the incorrect segregation of chromosomes during mitosis and focal CNVs occur more likely from DNA repair errors. (Van Gent et al., 2001; Zhang et al., 2016)

The concept of focality has gained significance as a criterion for distinguishing true tumor-driving alterations from functionally insignificant changes in cancer-related CNV analysis. Differing from broad CNVs, focal CNVs are more likely to be a consequence of selective pressure during cancer development. (Guichard et al., 2012) The recent publications used the threshold of focal CNVs as copy number changes which are below 1 or 3 Mb in length. (Bierkens et al., 2013; Bignell et al., 2010) These focal events have been observed in various cancers, including lung, colon, and breast cancer, where recurrent focal copy number variations (CNVs) below this thresholds are associated with well-known cancer driver genes such as PTEN, CDKN2A, and RB1.(Bierkens et al., 2013; Garnis et al., 2006; Leary et al., 2008; Zhang et al., 2016)

Similar to identifying potential cancer drivers by point mutations, it is still a great challenge for driver mining in CNVs. For example, according to Beroukhim et al., a total of 150 focal regions have been identified as potential hotspots for cancer driver genes, from which a broad analysis of more than 3,000 patients but only one fourth of the regions are linked to oncogenes or TSGs which are already known. (Beroukhim et al., 2010) Many theories have tried to explain this phenomenon. The distribution of focal CNVs is quite imbalanced in regions of chromosomes. These events may not act as traditional TSG or oncogenes but they affect tumor promotion with a collective of minor functions from different places in the genome. (Solimini et al., 2012)

1.4 Overview of the Study

Throughout the research of this thesis, which is based on the massive cancer cohort data and high-quality analysis pipeline from the PCAWG project, I conducted a series of in-depth analyses of correlations between genomic mutations in cancer. My idea was to start with the identification of mutations in cancer driver genes and then explore the origin of cancers step by step.

First of all, I developed highly efficient data storage, query and visualization solutions. After that, I used mutation recurrency enrichment algorithms to deeply analyze potential cancer driver mutations in both gene coding regions and non-coding DNA. After that I established the overall landscape for cancer driver mutations, I further studied the function of biallelic inactivation in cancer development. I applied our knowledge of variations in non-coding regions to the studies of promoter mutations and lncRNAs. Besides analyzing single-gene mutations, I also turned our attention to the interactions between multiple genes, using mutually exclusive analysis to the study of synthetic lethality among cancer gene mutations, providing support for screening new generations of cancer drug targets. In another chapter of the thesis, I looked into cancer-related focal CNVs. To better understand and visualize them, I developed a new bioinformatics tool called GenomeTornadoPlot. In addition, I also studied the PCAWG data to support the study of activated telomere maintenance mechanisms. Moreover, a random forest-based classifier has been implemented to distinguish between telomerase and ALT patients. This informative research insight will provide new explanations for the complex story of cancer development and gene mutation in several aspects.

2 Computational material and data preprocessing

2.1 Computational Materials

The operating system I used on the local computer is openSUSE Linux version 2.6.37.6-24-default and later changed into CentOS-7.8-DKFZ.

I used the DKFZ TBI-cluster with 44 nodes of AMD Opterons, 16-256 GB RAM, 16 GB swap space, and a clock frequency of 2.0-2.7 GHz in the first phase. Later, I changed it into the DKFZ LSF-cluster.

The programs of this study are mainly coded in R and Python. I coded R in the R studio on the DKFZ server and the version was 3.5.1. The Python code is under version 2.7. The used R and python packages will be introduced in each section.

2.2 Data

The PCAWG dataset was collected and unified by ICGC. The original data set contains more than 2,900 patients from 48 cancer research projects. The dataset comprises WGS data from clinical centers around the world.

The WGS data of each patient includes at least one tumor sample and one control sample, which were sequenced by Illumina HiSeq with 100-150bp paired-end sequencing reads and an average coverage of at least 30 reads in the tumor and 25 in the control sample. The control samples were collected from blood or healthy tumor-adjacent tissues depending on cancer type. In some cases, additional samples were collected. For all patients, clinical data, including patient ID, age, sex, and survival time, are also concluded. Moreover, the dataset provides RNA sequencing data for approximately two-thirds of the patients. The data is available in ICGC data portal (https://dcc.icgc.org/).

2.3 Data Preprocessing

The WGS data was downloaded from the ICGC-PCAWG data portal (https://dcc.icgc.org/pcawg) and WGS data were preprocessed by the DKFZ. The SNV/MNV/INDEL calling and functional annotation were performed by Ivo Buchhalter. (https://github.com/DKFZ-ODCF/SNVCallingWorkflow) The CNV and SV callings were performed by members of the previous Computational Oncology group. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020)

From the DKFZ variant calling pipelines, I obtained the VCF format files as output. Aiming for fast querying and highly efficient storage of the data, I converted the VCF files into a new data structure. In this data format, I encoded detailed information on all point mutations, CNVs, and SVs as simple strings in order to compress the file size.

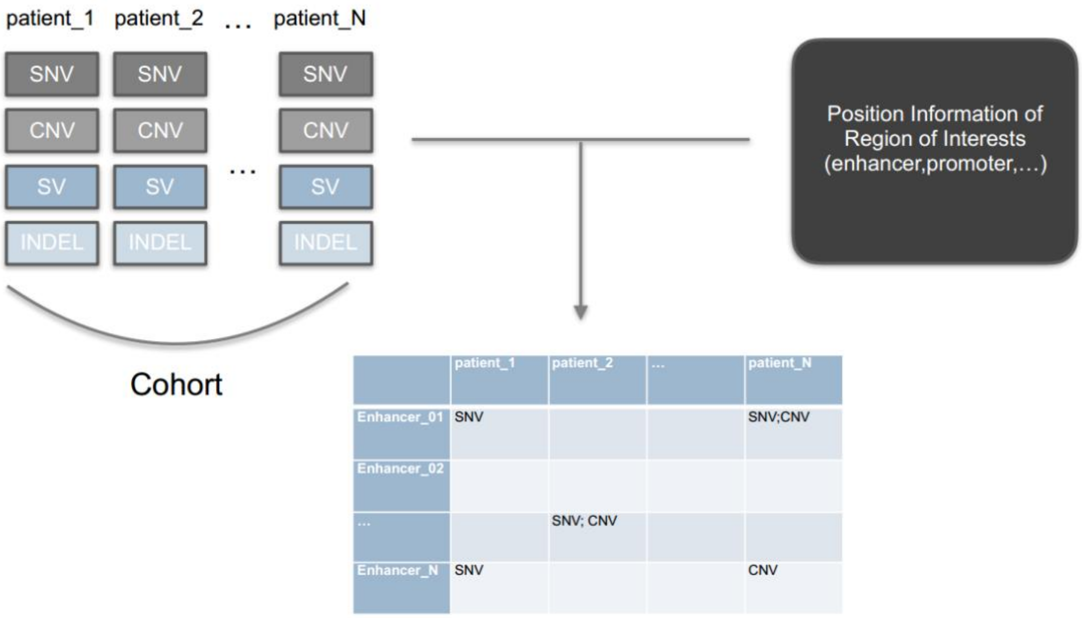
These strings are coded by the following format:

SNV_A_B_C; CNV_D_E,F

where A is the category of SNV/INDEL types:

0 - synonymous SNV, 1 - unknown INDEL, 2 - non-frameshift INDEL, 3 - frameshift INDEL, 4 - nonsynonymous SNV, 5 - stop-gain SNV; B is the chromosome index; C is the coordinate of SNV; D is the copy number; E and F are the starts and ends coordinate of CNV.

The WGS variations from all PCAWG patients are saved in one file. The data were stored in the form of a matrix. In this matrix, each row represents a gene and each column represents a patient. This gene-centric presentation makes it very easy to subtract interesting parts of data, no matter if the users wanted to analyze whole genome variants of certain cohorts or if they were interested in a specific gene's function among all cancer types. SNV/MNV/INDEL information and CNV/SV are combined in the same file so that it is convenient to analyze their interactions, such as bi-allelic inactivations. (Figure 5)


patient_1
patient_2
...
patient_N
SNV
CNV
SV
INDEL
Cohort
Position Information of Region of Interests (enhancer,promoter,...)

	patient_1	patient_2	...	patient_N
Enhancer_01	SNV			SNV;CNV
Enhancer_02				
...		SNV; CNV		
Enhancer_N	SNV			CNV

Figure 5: The data structure for storage of PCAWG mutation information.

In this matrix, each row represents a gene and each column represents a patient. The mutation information is stored in the cells with mutation types, positions, and other detailed information.(adapted from (Hong, 2016))

The preprocessing of RNA sequencing data was done in collaboration with Sandra Koser from the Applied Bioinformatics Department of DKFZ. Similar to WGS files, I preserved the Fragments Per Kilobase Million (FPKM) value of all data-available patients in a matrix format file. On the other hand, considering that the PCAWG dataset was collected from different projects all around the world, the batch effect of RNA sequencing cannot be ignored. Because of this, Sandra Koser and I normalized the FPKM values for each gene within every single cohort and we saved the z-scores as the adjusted expression level of genes. In a few cases, there are more than one RNA sequencing batch for one single patient. In this situation, the average of the expression value of this patient is calculated. The RNA sequencing data is saved in both python pickle files which can access with query in python script and matrix based data structure.

2.4 Code Availability

Part of my codes including python, R, shell scripts and documentations for the research project can be found in https://github.com/chenhong-dkfz.

3 Driver Mutations

Research on cancer driver genes is an important part of the PCAWG project. For this part, I explored the driver genes in the pan-cancer NGS data in terms of point mutation frequency, biallelic inactivation, chromosomal copy number variation, promoter mutations, etc. with multiple statistical methods, and tried to discover new driver gene candidates and explored their potential in cancer therapy.

3.1 SNV/CNV recurrence analysis of multiple cancer types

3.1.1 Overviews of SNVs in PCAWG data

I statistically analyzed the number of contributions of SNVs in each patient in different cancer types. (Figure 6-7) As it is shown in figure 6, I noticed that the number of SNVs per cohort differs significantly among different cohorts.

SKCM-US (Skin Cutaneous melanoma), LUSC-US (Lung Squamous Cell Carcinoma) and ESAD-UK (Esophageal Adenocarcinoma) have extremely high frequencies of SNVs in patients (mean > 30000 SNVs/patient). This might be the result of artifacts from SNV calling. Breast cancer (BRCA-US, BRCA-UK), Acute Myeloid Leukemia (LAML-US, LAML-UK) and ovarian cancer (OV-US, OV-AU) share similar SNV frequencies respectively. This similarity clearly indicates that cohorts of the same cancer types which were independently sequenced on different continents show low technical bias.

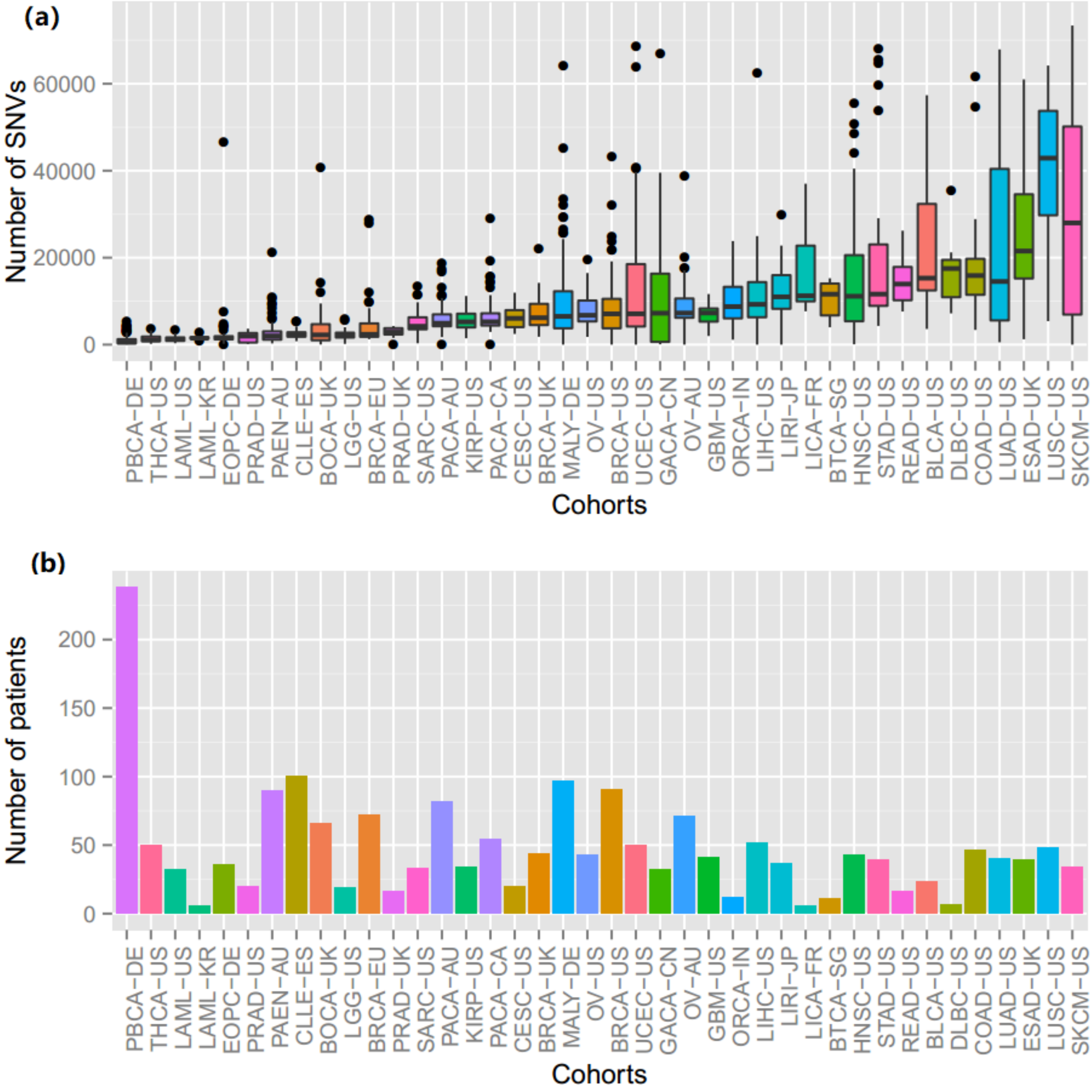


Figure 6: Distribution of SNV numbers per patient.

(a) Boxplot of total numbers of SNVs in WGS by PCAWG cohorts. Cohorts are ordered by the median number of SNVs per sample. (b) Barplot of total patient numbers of each PCAWG cohort. Cohorts are ordered as the same as that in (a). (Adapted from (Hong, 2016)*)*

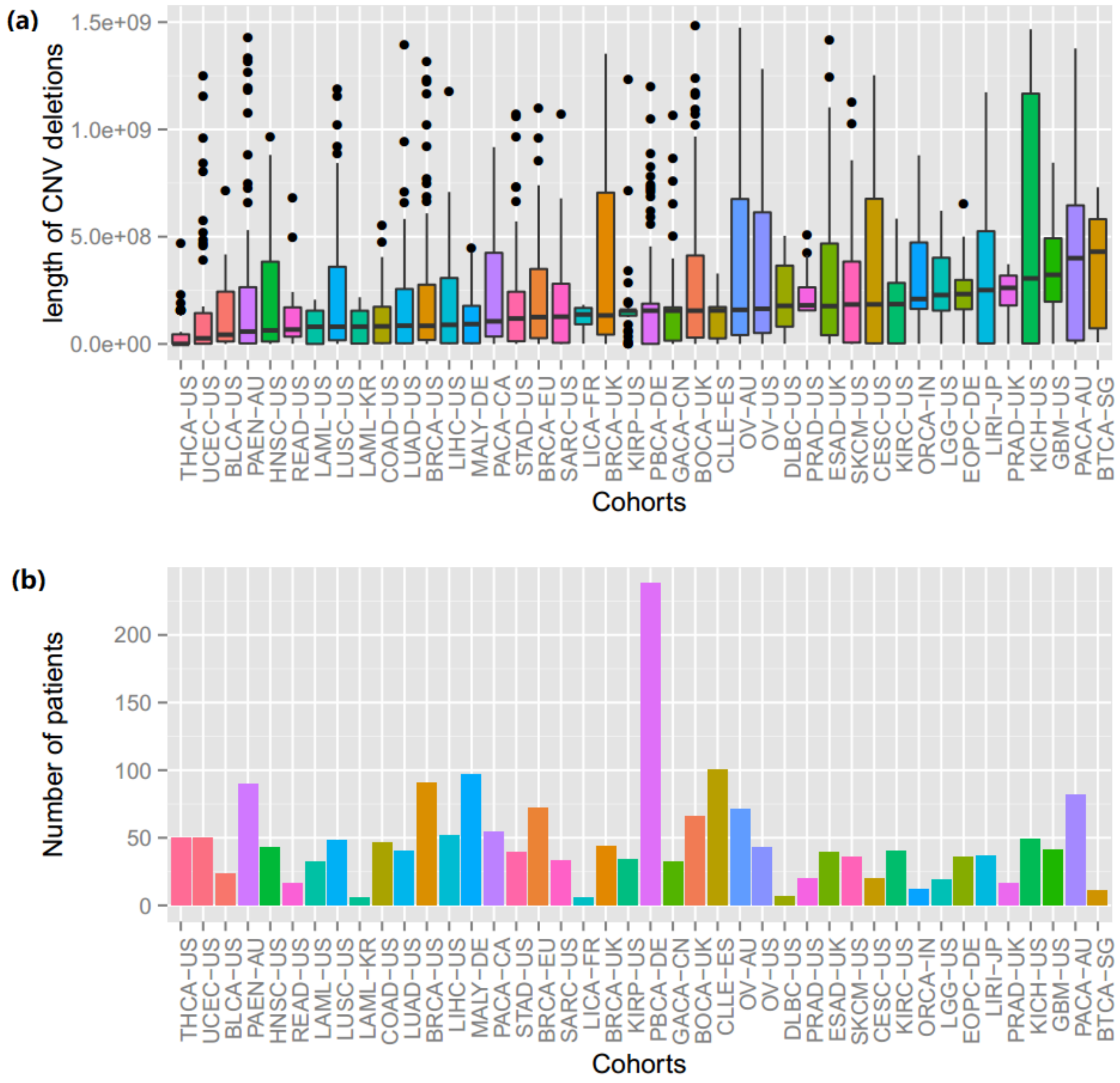


Figure 7: Distribution of CNV lengths per patient.

(a) Boxplot of total lengths in bp of CNV deletions in WGS by PCAWG cohorts. Cohorts are ordered by the median total lengths of CNV deletions per sample. (b) Barplot of total patient numbers of each PCAWG cohort. Cohorts are ordered as the same as that in (a). (Adapted from (Hong, 2016)*)*

3.2 Biallelic Inactivation Analysis and Applications

The inactivations of tumor suppressor genes can be carcinogenic. In diploid cells, the two alleles of genes denote expression levels. Mutations on either allele are possible to inactivate the gene and decrease expression. Compared to monoallelic inactivations, biallelic inactivations have stronger effects, as they are guaranteed to lead to a loss of genetic information.

It is not trivial to filter out biallelic mutations from VCF files, especially in cohort-based analysis. A Variant Call Format (VCF) file is a file format used in bioinformatics for storing genetic sequence variants. It records information including the technical information of the sequencing experiment as well as the position, reference, alteration, quality, and other information of each mutation. In this study, I focused on a special case of biallelic inactivations where structural variation, deletion (or copy number variation loss) and an SNV occurring in one gene region at the same time. In principle, the lengths of gene element regions are shorter than normal sizes of structural variations or copy number variations. So, in the case that both SV deletion (or CNV loss) and an SNV occur simultaneously in the region, it is unlikely that both of them are in one allele. Firstly, I define the “disruptive” SNVs as point mutations that actually inactivate the function of gene elements and they are defined slightly differently in different types of gene elements. Then I defined that if there exists at least one “disruptive” SNV and one deletion overlapping together in a gene element in one sample, this sample is a “potential biallelic inactivation” sample for this gene element. A homozygous CNV deletion is also considered as a potential biallelic inactivation because the function of genes is removed both alleles. In coding region sequences (CDS), nonsynonymous, splicing, and stop gain SNVs are involved. In 3’ UTR and 5’UTR regions, UTR SNVs are also considered as effective. While in core promoters and enhancers, all SNVs belong to “effective SNVs”.

If there exist copy number gains overlapping the region of a gene element, this sample is defined as a “potential gene amplification” sample for this gene element.

According to the copy number count of the region, I divided the potential gene amplification samples into three groups: genome duplication (CNV gain 3-4), double genome duplication (CNV gain 5-8) and "octoploid and events that are high-level amplifications" (CNV gain over 8). In general, the over-expression of genes is often linked to high copy numbers. The abnormally high expression of these genes is very interesting in cancer research.

Analysis of potential biallelic inactivations and gene amplifications provides insight between gene mutations and expressions and it is very helpful for predicting new cancer drivers.

3.2.1 Overview of Biallelic Inactivations in PCAWG Data

3.2.2.1 Overview

To discriminate pan-cancer drivers from mutations that are entity-specific, I counted the total number of potential biallelic inactivations in all patients and calculated the entropy of inactivation distributions among cohorts.

The Shannon entropy is calculated by:

$$H = -\sum_{i=1}^{m} \quad p_i log_2 p_i$$

where p_i is the percentage of patients of cohort i in total patients.

For each gene, the entropy is calculated. If the Shannon entropy is high, the corresponding genes are likely to involve CNVs distributed in different cohorts.

From Table 1, there are some known TSGs such as TP53, CDKN2A, and PTEN in the list. And some of their neighbor genes appear on this list as well because they are affected by the homozygous deletion. Therefore, it is necessary to rule out the effect of broad CNV deletion, thereby eliminating false-positive candidates. Meanwhile, I also looked into the top candidates from each chromosome. There are several interesting genes which have more bi-allelic inactivations than any other genes in the same chromosome, such as CSMD1, WWOX, CCSER1, and MACROD2.

I calculated the entropy of histology distribution of bi-allelic inactivations of CGC genes in the PCAWG data (Figure 8). CDKN2A and TP53 have the most events and they have high entropies, which suggests that they are more likely to distributed dispersedly in different histologist. While the genes like SMAD4, VHL, DCC, and PBRM1 have the trends of focus in a few specific histological types.

Table 1: Top 30 bi-allelic inactivated gene.

Gene name	Chr	Gene start	Gene end	Entropy	Sum	Role
RP11-145E5.5	9	21802635	22032985	2.20914072244596	307	NA
CDKN2A	9	21967751	21995300	2.21260178249442	304	TSG
C9orf53	9	21967137	21967738	2.23744325098634	265	NA
CDKN2B-AS1	9	21994777	22121096	2.21702864289586	262	NA
RP11-149I2.4	9	21995481	21996012	2.21662955697574	260	NA
CDKN2B	9	22002902	22009362	2.20632524516232	256	NA
UBA52P6	9	22012154	22012535	2.18959444788354	248	NA
MTAP	9	21802542	21931646	2.1939739073248	243	NA
RP11-149I2.5	9	21929456	21931072	2.19235275972528	240	NA
TP53	17	7565097	7590856	2.29055453558936	239	oncogene, TSG, fusion
RP11-70L8.4	9	21858909	21861925	2.15601998687196	207	NA
TUBB8P1	9	21811620	21812346	2.08179371332763	181	NA
RP11-408N14.1	9	22203989	22214671	2.16082816658821	168	NA
KHSRPP1	9	21695175	21696942	2.16285833634615	142	NA
DMRTA1	9	22446840	22455739	2.14295972979971	132	NA
FHIT	3	59735036	61237133	1.70019938189912	132	TSG, fusion
RP11-344A7.1	9	21638284	21638675	2.17837845334424	130	NA
MIR31HG	9	21455641	21559668	2.25844393703536	124	NA
RP11-399D6.2	9	22646199	22824212	2.15036895763009	113	NA
RP11-370B11.1	9	22747699	22748233	2.13703134241512	107	NA
SMAD4	18	48494410	48611415	1.26108178914847	107	TSG
IFNE	9	21480841	21482312	2.15298169834888	106	NA
RP11-370B11.3	9	22767174	22768315	2.13798235585408	105	NA
PTEN	10	89622870	89731687	2.37614595831218	102	TSG
IFNWP19	9	21455483	21456048	2.17068334238048	101	NA

IFNA8	9	21409146	21410184	2.17258343931578	100	NA
IFNA1	9	21440440	21441315	2.17272951810876	99	NA
IFNA2	9	21384254	21385396	2.19129070055303	99	NA
IFNWP2	9	21420233	21420812	2.17272951810876	99	NA
IFNA11P	9	21398613	21399138	2.1841948403504	98	NA

Notes: Each column represent gene names, chromosome, gene start position, gene end position, bi-allelic inactivation entropies, total numbers of bi-allelic inactivations and the roles of gene in CGC list.

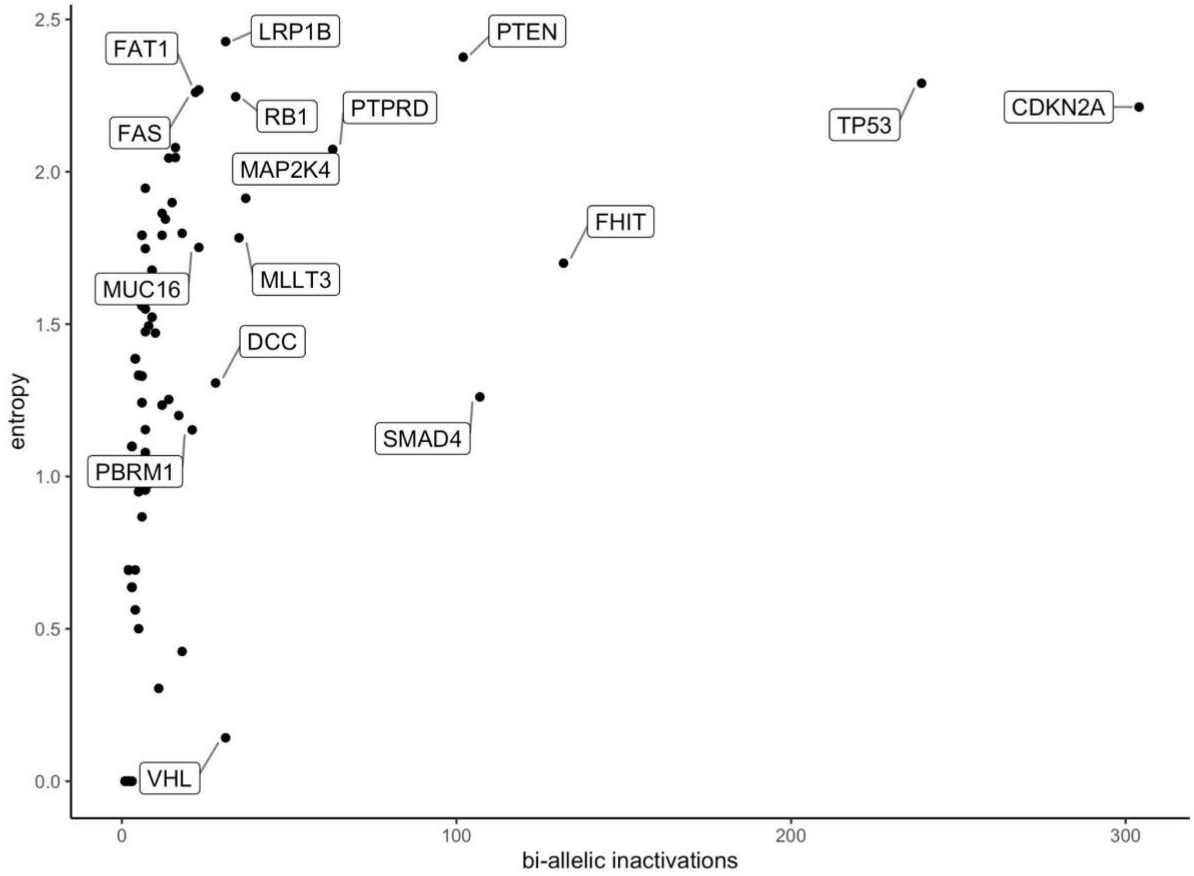


Figure 8: Bi-allelic inactivation and entropy of CGC genes.

The x-axis represents the total number of bi-allelic inactivations in all PCAWG patients. The y-axis represents the entropy of histology distribution of these bi-allelic inactivations.

3.2.1.2 Discussion

Similar to any other mutation-based study, the influence of lengths cannot be neglected from the bi-allelic inactivation analysis because the length the gene is the more likely the gene get mutated, even by 2 different variations. More studies are needed to establish a convincing correlation between these genes with high bi-allelic inactivations and cancers.

3.2.2 Biallelic Inactivation and Cancer Driver Long, Non-coding RNAs

3.2.2.1 Motivation

The Cancer LncRNA Census (CLC) genes are defined by Joana Carlevaro-Fita et al. as a set of LncRNAs that are directly impacting cancer progress as shown by experiment or genetic proof. In the CLC gene list, there are 122 LncRNA genes, including 77 oncogenes, 36 tumor suppressor genes and 9 genes with both functions. (Carlevaro-Fita et al., 2020)

In this study, I wanted to determine the biallelic inactivation and copy number amplification features of these LncRNAs.

3.2.2.2 Biallelic Inactivation on CLC genes

The CLC genes were preliminarily predicted and classified as T (tumor suppressor) genes, O(onco-) genes, and pt (complex functions) genes. 100 out of 117 CLC genes are long, noncoding RNAs. I calculated the potential biallelic activations, potential monoallelic activations and gene amplifications for all lncRNAs in gencode v19 (13848 genes) and compared the CLC genes to the remaining lncRNAs.

In this study, I observed that both CLC tumor suppressor genes and CLC oncogenes have more potential biallelic activations than background (unadjusted p-value: 0.000136 for O genes against background, 0.000749 for T genes, one-sided Wilcoxon rank-sum test). For potential monoallelic inactivations, only T genes have a significant difference (unadjusted p-value: 0.0176, one-sided Wilcoxon rank-sum test) and are higher than the background. (Figure 9)

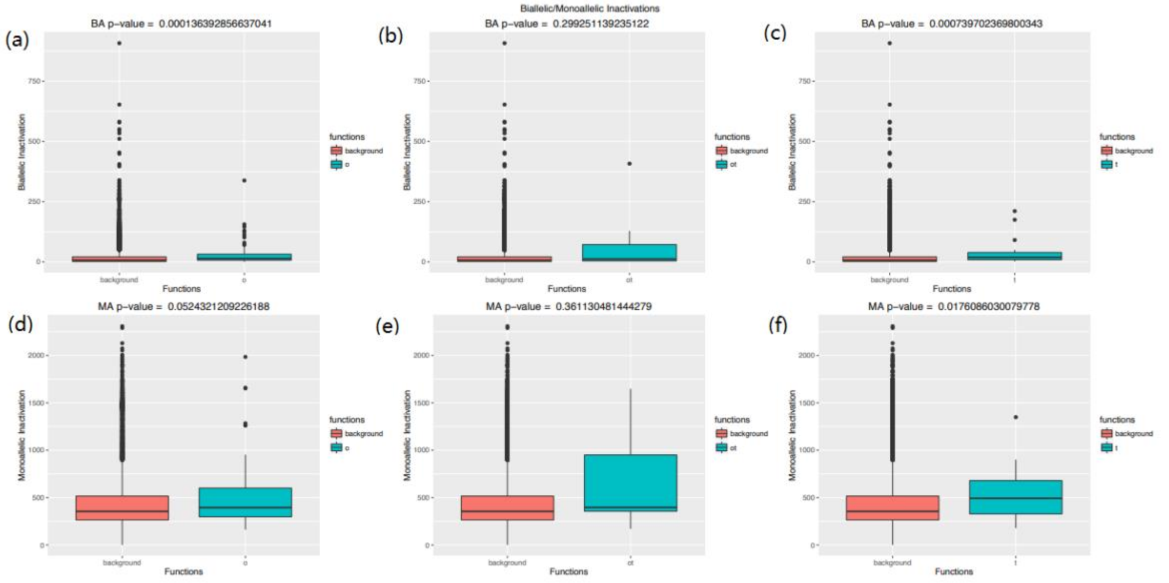


Figure 9: Comparison of counts of biallelic/monoallelic inactivations between predicted CLC genes and background.

Y-axis represents the count of bi/mono-allelic inactivations of genes in all patients. (a) bi-allelic inactivations between background (red) and oncogenes (cyan). (b) bi-allelic inactivations between background (red) and complex functional genes (TSG+oncogenes, cyan). (c) bi-allelic inactivations between background (red) and TSGs (cyan). (d) mono-allelic inactivations between background (red) and oncogenes (cyan). (e) mono-allelic inactivations between background (red) and complex functional genes (TSG+oncogenes, cyan). (f) mono-allelic inactivations between background (red) and TSGs (cyan).

Between CLC genes and background, I didn’t observe significant differences of gene amplifications with lower than 8 copies. For hyper gene amplification with more than 8 copies, however, the O genes have more amplification cases than the background group (unadjusted p-value: 0.0371, one-sided Wilcoxon rank-sum test) (figure 10). The result validated the prediction of oncogenes because these genes are very possible to unregulated in cancer samples. (Figure 10)

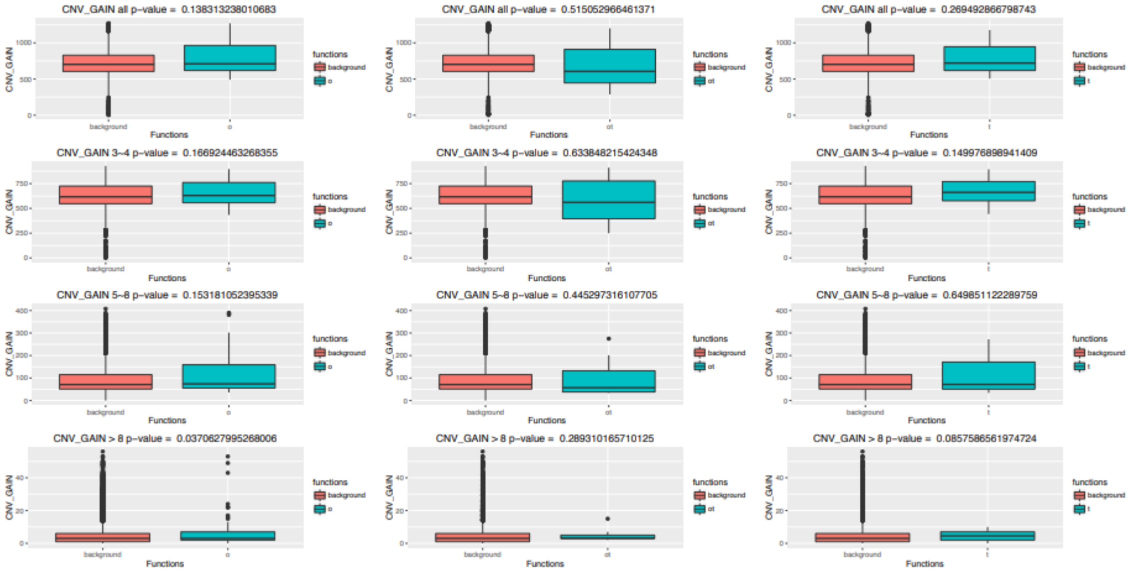


Figure 10: Comparison of counts of different levels of gene amplifications between predicted CLC genes and background.

Y-axis represents the count of CNV amplification of genes in all patients. The p-values of one-sided Wilcoxon rank-sum test are not significant between any cancer-related groups and controls except oncogenes vs control group with CNV gain above 8 (bottom left).

3.2.2.3 Most frequently bi-allelic inactivated and amplificated genes

I listed the top candidates in both CLC and background groups. The predicted CLC tumor suppressor genes with frequent potential biallelic inactivation and oncogenes with frequent hyper gene amplifications are more convincing cancer related lncRNA candidates.

Table 2: Top 10 LncRNAs with most frequent bi-allelic inactivations.

Ensembl gene ID	**gene name**	**role**
ENSG00000253535	RP11-624C23.1	NA
ENSG00000271860	RP11-436D23.1	NA
ENSG00000237647	ERICH1-AS1	NA
ENSG00000253642	RP11-436D12.1	NA
ENSG00000240498	CDKN2B-AS1	oncogenes in CLC
ENSG00000231535	LINC00278	NA
ENSG00000251574	RP11-6N13.1	NA
ENSG00000180910	TTTY11	NA
ENSG00000229308	AC010084.1	NA
ENSG00000265533	RP11-638L3.1	NA

Notes: Each column represents ensemble gene ID, gene name and predicted roles in CLC list.

For biallelic inactivations analysis, I found that many genes in the X chromosome have higher biallelic mutations due to the phenomenon of X chromosome hypermutations. Interestingly, it has been documented that the non-functioning X chromosome in many cancer genomes of female patients shows a significant increase in somatic mutations rates. (Jäger et al., 2013) Moreover, due to the specific nature of the X chromosome (males have only one X chromosome while in females one of the X chromosomes is randomly inactivated), I excluded X chromosome genes from the top candidates to remove the false positives. In the top 10 genes, I found one oncogene (CDKN2B-AS1) in the CLC list.

Table 3: Top 10 LncRNAs with most frequent amplifications (CN>5).

Ensembl gene ID	gene name	role
ENSG00000249375	CASC11	cancer susceptibility candidate
ENSG00000249859	PVT1	oncogenes in CLC
ENSG00000246228	CASC8	cancer susceptibility candidate
ENSG00000247844	CCAT1	oncogenes in CLC
ENSG00000253929	CASC21	cancer susceptibility candidate
ENSG00000254166	CASC19	cancer susceptibility candidate
ENSG00000253264	PCAT2	prostate cancer-associated transcript 2
ENSG00000254275	LINC00824	NA
ENSG00000253438	PCAT1	oncogenes in CLC
ENSG00000254286	RP11-89K10.1	NA

Notes: Each column represents ensemble gene ID, gene name and predicted roles in CLC list.

For gene amplification, I found three CLC oncogenes in the top 10 list as expected. This result would support the prediction of CLC genes. The rest of the candidates are also interesting, including CASC11, CASC8, CASC8, CASC19, and PCAT2, which are already known as cancer susceptibility candidates genes and prostate cancer-associated transcripts.

3.2.2.4 Discussion

Long non-coding RNAs play some roles in cancer development. However, the functions of most LncRNAs are still unclear. It is not even clear if some LncRNAs are primary “driver” genes whose mutations lead to cell tumorigenesis or if they are just playing a downstream role which accelerates tumor development and progresses with changing the expression levels (secondary driver). Therefore, I computed statistics of the recurrence of bi-allelic inactivations or hyper amplifications in these gene regions. Due to the uncertainty of function and alternative splicing, it is still hard to address the functional changes of SNVs or CNVs in the LncRNAs. To this end, a thoroughly conducted analysis of the secondary structure of the mutated lncRNAs would be required. My results can help to prioritize new targets for this type of analysis.

Comparing the CLC gene set and background, I found that tumor suppressor CLC genes have a significantly higher bi-allelic inactivation frequency (p-value=0.0007) than the background group while the mono-allelic inactivation frequency is only slightly significant different (p-value=0.0176). It may mean that mono-allelic inactivation could not change the functions of tumor suppressor lncRNAs, and thus, do not provide a selective advantage for cancer cells. Interestingly, I found that the CLC oncogenes also have significantly more bi-allelic inactivation than the background. This observation is counter-intuitive and requires further investigation. A detailed study of each entry is necessary in the future.

In the bi-allelic activation gene group, RP11-624C23.1 is involved in DNA damage response and is downregulated in childhood acute lymphoblastic leukemia. (Gioia et al., 2017) ERICH1-AS1 is an negative prognostic factor for gastric cancer and it is already used as a biomarker for predicting non-small-cell lung cancer. (Chen et al., 2020; Tang et al., 2015) LINC00278 is involved in esophageal squamous cell carcinoma and influence the androgen receptor signaling pathway. (Wu et al., 2020)

In the amplification group, I found several cancer susceptibility candidates in the list. These genes are not included in the CLC are awaiting more experimentally proof. CASC11, which is involved in WNT pathway, is linked to Glioma Susceptibility 1, Hepatocellular Carcinoma, colorectal cancer, and ovarian and rectum squamous cell carcinoma. (Luo et al., 2017) CASC8 is reported to be upregulated in colorectal cancer and breast carcinoma. (Wang et al., 2020; Yao et al., 2015) CASC21 is linked with chronic lymphocytic leukemia and prostate cancer. (Calin and Croce, 2009; Kim and Croce, 2018) CASC19 and PCAT2 are prostate cancer-related LncRNAs. (Bawa et al., 2018; Ramnarine et al., 2019; Wang et al., 2019)

In summary, for the top candidates in the bi-allelic inactivation list and amplification list, further research is required. Some genes in the list are also interesting and have already shown some functional correlation to cancers.

3.2.3 Synthetic lethality of cancer driver genes

3.2.3.1 Motivation

In the past decades of development, people have gradually realized the significance of synthetic lethality analysis for finding new cancer drugs. In this study, I mainly conducted an analysis on the PTEN gene, trying to find drug target genes that potentially can be used to cure PTEN-deficient cancer cells.

Phosphatase and tensin homolog (PTEN) is a protein encoded by the PTEN gene in the human genome. Mutations in this gene often link to the progression of many cancers. In general, the PTEN plays a central role in RET signaling, PI3K/AKT activation, and p53 signaling pathways. In the sense of biological functions, PTEN regulates cell proliferation, cell survival, cell migration, genome stability, and stem cell self-renewals. In many types of cancers, such as prostate cancer, breast cancer and liver cancer, the functions of PTEN are often inactivated and thus the cells transform into cancer cells.

Although PTEN is a very important gene, it is still not druggable. However, now the synthetic lethality analysis provides us with the probability to find PTEN’s SL partner. Once silencing the SL partner genes, the PTEN-deficient cell will die.

3.2.3.2 Pipeline design

In different cancer types, PTEN gene mutation rates differ. To avoid statistical bias, it is necessary to select the cancer cohorts which are enriched in PTEN-deficient patients.

Within the PCAWG data, there are originally 48 cohorts. I first calculated the PTEN loss-of-function mutation rate of each cohort. In this step, I defined three types of loss-of-function mutation levels and two types of amplification levels.

Loss-of-function mutation means there are functional point mutations or copy number deletions in this gene region:

- Monoallelic inactivations: a functional point mutation or a heterozygous copy number deletion occur in the same position in the gene region.
- Potential biallelic inactivations: (1). a point mutation and a heterozygous copy number deletion occur in different positions without overlaps in the gene region, and (2) two functional point mutations occur in the gene region. In these situations, it is not necessarily possible to identify if the mutations are in the same allele, so I defined them as ‘potential’ biallelic inactivations.
- Biallelic inactivations: (1) a functional point mutation and a heterozygous copy number deletion occur in the same position in the gene region, (2) a homozygous copy number deletion in the gene region.

Amplification means the copy number gain of the gene region.

- Copy number between 5 and 8 (5 and 8 inclusive)
- Copy number > 8

Copy number gain less than 5 was not considered as amplification here to make sure of the confidence of events because it may not have any effect to expression change.

For PTEN, a significant tumor suppressor, I was interested in the deletions and SNVs. Because of this, I calculated the mutation rates of each loss-of-function mutation type in all cohorts. (Figure 11)

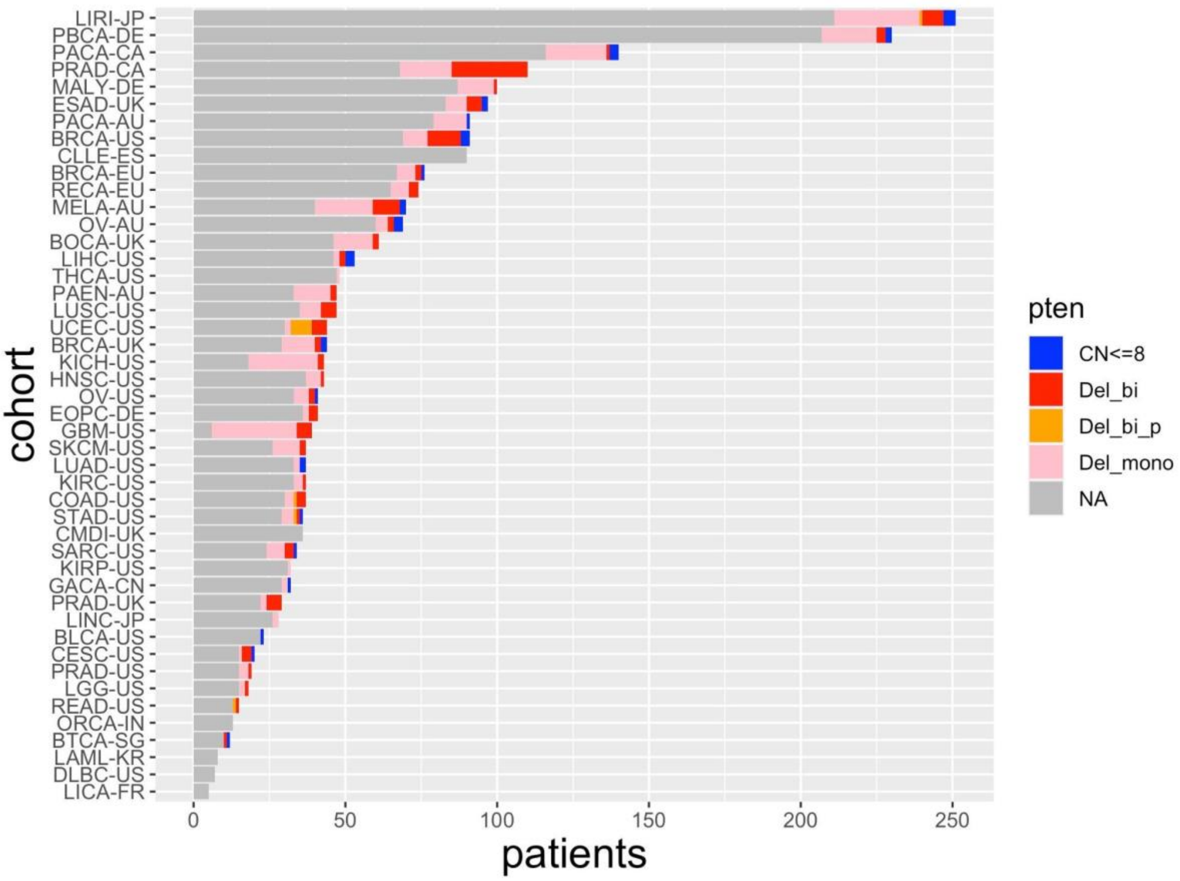
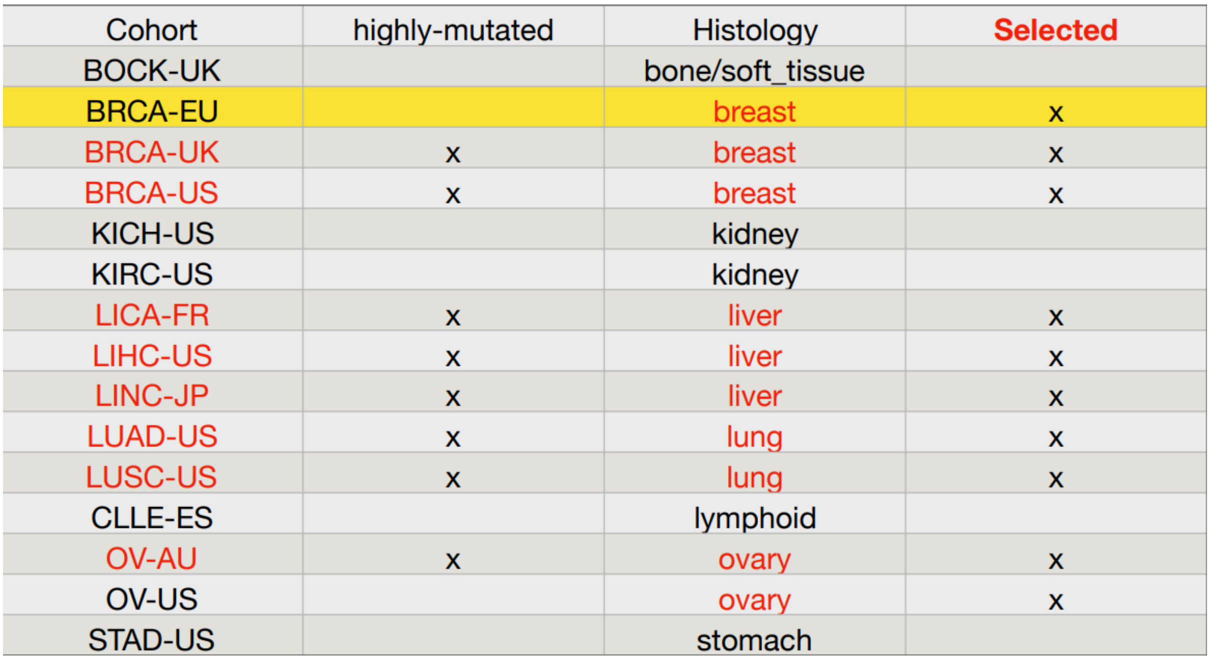


Figure 11: Overview of different types of mutations of PTEN in cohorts.

Blue represents for amplifications between copy numbers 4 to 8, pink represents monoallelic inactivations, orange stands for potential biallelic inactivations, and red stands for real biallelic inactivations.

Any cohorts in which at least 15% of patients have loss-of-function mutations of PTEN were selected as highly frequently mutated cohorts. To ensure that a sufficient number of patients are involved into the analysis, I also selected other cohorts which are histologically similar to the highly frequently mutated cohorts. Next, I performed analysis for all patients in the selected cohorts. For example, the proportion of mutated samples in the BRCA-EU project does not reach the threshold but this was still selected because other breast cancer cohorts met the requirement.

Table 4: an example of meta-cohort selection.



Cohort	highly-mutated	Histology	**Selected**
BOCK-UK		bone/soft_tissue	
BRCA-EU		breast	x
BRCA-UK	x	breast	x
BRCA-US	x	breast	x
KICH-US		kidney	
KIRC-US		kidney	
LICA-FR	x	liver	x
LIHC-US	x	liver	x
LINC-JP	x	liver	x
LUAD-US	x	lung	x
LUSC-US	x	lung	x
CLLE-ES		lymphoid	
OV-AU	x	ovary	x
OV-US		ovary	x
STAD-US		stomach	

Finally, I selected 13 histological meta-cohorts with 868 samples. The meta-cohorts are listed in table 5.

Table 5: the selected meta-cohorts for synthetic lethality analysis project.

Bone/soft tissue	Ovary
Breast	Pancreas
Cervix	Prostate
CNS	Skin
Colon/ Rectum	Stomach
Kidney	Uterus
Lung	

3.2.3.3 Mutual Exclusive Analysis

I apply mutual exclusivity analysis for PTEN with 3 loss-of-function alteration combinations against all mutations of all other genes.

They include three basic types of mutations: bi-allelic inactivations (bi), potential bi-allelic inactivations (po_bi) and mono-allelic inactivations (mi). According to the credibility of the mutation level, the PTEN mutations can be divided into three combinations: bi, bi+po_bi and bi+po_bi+mi. Each level is a subset of next level. The most convincing level is only the real bi-allelic inactivations. (Table 6)

Table 6: Schematic of mutually exclusive matrix.

	Gene X bi+po_bi+mi (mut)	Gene X bi+po_bi (pb)	Gene X bi (rb)	Gene X amp>4 (nam)	Gene X amp>8 (pam)
PTEN bi+po_bi+mi (mut)	x	x	x	x	x
PTEN bi+po_bi (pb)	x	x	x	x	x
PTEN bi (rb)	x	x	x	x	x

Notes: *bi = bi-allelic inactivations, po_bi = potential bi, mi = mono-allelic inactivations, amp = amplifications.*

For each combination of any gene and PTEN, I calculate the p-values of the hypergeometric distribution, which identifies mutually exclusive situations in the cohorts. I get 3x5=15 p-values for each gene pair. I didn’t apply multiple testing because the purpose of this step is to rank the genes and find a few top candidates.

The hypergeometric distribution is used for sampling without replacement. The density of this distribution with parameters m,n,k is given by:

$$p(x) = \binom{m}{x}\binom{n}{k-x}\Big/\binom{m+n}{k}$$

for x=0,...,k, where m is the number of gene A deficient patients in the cohort, n is the number of gene B deficient patients in the cohort and k is the number of patients with both A and B deficiency.

I applied the phyper function from the basic stats package in R to our data and calculated the p-values for all possible gene pairs.

3.2.3.4 Filters

After calculating the p-values, we designed four filters to screen out target gene candidates.

- All protein coding genes in pan-cancer list, events >= 5 & p-values >0.77 in bi and potential bi list.
- Genes on X only if mutated in prostate
- No large overlap in one indication (all OR < 2)
- >=6 events in selected indications of interest

There are 54 genes that meet all four requirements. (Supplemented Table 5) As a follow-up, these candidates were studied by analyzing pathway information, their mutational patterns in the cBIO database and the scientific literature.

3.2.3.5 Pathway Information, cBIO, and Literature.

The following databases were used for the functional annotation of all candidate genes and done by my colleagues Tobias Bauer and Daniel Hübschmann in DKFZ.

1. Gene function and annotation – GeneCards(https://www.genecards.org/)

2. Detailed gene function – PubMed

3. Literature link to cancer – PubMed

4. Confirmation of ME in other/larger cohorts of relevant indication – CBioPortal(https://www.cbioportal.org/)

5. Assessment of co-deletion events with known TSs – UCSC Genome Browser and CBioPortal(https://genome.ucsc.edu/; https://www.cbioportal.org/)

6. Availability of tool molecules/competition and available X-ray structures– ClinTrials and PDB(https://www.clinicaltrials.gov/;https://www.rcsb.org/)

3.2.3.6 Result

Finally, we kept 10 genes as the final candidates. (Supplemented table 5)

MED12 (mediator complex subunit 12), located on chromosome X, is amplified in neuroendocrine prostate cancer and castration-resistant prostate cancer. It may be oncogenically L1224F mutated in prostate cancer. This gene participates in the structure of the mediator which activates the CDK8 kinase and plays an important role in transcription regulation. (Zhang et al., 2020)

BAP1 (BRCA1 Associated Protein 1) binds to the RING finger domain of BRCA1, which is encoded by the BAP1 gene in chromosome 3. (Jensen et al., 1998) It locates in the same chromosome arms with VHL and they are often co-deleted. In this study, BAP1 gene is mutated in 9 (11%) kidney cancer patients and 10 samples in other cancer types.

CBFB (Core-Binding Factor Subunit Beta), located in chromosome 16, forms a heterodimeric complex core-binding factor (CBF) with RUNX family proteins. CBFB genes are significantly mutated in cancers such as breast cancer and intestinal cancer. The CBFB gene is a highly mutated driver in many human cancers, including breast cancer. (Malik et al., 2019; Speck, 2001)

GPR98, located in chromosome 5, encodes the G protein-coupled receptor 98, whose aberrant expression and activity of G proteins and G protein-coupled receptors (GPCRs). These proteins are frequently associated with tumorigenesis. Nearly 20% of human tumors have mutations in GPCRs. GPR98 is one of the most frequently mutated GPCRs in cancer. (O’Hayre et al., 2013; Sriram et al., 2019)

WNK3 is a gene located on the X chromosome that encodes a protein belonging to the 'with no lysine' (WNK) family of serine-threonine protein kinases. Proteins within this family are characterized by the absence of the catalytic lysine in subdomain II and instead possess a conserved lysine in subdomain I. Although there are not many studies about the direct crosslinks between WNK and cancers, WNK proteins, along with their associated proteins, play a role in the modulation of several signaling pathways, including PI3K-AKT, TGF-β, and NF-κB signaling. They interact with these pathways and can influence their activity, which are known as tumor-related pathways. (Gallolu Kankanamalage et al., 2018; Lai et al., 2014; Moniz and Jordan, 2010)

SMARCA1 locates on chromosome X. The protein encoded by SMARCA1 belongs to the SWI/SNF family. SWI/SNF family actives ATPase and helicase. It has ablity to change the chromatin structure nearby. Disordered chromatin remodeling regulation is related to cancer initiation. SMARCA1 plays both oncogene and TSG roles in many types of cancers, such as stomach, breast, lung, and cervical cancers. (A Patil et al., 2018; Li et al., 2021)

ZMYM3, also known as zinc finger, myeloproliferative, and mental retardation-type 3, is a gene that codes for a chromatin-interacting protein. It plays a critical role in promoting DNA repair through the process of homologous recombination. ZMYM3 is involved in regulating the localization of BRCA1, which is a key protein involved in DNA repair in damaged chromatin, thereby facilitating efficient DNA repair mechanisms. This gene is located on the X chromosome and is subject to X inactivation. (Brancaleoni et al., 2016) ZMYM3 deficiency resulted the instability of DNA repairing and possible cancer development, including ovarian and breast cancers. (Leung et al., 2017)

CYSLTR2, Cysteinyl leukotriene receptor 2, locates on the long arm of chromosome 13. It can bind to cysteinyl LTs such as LTC4, LTD4, LTE4) which are envolved in functions in endocrine and cardiovascular systems. It recurrently mutated in uveal melanoma. (Möller et al., 2017; Nell et al., 2021)

NCOR1 (nuclear receptor corepressor 1) is known to be implicated in thyroid development and thyroid cancer. (Fozzatti et al., 2013) This gene locates on chromosomes 17 and pseudogenes of this gene are found on chromosomes 17 and 20. It is linked to not only thyroid cancer, but also bladder cancer, breast cancer and prostate cancer. (Fozzatti et al., 2013; Lin et al., 2021; Noblejas-López et al., 2018; Tang et al., 2020)

STS on chromomsome X encodes Steroid sulfatase which is involved in the metabolism of steroids. Recent studies revealled that STS functions are involved in the pathways of a few cancers espacially hormon-dependent cancers such as breast cancer and prostate cancer. (McNamara et al., 2013; Shimizu et al., 2018) The inhibition of STS can lead to significance decrease hormone levels for cancer cells and therefore can be a potential drug target. (Daśko et al., 2020; McNamara et al., 2013; Shimizu et al., 2018)

Events in these genes are not only mutually exclusive with PTEN deletions, but they are also linked with cancers or other important biological functions. Additionally, according to literature, they have a large likelihood to be druggable. As a result, they represent promising PTEN synthetic lethality partners for cancer drug design.

3.2.3.7 Discussion

The bioinformatics screening of synthetic lethality partners of the cancer suppressor gene PTEN is the first step of new drug target gene discovery. It speeds up the progress of searching for new targets and saves time and resources for biological validation.

However, this method still has some disadvantages. The hypergeometric distribution test works only for frequently mutated genes If a “real” synthetic lethality partner is rarely mutated, the p-value would not be significant due to a lack of statistical power, so it will be missing. An increase of the cohort size would compensate for that effect.

Moreover, synthetic lethality analysis for new targets is not a perfect idea. It is known that human cells can have alternative repair pathways and they are likely to become drug resistant cells to the new drugs. Also, the new target genes are possible to play other important roles in related pathways, and therefore treatments with inhibitors might cause extra cell damage to normal cells.

Similar to other new drug screening projects, this study needs more detailed analysis of target gene biology to strengthen the result. The next steps are operated by our partner Proteros Biostructures GmbH.

3.3 CaSINo - A Scoring Algorithm to Interpret Mutation Significance

3.3.1 Motivation

Gene mutations play an important role in cancer initiation and progression. A few bioinformatics tools, such as Genome Music and MutSigCV, are developed for identifying significant cancer driver mutations in genomes, especially for coding genes. However, these tools are not designed for non-coding regions, codons or single nucleotides. Applying these existing tools to these objects is not straightforward. (Hong, 2016)

3.3.2 Method and pipeline implementation

In this study, huge quantities of NGS data from multiple cancer cohorts were obtained from The Pan-Cancer Analysis of Whole Genomes project of the International Cancer Genome Consortium (ICGC-PCAWG). Based on this data, I parsed all SNV information from 2962 patients of 48 different cohorts. ANNOVAR was then used for mutation functional annotation. I formatted the data into BED files for the follow-up analysis.

I implemented the CaSINo (Cancer SIgnificance analysis for Non-coding genomes) which calculates scores for each position/codon based on point mutation frequencies, but weights the information content by the mutation burden of the samples in which mutations are observed. Therefore, mutations from samples with few overall mutations have a stronger impact on the CaSINo score than mutations from samples with many mutations. I then selected the top-100 candidates with the highest scores as possible cancer drivers. (Hong, 2016)

To assess the ability of the CaSINo score to identify relevant mutations, I compared these top candidates with known cancer driver genes from Cancer Gene Census. Additionally, for the remaining candidates, I performed Qiagen Ingenuity Pathway

Analysis (IPA) and sorted out the related pathways and diseases. The result indicates a strong enrichment of the selected candidates in cancer-related pathways. (Hong, 2016)

Table 7: Selected cohorts for CaSINo analysis.

Cohort	Size	Cohort	Size	Cohort	Size
BLCA-US	23	KICH-US	49	OV-US	45
BOCA-UK	76	KIRC-US	40	PACA-AU	99
BRCA-EU	79	KIRP-US	34	PACA-CA	150
BRCA-UK	46	LAML-KR	10	PAEN-AU	105
BRCA-US	92	LAML-US	33	PBCA-DE	251
BTCA-SG	12	LGG-US	19	PRAD-CA	124
CESC-US	20	LICA-FR	6	PRAD-UK	83
CLLE-ES	100	LIHC-US	54	PRAD-US	20
CMDI-UK	70	LINC-JP	31	READ-US	16
COAD-US	46	LIRI-JP	269	RECA-EU	95
DLBC-US	7	LUAD-US	42	SARC-US	34
EOPC-DE	71	LUSC-US	48	SKCM-US	38
ESAD-UK	100	MALY-DE	101	STAD-US	39
GACA-CN	42	MELA-AU	70	THCA-US	50
GBM-US	41	ORCA-IN	13	UCEC-US	51
HNSC-US	44	OV-AU	73	READ-US	16

The SNVs are called through the DKFZ SNV calling workflow by Ivo Buchhalter and annotated through ANNOVAR. (Jäger et al., 2013; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020) In this study, all of the SNVs with a confidence below 8 were neglected.

In this study, I developed a statistical model called "CaSINo" to assign scores to each position, codon, gene, and other relevant elements. Regions showing a significantly higher point mutation rate across the entire cohort receive higher scores. Conversely, mutations in patients with fewer overall mutations are considered more informative. The output of this algorithm provides a list of candidate functional elements, ranked based on their CaSINo scores. (Figure 12) (Hong, 2016)

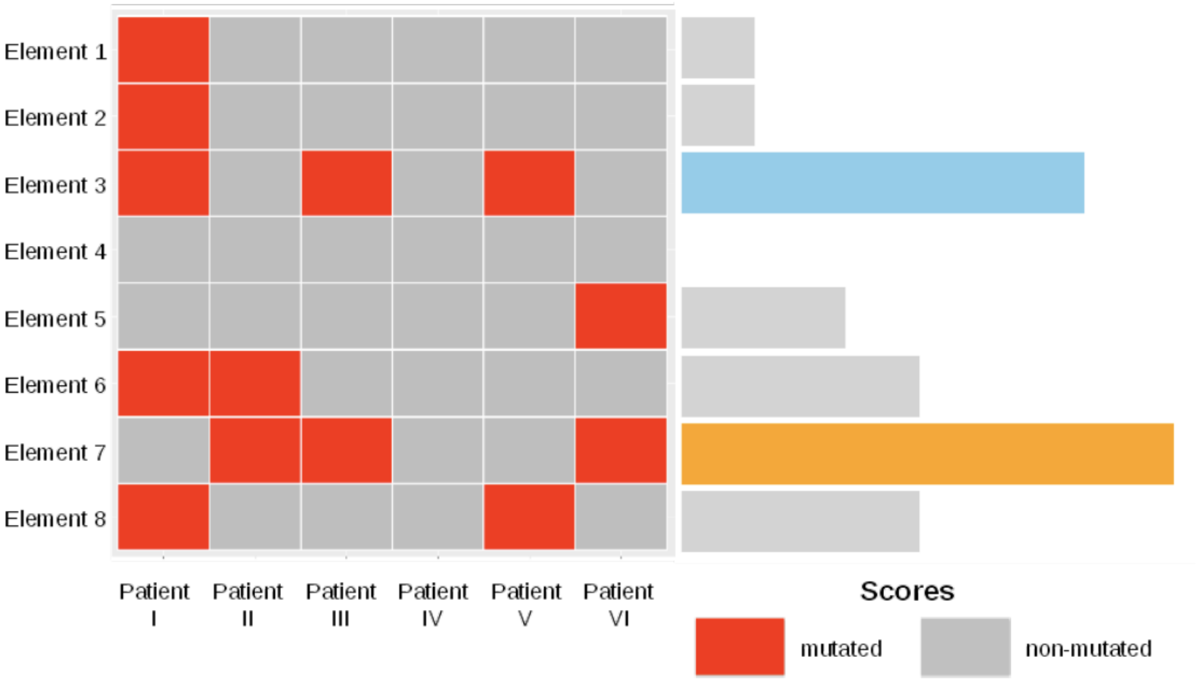


Figure 12: Schematic diagram of CaSINo.

For each functional element, a score was calculated not only by the cohortwise mutation recurrences but also by the individual patient background mutation rate. (Adapted from (Hong, 2016))

$$Score(r,cohort)=\frac{\Sigma_{p}^{cohort}\left(-log\left(\frac{SNVs(p)-SNVs(r,p)+1}{SNVs(p)+1}\right)\right)}{N_{cohort}}\times 1000$$

In this formula, *r* refers to a genome region, which can be either a single nucleotide, a codon, a non-coding element, or a gene. *p* refers to mutations from one patient mutation/normal pair. "*cohort*" refers to a set of patient samples. *SNVs(p)* refers to the total number of SNVs in one patient. *SNVs(r,p)* refers to the number of functional mutations that occur in region *r* and N_{cohort} refers to the total patient number of the cohort. (Hong, 2016)

In this study, I used functional SNVs, such as non-synonymous mutations, stop-gain or stop-loss mutations as *SNVs(r,p)* and used all types of SNVs as SNVs(p) as the background mutational rate.

3.3.3 Results

I calculated the CaSINo score for both mutated positions and mutated codons. (Table 8-9)

Table 8: Top 20 mutated positions with highest CaSINo scores.

gene	position	score	role
TIPIN	15:66641448:TIPIN	0.00557872475374628	NA
BRAF	7:140453136:BRAF	0.00513523465793539	oncogene, fusion
KRAS	12:25398284:KRAS	0.00477408641543645	oncogene
FAM174B	15:93198688:FAM174B	0.00438019804840099	NA
JAK2	9:5073770:JAK2	0.00435965029916284	oncogene, fusion
FAM174B	15:93198687:FAM174B	0.00280929488876375	NA
GBP4	1:89652088:GBP4	0.0021676901083092	NA
RP1L1	8:10467628:RP1L1	0.00216157424648286	NA
GBP4	1:89652090:GBP4	0.00213293319459771	NA
IGFN1	1:201180317:IGFN1	0.00204246380040236	NA
TPRXL	3:14106332:TPRXL	0.00202385850681739	NA
PIK3CA	3:178952085:PIK3CA	0.00199396549476417	oncogene
HLA-DRB1	6:32552060:HLA-DRB1	0.00185079665532673	NA
QRICH2	17:74288410:QRICH2	0.00172987345714036	NA
HLA-DQA1	6:32609271:HLA-DQA1	0.00171383483680916	NA
ATP1A3	19:42470962:ATP1A3	0.00170070475958498	NA
GBP4	1:89652087:GBP4	0.00169760940310457	NA
HLA-DQA1	6:32609278:HLA-DQA1	0.00167795226265418	NA
KRTAP9-1	17:39346622:KRTAP9-1	0.00167076299295663	NA
HLA-DQA2	6:32714125:HLA-DQA2	0.00166051511514349	NA

Notes: Each column represents the gene names, mutation positions, CaSINo socres and the roles of genes in CGC list.

Table 9: Top 20 mutated codons with highest CaSINo scores.

gene	codon	score	role
FAM174B	FAM174B:p.S68	0.00718949293716475	NA
KRAS	KRAS:p.G12	0.00603122834678532	oncogene
TIPIN	TIPIN:p.R142	0.00557872475374628	NA
TIPIN	TIPIN:p.R41	0.00557872475374628	NA
BRAF	BRAF:p.V600	0.00515421412977671	oncogene, fusion
BRAF	BRAF:p.V28	0.00515421412977671	oncogene, fusion
JAK2	JAK2:p.V617	0.00435965029916284	oncogene, fusion
JAK2	JAK2:p.V468	0.00435965029916284	oncogene, fusion
GBP4	GBP4:p.M545	0.0043006233029069	NA
DSPP	DSPP:p.D673	0.00282263267249971	NA
HLA-DQA1	HLA-DQA1:p.A92	0.00252451047643133	NA
HLA-DQB1	HLA-DQB1:p.G121	0.00229541431993378	NA
RP1L1	RP1L1:p.T1327	0.00223969938866847	NA
IDH1	IDH1:p.R132	0.00211030292005979	oncogene
PIK3CA	PIK3CA:p.H1047	0.00207690802875385	oncogene
IGFN1	IGFN1:p.E2099	0.00204246380040236	NA
TPRXL	TPRXL:p.S219	0.00202385850681739	NA
HLA-DRB1	HLA-DRB1:p.S66	0.00185079665532673	NA
C2orf82	C2orf82:p.A10	0.00184432670543624	NA
DPP7	DPP7:p.L47	0.00184432670543624	NA

Notes: Each column represents the gene names, names of codons, CaSINo scores and the roles of genes in CGC list.

To check the plausibility of the results, I compared the top 100 candidates of both mutation levels to the Cancer Gene Census list. The Cancer Gene Census (CGC, https://cancer.sanger.ac.uk/census) provides a list to catalogue the cancer driver genes and this list explains the roles these genes play in cancer development. The CGC list contains two tiers of confidences. I used both categories in the comparison.

(The CGC gene list mentioned in the paper refers to both tier 1 and tier 2 genes in the list. The version of CGC used in the thesis is v95.)

As of November 2020, the list includes 723 cancer related genes containing oncogenes, tumor suppressor genes and other genes with complicated functions.

In our study, I found 37 codon candidates out of our top 100 that can be found in the CGC list of genes. For the top 100 single nucleotide positions, I found 18 of them are involved in the CGC list.

Considering the CGC list is a very strict standard, I selected all the functional regions which are not listed in CGC from the top 100 lists and analyzed them in Qiagen Ingenuity Pathway Analysis (IPA® , QIAGEN Redwood City,www.qiagen.com/ingenuity).

IPA is a web-based pathway analysis software, which reports pathway details, biological processes and molecule function networks for a given gene list. The genes are annotated with high-quality gene ontology information and mapped into known pathways. A P-value of Fisher's Exact Test is calculated for each pathway and function. The P-value cutoff was set to 0.05 and a P-value closer to 0 implies that sample genes are enriched in the pathway or function. (Table 10-11)

Table 10: Functions and disease associations for the unknown genes from the top 20 codon gene list.

Diseases or Functions Annotation	p-value	Molecules
Susceptibility to multiple sclerosis	9.86E-06	HLA-DQB1,HLA-DRB1
Head and neck squamous cell carcinoma	1.14E-05	ACTC1,ACTR3C,AKAP17A,ATP1A3,DSPP,FAM186A,FAM8A1,HLA-DQA1,HLA-DQA2,HLA-DQB1, HLA-DRB1,IGFN1,KRTAP1-1,KRTAP4-5,KRTAP9-1,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1

Hypersomnia	3.28E-05	HLA-DQB1,HLA-DRB1
Extrapulmonary squamous cell carcinoma	1.25E-04	ACTC1,ACTR3C,AKAP17A,ATP1A3,DSPP,FAM186A,FAM8A1,HLA-DQA1,HLA-DQA2,HLA-DQB1, HLA-DRB1,IGFN1,KRTAP1-1,KRTAP4-5,KRTAP9-1,MUC22,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Pediatric inflammatory bowel disease	2.95E-04	HLA-DQA1,ICOSLG/LOC102723996
Embryonal rhabdomyosarcoma in striated muscle	4.53E-04	DSPP,FAM186A,HLA-DRB1,IGFN1,PLIN4,RP1L1
Muscle tumor	4.63E-04	ACTC1,DSPP,FAM186A,HLA-DQA1,HLA-DRB1,IGFN1,PLIN4,RP1L1
Dystrophy of muscle	4.81E-04	ACTA1,HLA-DQA1,HLA-DQB1,HLA-DRB1,IGFN1
Familial restrictive cardiomyopathy	4.94E-04	ACTA1,ACTC1
Hereditary myopathy	1.08E-03	ACTA1,ACTC1,ATP1A3,HLA-DQA1,HLA-DQB1,HLA-DRB1,IGFN1
Early-onset high myopia	1.21E-03	DSPP,KRTAP9-1
Celiac disease	1.39E-03	HLA-DQA1,HLA-DQB1
Duchenne muscular dystrophy	1.56E-03	HLA-DQA1,HLA-DQB1,HLA-DRB1
Acute myeloid leukemia	1.66E-03	ACTC1,ACTR3C,AKAP17A,DSPP,HLA-DQA1,HLA-DQB1,HLA-DRB1,KRTAP1-1,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1

Notes: The p-value here is calculated by Fisher's Exact Test and adjusted with methods based on the Benjamini-Hochberg. The detail algorithm can be found in IPA manual online. (https://qiagen.secure.force.com/KnowledgeBase/KnowledgeIPAPage?id=kA41i000000L5nQCAS)

Table 11: Functions and diseases for the unknown genes from the top 20 mut positions gene list.

Diseases or Functions Annotation	p-value	Molecules
Squamous-cell carcinoma	2.26E-07	ADGRE2,ATP1A3,ATXN1,C6,CES1,DSPP,FAM186A,FAM8A1,HLA-DQA1,HLA-DQA2,HLA-DQB1,HLA-DRB1, IGFN1,IGSF9B,KRT18,KRTAP4-5,KRTAP9-

		1,MACF1,MUC17,OR1L4,OR4A16,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Skin carcinoma	5.42E-07	ADGRE2,ATXN1,C6,DSPP,FAM186A,FAM8A1,GBP4,HLA-DRB1,IGFN1,KRT18,KRTAP9-1,MUC17,OR1L4,OR4A16,RP1L1
Extrapulmonary squamous cell carcinoma	8.19E-07	ADGRE2,ATP1A3,ATXN1,C6,CES1,DSPP,FAM186A,FAM8A1,HLA-DQA1,HLA-DQA2,HLA-DQB1,HLA-DRB1,IGFN1,KRT18,KRTAP4-5,KRTAP9-1,MACF1,MUC17,OR1L4,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Head and neck squamous cell carcinoma	1.36E-06	ATP1A3,ATXN1,DSPP,FAM186A,FAM8A1,HLA-DQA1,HLA-DQA2,HLA-DQB1,HLA-DRB1,IGFN1,KRT18,KRTAP4-5,KRTAP9-1,MACF1,MUC17,OR1L4,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Embryonal rhabdomyosarcoma in striated muscle	4.54E-06	ATXN1,DSPP,FAM186A,HLA-DRB1,IGFN1,MACF1,PLIN4,RP1L1
Cancer of cells	4.57E-06	ADGRE2,ATP1A3,ATXN1,C6,CES1,DSPP,FAM186A,FAM8A1,GZMB,HLA-DQA1,HLA-DQA2,HLA-DQB1,HLA-DRB1,IGFN1,IGSF9B,KRT18,KRTAP4-5,KRTAP9-1,MACF1,MUC17,OR1L4,OR4A16,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Susceptibility to multiple sclerosis	9.34E-06	HLA-DQB1,HLA-DRB1
Hypersomnia	3.11E-05	HLA-DQB1,HLA-DRB1
Muscle tumor	6.22E-05	ATXN1,DSPP,FAM186A,HLA-DQA1,HLA-DRB1,IGFN1,MACF1,PLIN4,RP1L1
Soft tissue sarcoma	1.04E-04	ATXN1,DSPP,FAM186A,HLA-DRB1,IGFN1,KRT18,MACF1,MUC17,OR4A16,PLIN4,RP1L1
Severe COVID-19	1.74E-04	C6,GZMB,HLA-DQB1,HLA-DRB1
Familial combined hyperlipidemia	2.03E-04	ATXN1,HLA-DQA1
Malignant myeloid neoplasm	2.06E-04	ATXN1,C6,DSPP,GZMB,HLA-DQA1,HLA-DQB1,HLA-DRB1,KRT18,OR1L4,OR4A16,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1
Bone marrow cancer	2.19E-04	ATXN1,C6,DSPP,GZMB,HLA-DQA1,HLA-DQB1,HLA-DRB1,KRT18,OR1L4,OR4A16,PABPC3,PLIN4,QRICH2,RP1L1,SIRPB1

Notes: The p-value here is calculated by Fisher's Exact Test and adjusted with methods based on the Benjamini-Hochberg. The detail algorithm can be found in IPA manual online. (https://qiagen.secure.force.com/KnowledgeBase/KnowledgeIPAPage?id=kA41i000000L5nQCAS)

These genes are involved in quite a number of cancer pathways with significant p-values. It provides strong evidence that our findings are interesting.

3.3.4 Discussion

CaSINo scores are based on basic statistics and can analyze point mutations that occur in any region of the gene. This flexible method enables the analysis of many genetic elements including promoters, gene bodies or enhancers, and even codons. By downweighting information from hypermutated cases, the score improves over a pure mutation frequency analysis. Clearly, mutations, especially those in non-coding regions, have a complex impact on gene function. While it can't definitively tell us which genes are "potential cancer driver genes," it can tell us which genes are more likely candidates by ranking the scores.

Other approaches such as MutSigCV require very specific information that is not available for all cell types. Furthermore, this information is highly parametrized and therefore represents a strong bias. In contrast, CaSINo relies on very few assumptions and still is very powerful.

However, the CaSINo algorithm does not take into account the more complex biological functions of genes and does not explore complex covariant contexts. This makes this method currently only available as a complement to other feature-based methods. To further improve the performance of the algorithm, firstly, it is necessary to establish a more credible background mutation frequency model by means of simulation. Secondly, it is important to further quantify the biological significance of all functional modules. Finally, it is also necessary to quantify the obtained Score to establish a reasonable threshold.

3.4 Functional Promoter Mutations

Certain point mutations in promoters can significantly alter gene expression levels. If these mutations happen to act on cancer genes, they have potential implications for cancer development. The understanding of functional promoter mutations gives us an opportunity to improve our understanding of cancer etiology and may lead to new therapeutic options. It is valuable work to find similar cases based on big data from PCAWG. In this project, I set up an analysis pipeline to screen for specific promoter mutations that may upregulate gene expression.

3.4.1 Promoter Mutations and Expression Upregulation of TERT

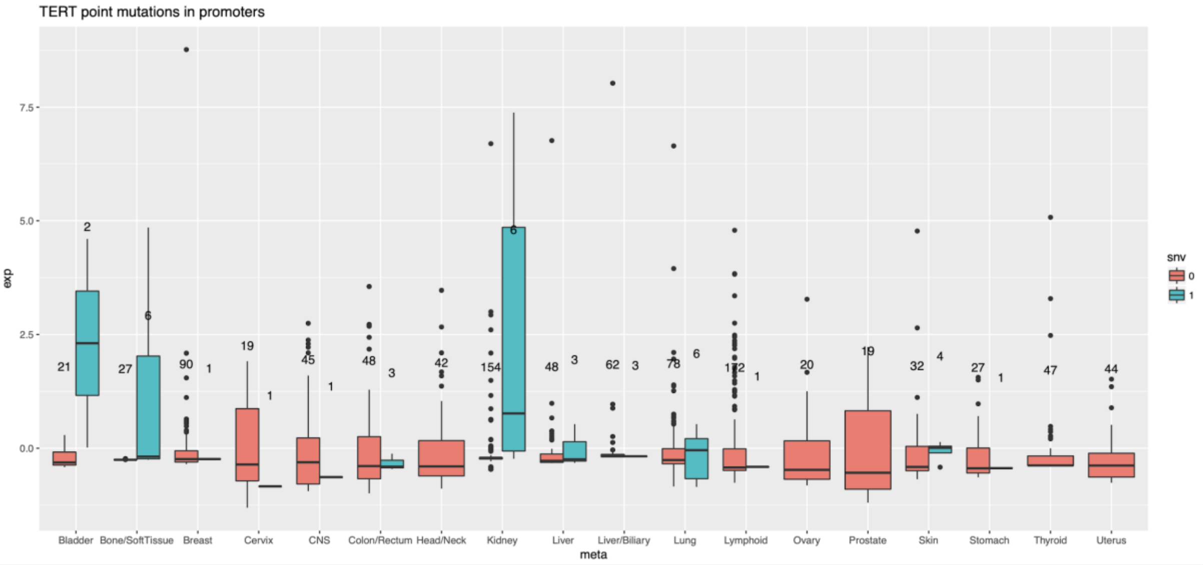


Figure 13: The promoter point mutations and expressions of TERT.

The x-axis represents the histological meta-cohorts and the y-axis stands for the expression value in z scores. The cyan boxplots represent samples with CNV in the TERT promoter and red boxplots represent the wild type. The numbers indicate how many samples there are in each group.

As shown in Figure 13, a few TERT promoter mutations are likely to be linked to expression upregulation, especially in bladder, bone, soft tissue, and kidney cancer. However, there are not too many patients with promoter mutations, which suggests that the promoter functional mutations are low-frequent.

3.4.2 Pipeline Design

With this idea, I developed a filter-based pipeline to screen out candidates of genes which fulfill all the following criteria:

- Point mutations occur in the promoter region.
- Expression is upregulated significantly.
- The gene coding region is not covered by CNV gains
- The point mutations change the motifs.

3.4.2.1 Data Preprocessing

Firstly, I defined the range of promoters in the whole genome. According to the gene coordinates in the HG19 database, I confirmed the boundaries of the TSS for each gene. I defined a promoter to start 2000bp upstream of the TSS. Then, I applied BED tools to identify point mutations that occur in the promoter regions in the PCAWG dataset.

The expression data was originally in FPKM format. In order to avoid batch effects from different cohorts, the FPKM values should be normalized. The normalization is based on:

$Z = (X_i - M_{cohort})/\sigma_{cohort}$

where X_i is the expression value in FPKM of a given gene for patient i, *Mcohort* is the mean expression value of the given gene in the cohort and the σ_{cohort} is the standard deviation of expression value from each patient within the cohort.

The expressions is often upregulated by copy number duplications in gene regions. To mitigate the impact of CNVs, I excluded the gene/patient-pairs in which the gene coding region is affected by a copy number change above 2.

3.4.2.2 FPKM selected list

In this study, I studied the 37,139 original whole genome genes, which are listed in the HG19 database. To filter for the false positive candidates, which are statistically upregulated but lack abundance for biological meaning, if the average expression value in FPKM of a gene is lower than 1. In this step, I kept 20,681 genes in our FPKM selected list.

3.4.2.3 Expression Filter

I set the Z-score threshold to 2. Genes exceeding this threshold were considered significantly upregulated, and SNVs and Indels in the corresponding promoter were kept in the potential candidate list.

3.4.2.4 Amplification Filter

The patient-gene-pairs that are annotated with CNV amplification above 4 were excluded from our candidate list. Amplifications are a stronger mechanism to upregulate genes than SNVs. Therefore, the Amplification Filter should exclude SNVs in promoter regions of amplified genes.

3.4.2.5 CGC Filter

In this study, I was interested in cancer driver genes so I have to narrow down our candidates in this direction. One of the best golden standards of cancer genes is the Cancer Gene Census (CGC) list, which is created by the Catalogue of Somatic Mutations in Cancer (COSMIC) (ttps://cancer.sanger.ac.uk/cosmic/download). In total 719 genes are recorded on the CGC list that have been shown to be cancer driver genes. (Sondka et al., 2018) The CGC Filter selects the mutations which lie in promoter regions of cancer-driving genes from the CGC list.

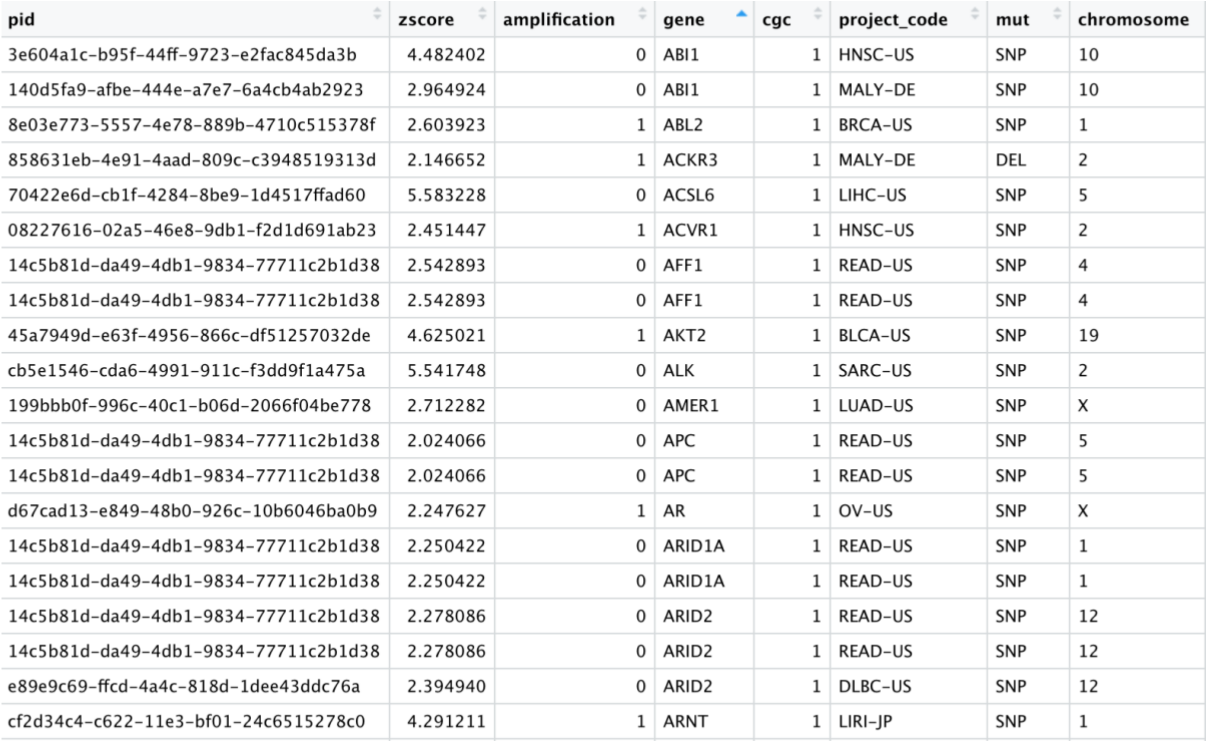
3.4.2.6 Motif Filter

The destruction or creation of a motif is considered to be the most important mechanism behind promoter mutational upregulation of gene expression. I applied the FIMO software with HOCOMOCO database to identify if the SNVs are located in

TFBS-related motifs. (Sondka et al., 2018)

3.4.3 Preliminary Results

Table 12: The preliminary result of functional promoter mutation analysis.



pid	zscore	amplification	gene	cgc	project_code	mut	chromosome
3e604a1c-b95f-44ff-9723-e2fac845da3b	4.482402	0	ABI1	1	HNSC-US	SNP	10
140d5fa9-afbe-444e-a7e7-6a4cb4ab2923	2.964924	0	ABI1	1	MALY-DE	SNP	10
8e03e773-5557-4e78-889b-4710c515378f	2.603923	1	ABL2	1	BRCA-US	SNP	1
858631eb-4e91-4aad-809c-c3948519313d	2.146652	1	ACKR3	1	MALY-DE	DEL	2
70422e6d-cb1f-4284-8be9-1d4517ffad60	5.583228	0	ACSL6	1	LIHC-US	SNP	5
08227616-02a5-46e8-9db1-f2d1d691ab23	2.451447	1	ACVR1	1	HNSC-US	SNP	2
14c5b81d-da49-4db1-9834-77711c2b1d38	2.542893	0	AFF1	1	READ-US	SNP	4
14c5b81d-da49-4db1-9834-77711c2b1d38	2.542893	0	AFF1	1	READ-US	SNP	4
45a7949d-e63f-4956-866c-df51257032de	4.625021	1	AKT2	1	BLCA-US	SNP	19
cb5e1546-cda6-4991-911c-f3dd9f1a475a	5.541748	0	ALK	1	SARC-US	SNP	2
199bbb0f-996c-40c1-b06d-2066f04be778	2.712282	0	AMER1	1	LUAD-US	SNP	X
14c5b81d-da49-4db1-9834-77711c2b1d38	2.024066	0	APC	1	READ-US	SNP	5
14c5b81d-da49-4db1-9834-77711c2b1d38	2.024066	0	APC	1	READ-US	SNP	5
d67cad13-e849-48b0-926c-10b6046ba0b9	2.247627	1	AR	1	OV-US	SNP	X
14c5b81d-da49-4db1-9834-77711c2b1d38	2.250422	0	ARID1A	1	READ-US	SNP	1
14c5b81d-da49-4db1-9834-77711c2b1d38	2.250422	0	ARID1A	1	READ-US	SNP	1
14c5b81d-da49-4db1-9834-77711c2b1d38	2.278086	0	ARID2	1	READ-US	SNP	12
14c5b81d-da49-4db1-9834-77711c2b1d38	2.278086	0	ARID2	1	READ-US	SNP	12
e89e9c69-ffcd-4a4c-818d-1dee43ddc76a	2.394940	0	ARID2	1	DLBC-US	SNP	12
cf2d34c4-c622-11e3-bf01-24c6515278c0	4.291211	1	ARNT	1	LIRI-JP	SNP	1

Notes: *Each column represents the patient id, expression in z score, amplification (0-normal, 1-amplificated), gene names, CGC list (0-not in CGC list, 1-in CGC list, here the non-CGC genes are neglected), project code in PCAWG, mutation types and chromosomes of genes.*

In the preliminary results, I obtained 566 SNVs in the candidate list. There are 551 non-recurrent SNVs in the whole database. Four recurrent SNVs were found. They are located in the promoters of the genes TERT, PIM1, BCL2 and EBF1. This result is currently referred to as a "preliminary result" since my colleague, Irina Glas, has taken over the subsequent analysis, and the new findings are not yet clear at the time of writing the thesis.

3.4.4 Discussion

In this project I used several different filters. These filters bring our results one step closer to real discovery. The first thing I have to say is that even with so many ways to eliminate irrelevant genes, there still exists a lot of problems. In terms of gene expression, in addition to our selected point mutations, gene expression is still affected by many other factors, such as epigenetics or the control of distant regulators. So here, I can't be sure that our filter will get the candidates which are needed without omission.

The first filter we employ here is the expression filter. I set a threshold of z larger or equal to 2. This relatively high threshold is set to reduce false-positive results caused by abnormal up-regulation of patient expression caused by other factors. On the other hand, since I need to actually find significant examples like TERT, after I evaluate the overall candidate size, I set 2 as the threshold.

In the design of the amplification filter, I considered that the causal relationship between amplification and gene expression up-regulation of many large copy numbers is more obvious. I think the up-regulation of gene expression with a copy number of 4 or more is likely to come from CNV. The lower copy number gains may not cause significant changes in gene expression. Based on this assumption, I screened out all genes that met this criterion.

The CGC filter is a very important step in the whole analysis process, because my research direction is limited to the research of cancer-related genes, so I hope to find the cancer driver genes that have been proven. This option may indeed give up some potential new SNVs, but given the sheer size of the potential candidates, I think adding a CGC filter is very sensible.

The motif filter is integral to this pipeline. I think it is only from this step that I can move from the black box of statistics to the analysis of biological functional modules. The motif scores report the likelihood of the potential binding sites or protein motifs and it can help us understand the biological function of sequence. Since my work was handed over to my colleagues at this step, I wasn't able to dig deeper into the performance and results of the motif filter, but I think this step will greatly eliminate the impact of false positives.

In the follow-up analysis, I think researchers should analyze from the impact of TFBS, these potential functional modules will greatly affect the level of gene expression. On the other hand, if it is possible to find high frequency mutation sites like TERT C228T and C250T, then this study will achieve greater scientific significance. Therefore, if possible, it is undoubtedly a very meaningful analysis to consider the mutation frequency of mutation points.

3.5 Focality Analysis of Copy Number Variations

3.5.1 Definition of Focal CNVs

In this study, focal CNVs are defined as the CNV deletions and amplifications that are longer than 1kb and shorter than 10Mb.

3.5.2 Overview of Focal CNVs in PCAWG Data

The lengths of CNVs differ greatly in genomes. The distribution diversity of CNV deletions in different cohorts is shown in Figure 14. Thyroid carcinoma and acute myeloid leukemia have shorter average CNV deletions while breast cancer and kidney cancer display longer deletions. This is possibly caused by different mechanisms of genomic variation in various cancer types. For example, in renal cell carcinoma, the short arm of chromosome 3 is often deleted and causes broad CNV deletion. (Quddus et al., 2019)

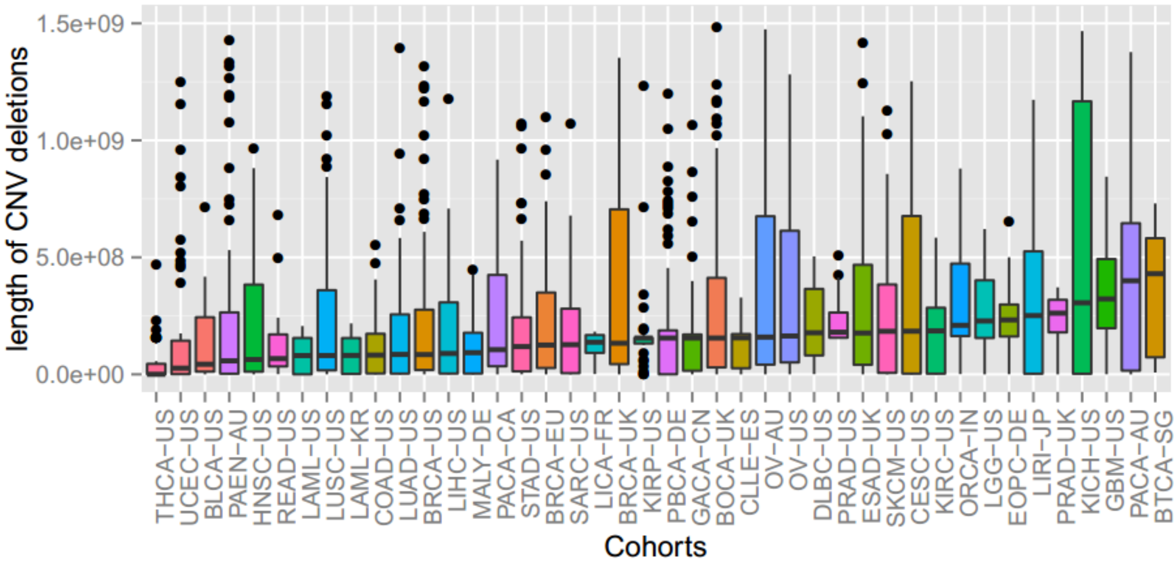


Figure 14: Overview of CNV deletion length distribution in different cohorts.

The y-axis represents the sum of CNV deletion length in one patient and the x-axis represents the cohort. The boxplots are sorted by the median length of CNV deletions. It is shown that CNVs influence differently among cancer types. (Adapted from (Hong, 2016)*)*

Focal CNVs are enriched in certain regions in the human genome. These regions are believed to be fragile sites inside chromosomes. (Le Tallec et al., 2013)

In the view of genes, I found that focal deletions are more abundant in cancer-related genes such as FHIT, CSMD1, WWOX, and PTPRD.

Table 13: Top 20 most frequently CNV loss genes.

chromosome	start	end	gene	Number of patients with deletion
3	59735036	61237133	FHIT	1104
8	2792875	4852494	CSMD1	889
16	78133310	79246564	WWOX	834
9	21802635	22032985	RP11-145E5.5	776
8	163186	182231	RPL23AP53	754
9	8314246	10612723	PTPRD	710
8	13947373	15095848	SGCZ	695
8	22570769	22857513	PEBP4	689
8	22877646	22926692	TNFRSF10B	686
8	24153327	24769586	RP11-624C23.1	685
8	22547663	22656129	RP11-459E5.1	683
8	22941868	22974950	TNFRSF10C	679
8	18384811	18942240	PSD3	677
8	22993101	23021543	TNFRSF10D	677
16	88781751	88851619	PIEZO1	677
8	22224762	22291642	SLC39A14	676
8	22925742	22941132	RP11-875O11.2	676
8	22928890	22932001	RP11-875O11.3	676
8	23047965	23082639	TNFRSF10A	676
8	15965387	16424999	MSR1	675
8	22132810	22215076	PIWIL2	675

Notes: Each column represents the chromosome, gene start position, gene end position, gene name and the nubmer of patients with deletion in gene.

Among the gene candidates shown in Table 13 in addition to some of the genes listed in CGC, I can also find other genes that are not related to cancer, but are only chromosomally close to these CGC genes. Therefore, to identify potential cancer driver genes, it is required to devise more precise statistical methods.

I also analyzed the lengths of focal CNV deletions in tumor suppressor genes

(TSGs) and genes which are not listed in CGC (non-cancer-related genes). I applied the Wilcoxon test to the counts of focal deletions (0-1Mb and 0-10Mb) and I found that shorter CNVs are more enriched in TSGs (2.2e-16) than longer ones (0.015). In both groups, the differences between the TSG group and the control group are significant. A shorter threshold causes more significant difference because longer events are more likely to effect neighbors of TSGs. (Figure 15)

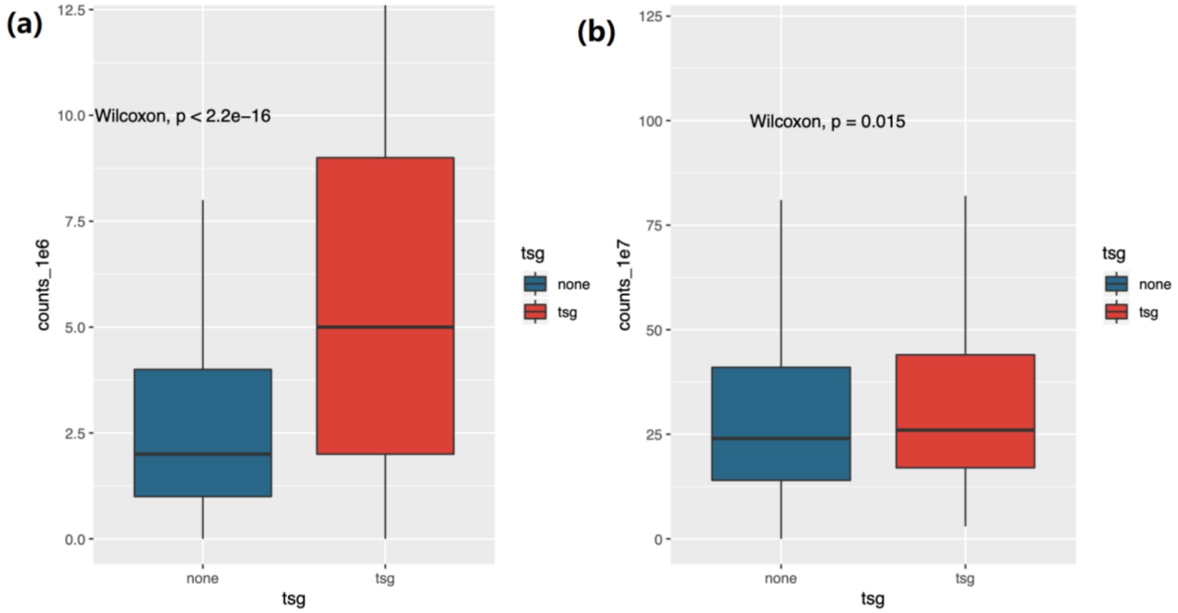


Figure 15: Focal CNV deletions in tumor suppressor genes (TSGs) and non-cancer-related genes with different length thresholds for focal CNV definition.

(a) CNVs shorter than 10^6 base pairs. (b) CNVs shorter than 10^7 base pairs. The difference of distributions of focal CNVs are significant between TSGs and non-cancer-related genes, especially if the definition length is shorter.

3.5.3 Focality Score Model

The statistical recurrence analysis indicates that focal CNV deletions tend to be enriched in TSGs. However, this study lacks accuracy. According to our previous study, shorter CNVs are more likely to be abundant than the longer ones. Additionally, even focal CNVs are sometimes comparatively longer than or similar to the sizes of genes. In this case, one CNV event could influence more than one gene.

If we use focal CNV enrichment to screen out possible cancer driver genes, it would result in a lot of potential false positives. I developed a simple mathematical method, which I called the ‘focality score’, to identify the short CNVs enrichment in cancer cohorts. The focality score is calculated as follows:

$$S = \sum_{i=1}^{m} \left(log(L_{max} - L_i)\right),$$

where *Lmax* is the defined focal CNV length upper limit (1Mb or 10Mb). *Li* is the length of the ith focal CNV which affects the gene.

The higher the focality score is, the more the gene is affected by shorter focal CNVs.

To solve the problems of neighboring genes sharing CNVs, I have to update the focality score by comparing the gene with its neighbors. So I defined the edge score:

score.edge = $(2*S_{gene} - S_{neighbour\_1} - S_{neighbour\_2})/2$

where neighbour_1 and neighbour_2 are neighboring genes of the target gene, if the target gene is at the edge of chromosome, the only neighbor gene counts as both neighbour_1 and neighbout_2.

In brief, the edge score represents the average difference between the standard focality score of a gene and the scores of its neighboring genes.

3.5.4 Focality Analysis for Driver Genes in PCAWG

I applied the focality score to the PCAWG dataset and I compared the focality score for CNV deletions of TSGs and the group of control genes.

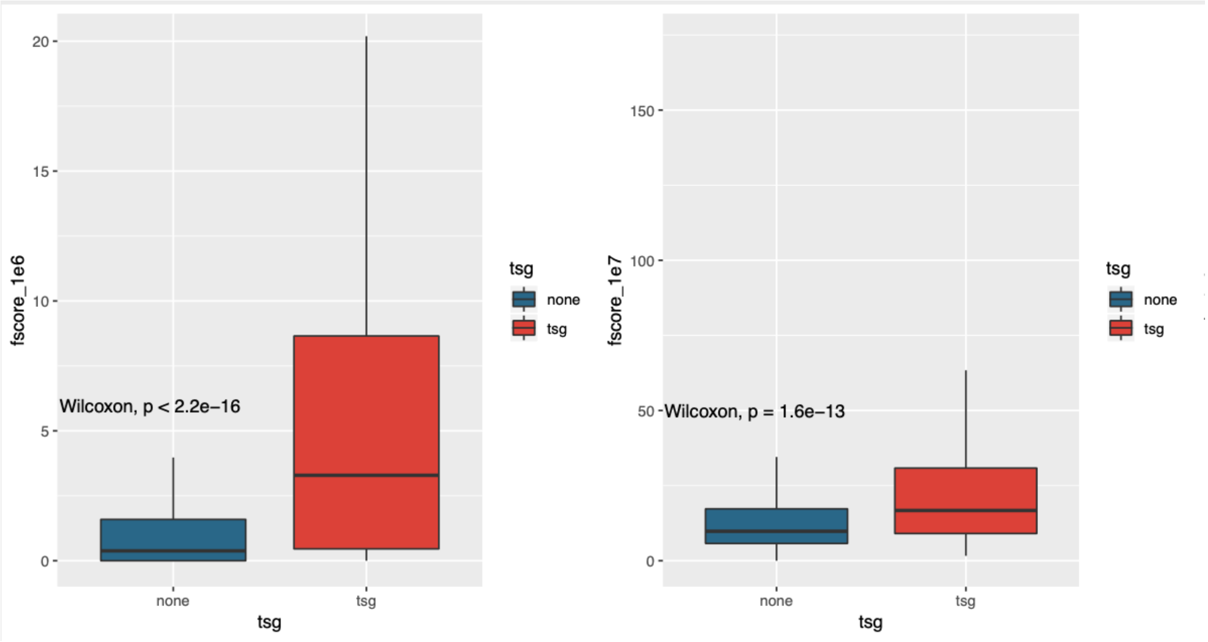


Figure 16: Focality scores in TSGs and non-cancer-related genes.

Significant differences between the two groups are shown in different length thresholds of 1Mb and 10 Mb.

The TSG genes have significantly higher focality scores than the control group (two-sided wilcoxon test p-value<2.2e-16 for CNVs shorter than 1Mb and p-value=1.6e-13 for CNVs shorter than 10Mb). The significance in Figure 16 is greater than the CNV deletion counts comparison of these groups which is shown in Figure 15.

I also tested the edge scores of the two groups. (Figure 17)

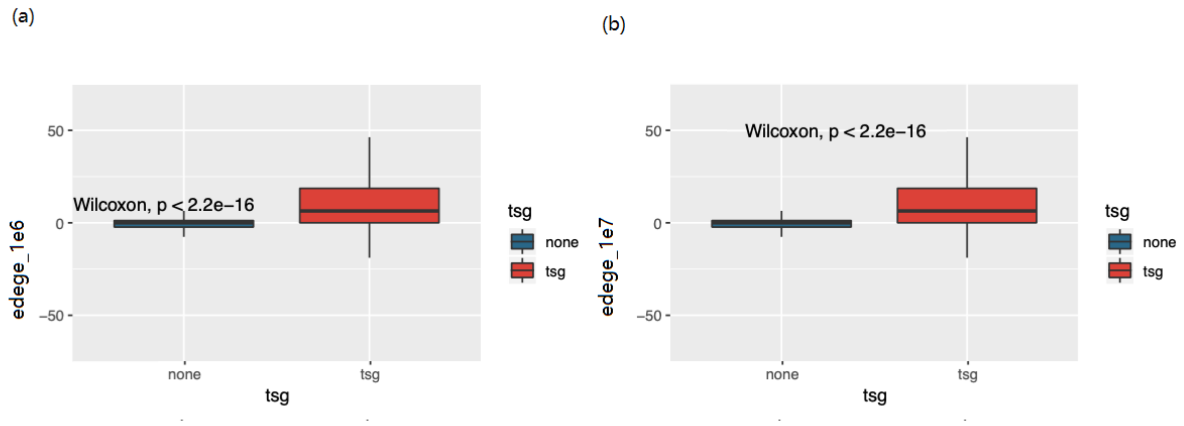
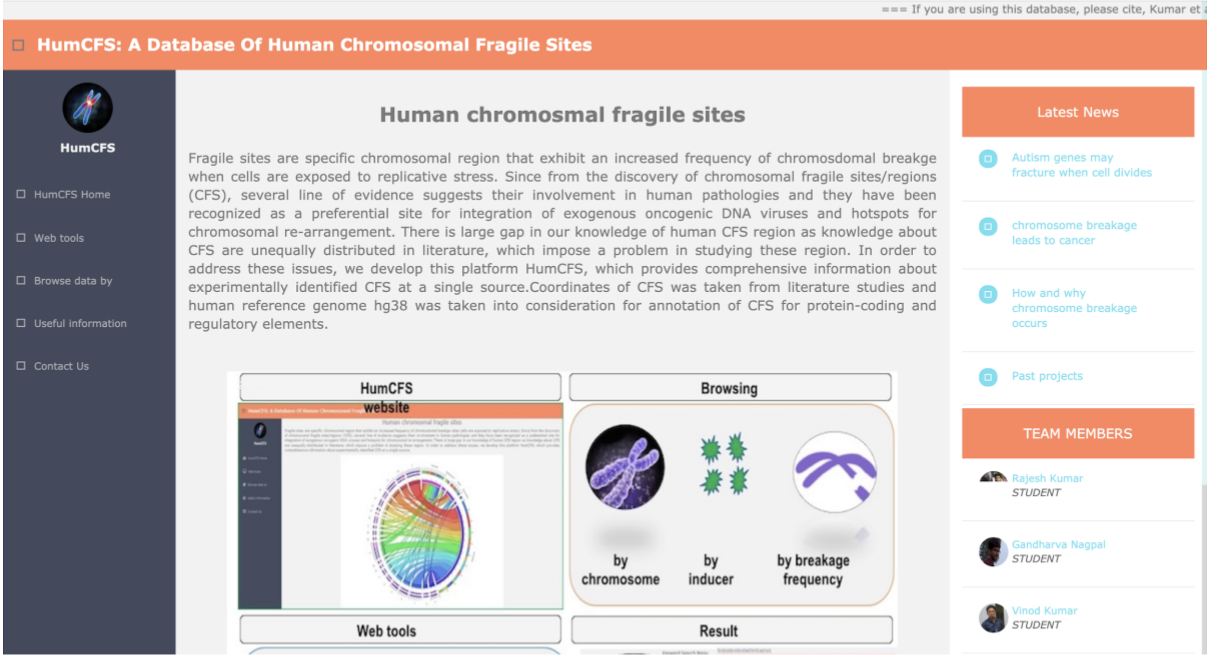


Figure 17: Edge scores in TSGs and non-cancer-related genes.

Significant differences between the two groups are shown in different length thresholds of 1Mb and 10Mb. The two-sided Wilcoxon test p-values are below 2.2e-16 in both tests.

According to these analyses, I concluded that focal CNV deletions are enriched in TSGs, especially the shorter events. Edge scores are more significant than original focality scores. Like SNV or CNV recurrence analysis, the focality score system can be used in potential cancer driver gene screening.

I applied the edge score to PCAWG data and ranked the gene list in descending order. I have to consider the fragile site of the chromosomes. “Chromosomal fragile sites are specific loci that preferentially exhibit gaps and breaks on metaphase chromosomes following partial inhibition of DNA synthesis.” (Durkin and Glover, 2007) Genes located in these regions can also be involved in CNV deletions but it might not relate to cancer progress. I have to filter out the genes in fragile sites from our candidate list. The fragile sites data can be found online. (Figure 18)



=== If you are using this database, please cite, Kumar et

□ HumCFS: A Database Of Human Chromosomal Fragile Sites

HumCFS

□ HumCFS Home

□ Web tools

□ Browse data by

□ Useful information

□ Contact Us

Human chromosmal fragile sites

Fragile sites are specific chromosomal region that exhibit an increased frequency of chromosdomal breakge when cells are exposed to replicative stress. Since from the discovery of chromosomal fragile sites/regions (CFS), several line of evidence suggests their involvement in human pathologies and they have been recognized as a preferential site for integration of exogenous oncogenic DNA viruses and hotspots for chromosomal re-arrangement. There is large gap in our knowledge of human CFS region as knowledge about CFS are unequally distributed in literature, which impose a problem in studying these region. In order to address these issues, we develop this platform HumCFS, which provides comprehensive information about experimentally identified CFS at a single source.Coordinates of CFS was taken from literature studies and human reference genome hg38 was taken into consideration for annotation of CFS for protein-coding and regulatory elements.

Latest News

Autism genes may fracture when cell divides

chromosome breakage leads to cancer

How and why chromosome breakage occurs

Past projects

TEAM MEMBERS

Rajesh Kumar
STUDENT

Gandharva Nagpal
STUDENT

Vinod Kumar
STUDENT

Figure 18: Webpage of HumCFS: A database of human chromosomal fragile sites.

This database provides coordinates and cytoband information of potential fragile sites. (Kumar et al., 2019)(https://webs.iiitd.edu.in/raghava/humcfs/)

Table 14: Top 30 genes with highest *focality scores.*

chr	gene	focality score	length	cgc	tsg	edge score
X	ARHGEF9	1605.16695187	150579	none	none	1507.1866407483
3	FHIT	983.081020351	1502097	cgc	tsg	723.06667374
16	WWOX	702.885882163	1113254	none	none	574.6395711025
20	MACROD2	419.008793511	2057827	none	none	392.0484454311
2	LRP1B	468.266307734	1900278	cgc	tsg	377.23439454875
4	CCSER1	514.096053126	1474378	none	none	370.346370954
16	PIEZO1	545.981268777	69868	none	none	359.916265197
9	PTPRD	523.842247213	2298477	none	none	358.722721107
5	PDE4D	513.872046231	1553082	none	none	337.7375715865
X	DMD	471.26291887	2241764	none	none	320.812832895
13	DCUN1D2	426.548420828	35133	none	none	300.301219457
6	PARK2	462.806541116	1380351	none	none	297.7487389495
9	RP11-143M1.7	801.29063767	20871	none	none	271.6443630845
9	RP11-145E5.5	816.881232864	230350	none	none	266.877221766
8	CSMD1	529.889931497	2059619	none	none	265.3359261175
19	OR4G1P	499.095780377	936	none	none	247.7431919
3	LSAMP	266.159193365	2194860	none	none	218.9999028209
22	TTC28	314.454388867	701849	none	none	213.4616675411
22	CECR2	216.463585853	197013	none	none	211.39076194082
9	CBWD1	815.021236355	67938	none	none	208.0689789125
18	RP11-451L19.1	551.401327872	4318	none	none	198.8106330885
16	CDT1	539.040613098	6045	none	none	174.8831704375
17	KSR1	185.391112767	169791	none	none	173.8951568951
3	NAALADL2	214.090323188	1367065	none	none	170.8018854681
10	PTEN	510.098596345	108817	cgc	tsg	169.5712707035
7	IMMP2L	243.489201756	899463	none	none	166.87010840805
13	TMCO3	432.504336804	59232	none	none	154.621341026
16	RBFOX1	200.942476258	1694245	none	none	154.2319972564
18	GREB1L	160.870797931	283175	none	none	151.8470848907
6	EYS	261.241127135	1987242	none	none	146.240906015

Notes: Each column represents chromosome, gene name, focality score, length of CNV, CGC list info, tumor suppressor gene info and edge score.

At the top of this list, I found a large number of known cancer driver genes, like FHIT, LRP1B and PTEN. These genes are highlighted in the CGC list. Moreover, the remaining genes are also interesting even if they are not listed in the CGC. For example, the ARHGEF9 gene is linked to the GPCR pathway, which plays a role in Pancreatic Adenocarcinoma and breast cancer regulation by Stathmin1. (Sriram et al., 2020)

3.5.5 Discussion

Copy number variation has strong effect on expression levels. From this logic, I can indeed deduce which genes may have a role in the development of cancer from the degree of CNV enrichment, especially the enrichment degree of focal CNV, which can reduce the interference of many false positive genes. The edge score method can effectively avoid genes with smaller lengths from being affected by neighboring genes, preventing us from misidentifying false positives around some hotspots in the analysis.

CNV deletion, thus present strong evidence for discovering new tumor suppressor genes. On the other hand, I found that the situation of CNV amplification is much more complicated, because it is difficult to determine what effect these CNVs appearing inside genes have on gene expression.

Moreover, the edge score and the CaSINO score, which were mentioned previously, complement each other and are based on some shared ideas. Both essentially use frequency information but take into account that the counts need to be weighted. In one case the overall mutation load is considered in the other case the length of the deletions.

I propose an algorithm for the quantification of enrichment for focal CNV deletion. This algorithm takes into account the length and frequency of copy number variation, and can effectively ignore the influence of excessively long CNVs to remove random interference. I realized that since this model does not take into account the effects of gene length and chromosomal local fragility, it does not perfectly represent the importance of CNVs. Even with edge score methods, it is difficult for us to find interesting candidates in some densely arranged shorter genes. To quantitatively analyze the impact of focal CNVs, more complex models and larger amounts of data is required.

I am delighted that many scientists have developed a strong interest in focal CNVs, which has also led me to a keen interest in how these analyses can be used for data visualization and standardization of software development. For this reason, I later developed a visualization tool GenomeTornadoPlot based on the R package, which allows each user to easily view the Focal CNV enrichment in a specific region of the chromosome and calculate its focus score. This work will be presented in Chapter 5.

4 Telomere Analysis

4.1 Background

Telomerase is up-regulated in about 85% of human cancers by different mechanisms, including TERT amplifications, structural variations or mutations in the TERT promoter. The remaining tumors utilize an alternative lengthening of telomeres (ALT) pathway, which involves DNA recombination of telomeric sequences. (Horn et al., 2013; Huang et al., 2013) Identifying distinct telomere maintenance mechanisms (TMMs) could provide greater insight into cancer initiation and progression. Based on different TMMs, scientists can try to develop corresponding diagnostic tools and anti-cancer therapies. Therefore, I can devise methods to determine the type of TMMs in a particular cancer patient through NGS data input. (Jafri et al., 2016)

At the molecular biology level, we can identify ALT with several different markers. However, most of them could not be directly detected in the short-read WGS data. While specific details regarding the ALT mechanism are not fully understood, it has been linked to loss-of-function mutations in chromatin remodeling genes ATRX (alpha-thalassaemia/mental retardation syndrome X-linked) and DAXX (death-domain associated protein). (Heaphy et al., 2020) In this study, we used ATRX and DAXX trunc mutations (ATRX/DAXXtrunc) as ALT indicators. I can use this as the ground truth to train machine learning models which train on features that can be measured in short-sequence NGS to effectively predict ALT. To be noticed, ATRX/DAXXtrunc samples do not cover all the ALL samples. (De Nonneville and Reddel, 2021) However they can be identified through NGS technology from PCAWG dataset. Therefore, I used these samples as a strong subset of ALT samples to train the classifier.

Non-supervised clustering of normalized TGAGGG, TCAGGG, TTGGGG, TTCGGG, and TTTGGG singleton repeat counts effectively separates most ATRX/DAXXtrunc samples from TERT modifications (TERTmod, such as amplifications, deletions, structural variations and point mutations) samples. The ATRX/DAXXtrunc clusters exhibit higher telomere content and a greater number of telomere insertions compared to the total number of breakpoints, which is not surprising. (Figure 19) (Sieverling et al., 2020)

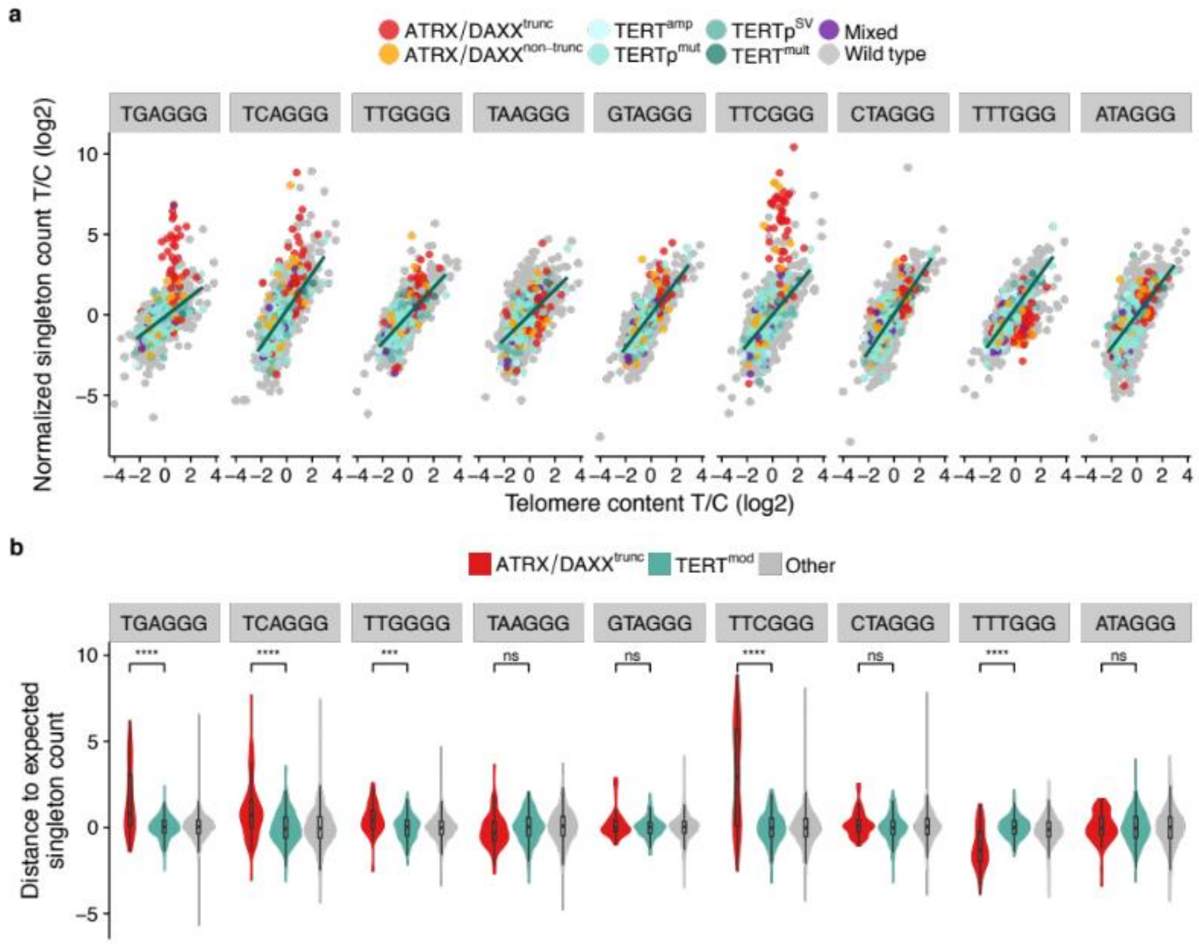


Figure 19: Telomere variant repeat in telomere.

*(a) Normalizaed count of tumor/control log2 ratios of all patients plotted against telomere content tumor/control log2 ratios for selected singletons. The regression line through the TERTmod samples is shown in green. (b) Distance to the expected singleton repeat count in ATRX/DAXXtrunc and TERTmod samples. ****p < 0.0001; **p < 0.01, two-sided Wilcoxon rank-sum tests after Bonferroni correction. (reprinted with original captions from Sieverling, Lina, et al. “Genomic Footprints of Activated Telomere Maintenance Mechanisms in Cancer.” Nature Communications, vol. 11, p. 733, 2020)*

4.2 Predictive Model for Identification of Telomerase/ALT

In this research, I built a random forest classifier. Through the R script, I used ATRX/DAXXtrunc (class I) and TERTmod (class II) samples as training datasets, with features including singleton divergence to expect count of TTTGGG, TTCGGG, TGAGGG, TCAGGG, TTGGGG, and breakpoint count, telomere insertion count and telomere content tumor/control log2 ratio for training. Considering that the sizes of two classes of sample are different, I used a down-sampling technique to subset telomerase samples before training, to solve the imbalance problem. After carefully tuning the parameters, I built a reliable classifier model. After 10-fold cross-validation, the classifier had an average area under the curve of 0.96, a sensitivity of 0.72, and a specificity of 0.98. The variables of highest importance for classification were the observed difference from expected counts of TTTGGG and TTCGGG singleton TVRs, the number of breakpoints, and the number of telomere insertions. (Sieverling et al., 2020)

Table 15: Feature importance in random forest classifier trained on ATRX/DAXXtrunc and TERTmod tumor samples.

Feature	Importance
TTTGGG singleton divergence to expected count	13.59
TTCGGG singleton divergence to expected count	11.92
Brekpoint count	11.01
Telomere insertion count	10.03
Telomere content tumor/control log2 ration	5.34
TGAGGG singleton divergence to expected count	5.02
TCAGGG singleton divergence to expected count	3.25
TTGGGG singleton divergence to expected count	2.83

The random forest model can not only classify, but also calculate a score for each test sample. This score can be interpreted as an ALT probability. As I expected, ATRX/DAXXtrunc had high ALT probability (mean = 0.92) in the test set, while TERTmod samples had low ALT probability (mean = 0.13). A total of 18 samples without ATRX/DAXXtrunc mutations had ALT probabilities over 0.9, two of which had non-truncating ATRX/DAXX mutations, and one had a frameshift insertion and TERT amplification in ATRX (11 TERT copies, triploid). (Figure 20-23) (Sieverling et al., 2020)

In the PCAWG dataset, the majority of cancer types displayed low probabilities of alternative lengthening of telomeres (ALT), indicating that their telomere maintenance mechanisms (TMMs) are primarily telomerase-based. This observation held true even for samples with ATRX/DAXX missense mutations, suggesting that these mutations may have a limited functional relevance and might be more incidental in nature. However, certain tumor types, such as leiomyosarcoma, osteosarcoma, and pancreatic endocrine tumors, exhibited high probabilities of ALT, which aligns with the well-documented prevalence of ALT in these particular cancer entities. (Sieverling et al., 2020)

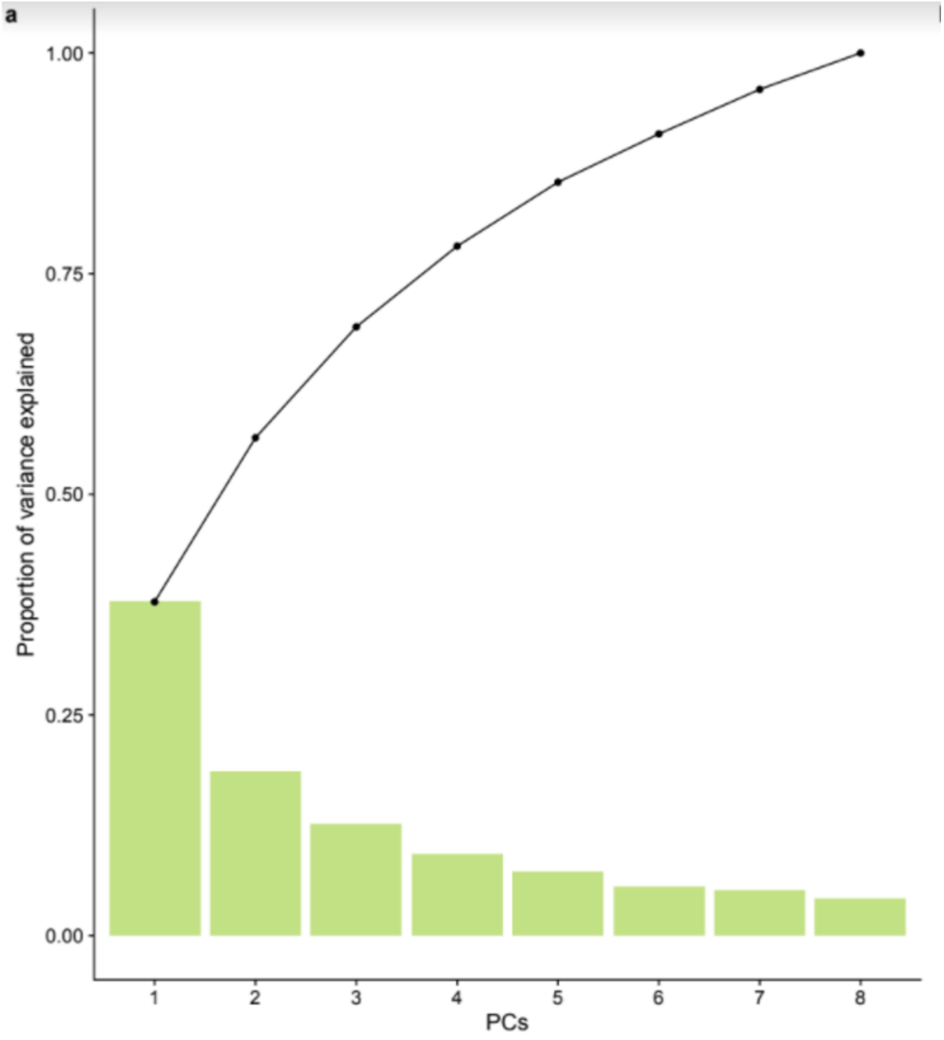


Figure 20: The proportion of variance explained in PCA analysis with selected features.

The selected features include telomere content tumor/control log2 ratio, number of telomere insertions, number of breakpoints and the distance of TGAGGG, TCAGGG, TTGGGG, TTCGGG and TTTGGG singletons. Top four Pcs explained more than 75% of variance.

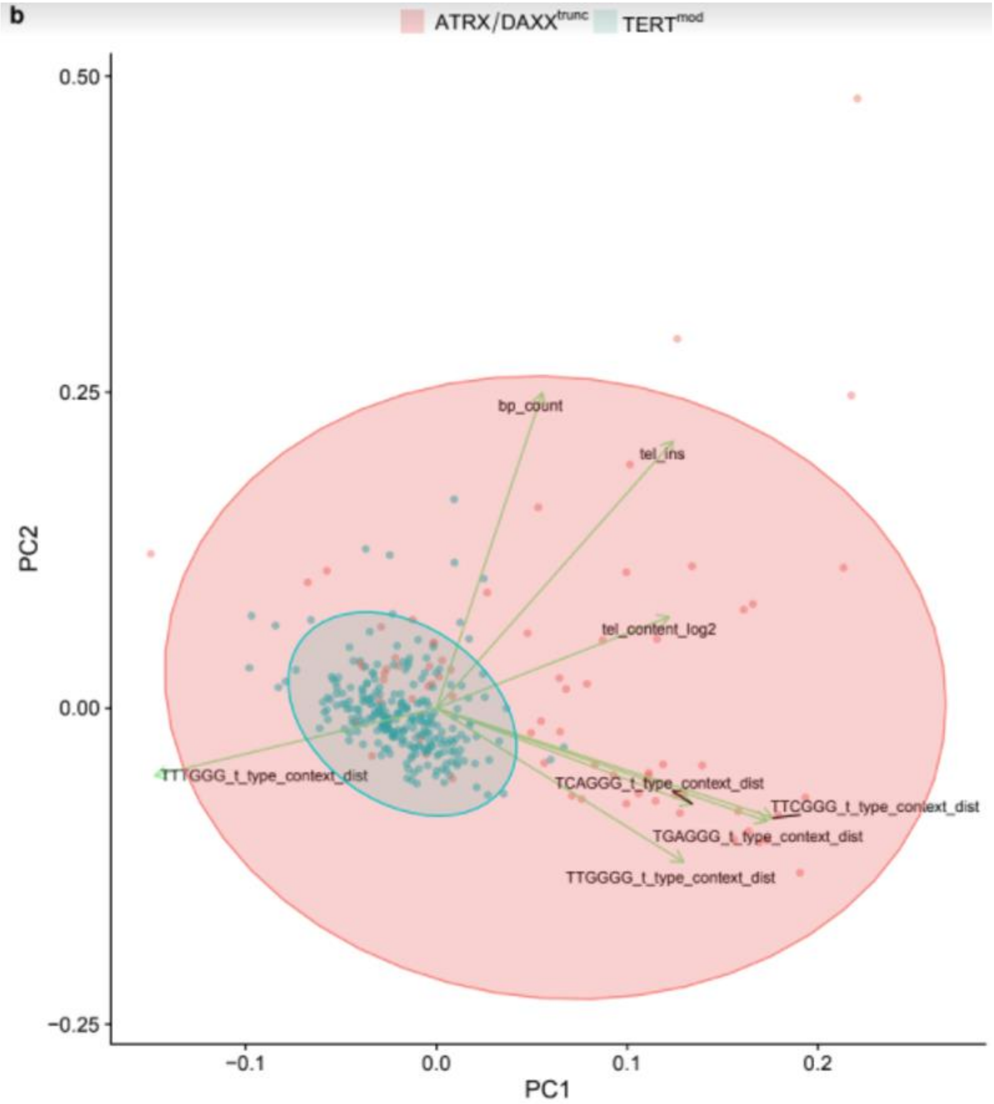


Figure 21: PCA analysis of the selected ALT features.

The selected features include telomere content tumor/control log2 ratio, number of telomere insertions, number of breakpoints and the distance of TGAGGG, TCAGGG, TTGGGG, TTCGGG and TTTGGG singletons. The red points represent ALT samples and cyan points for telomerase samples.

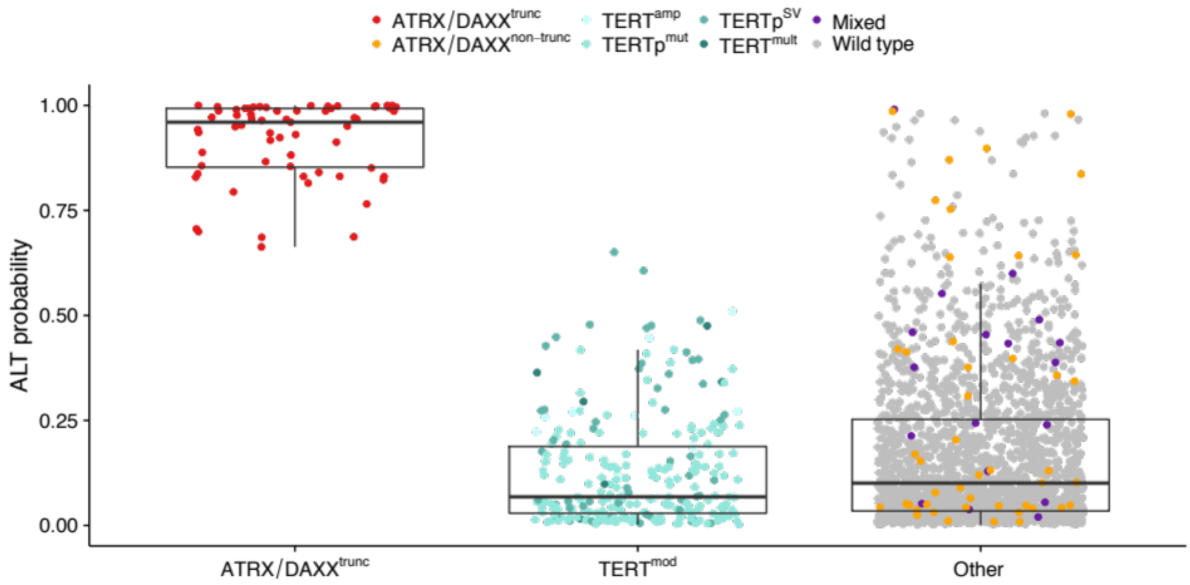


Figure 22: ALT prediction score of tumor samples with different TMM-associated mutations.

*The ALT probability was derived from a random forest classifier trained to distinguish ATRX/DAXXtrunc from TERTmod samples based on the following features: telomere content tumor/control log2 ratio, number of telomere insertions, number of breakpoints and the distance of TGAGGG, TCAGGG, TTGGGG, TTCGGG and TTTGGG singletons to their expected occurrence. The classifier was only applied to samples without missing data. The center line of the boxplot is the median, the bounds of the box represent the first and third quartiles, the upper and lower whiskers extend from the hinge to the largest or smallest value, respectively, no further than 1.5 * interquartile range (IQR) from the hinge. (reprinted with original captions from Sieverling, Lina, et al. “Genomic Footprints of Activated Telomere Maintenance Mechanisms in Cancer.” Nature Communications, vol. 11, p. 733, 2020)*

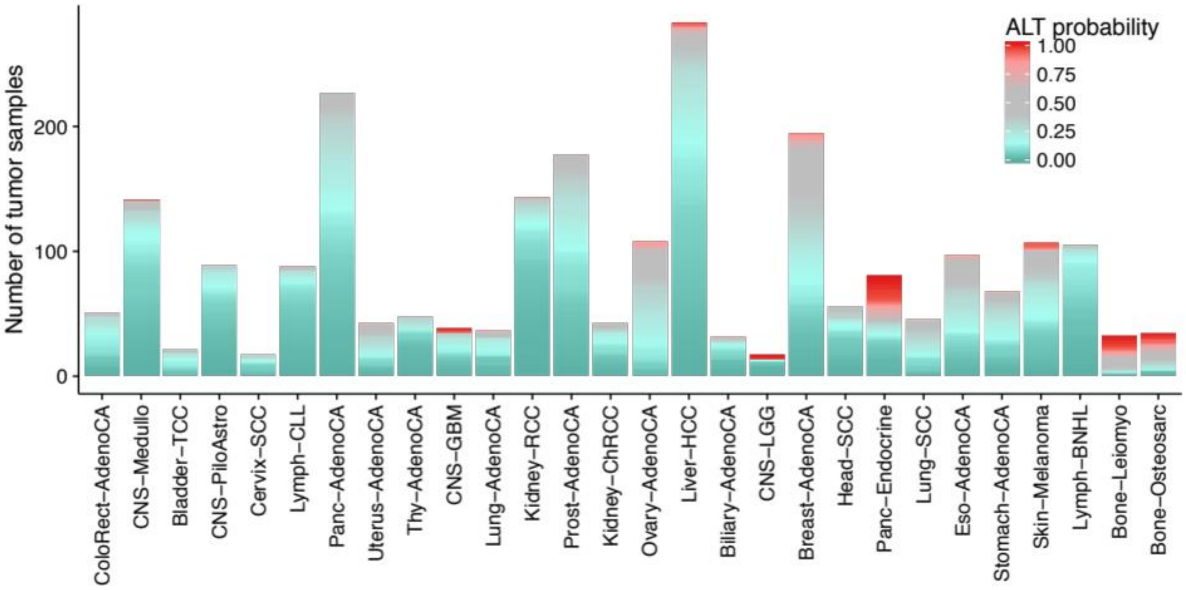


Figure 23: Prediction of ALT probability in different tumor types.

For each tumor sample, the ALT probability predicted by a random forest classifier is shown. Red represents high probability of ALT while cyan represents high probability of telomerase cases. The tumor types are sorted by mean telomere content tumor/control log2 ratio from left to right. Cohorts with sample sizes below 15 are not shown. (reprinted with original captions from Sieverling, Lina, et al. "Genomic Footprints of Activated Telomere Maintenance Mechanisms in Cancer." Nature Communications, vol. 11, p. 733, 2020)

4.3 Discussion

In the prediction of telomerase/ALT samples, I used a random forest model as a classifier and preprocessed the training set with a down-sampling method. The selected features are not only biologically molecularly meaningful, but also demonstrate representativeness by means of unsupervised clustering. The model has a good performance, achieving high AUC (0.96) and specificity (0.98).

However, in the calculation of this model, I found that the sensitivity is only 0.72, which is significantly lower than the specificity. This suggests that the model may mistake some ALT for telomerase samples. The reason for this result is probably due to the imbalance of the data or the scattered distribution of features in the ALT sample.

In addition, considering that the distribution of different TVRs in different cancer types is also different, in the case of sufficient data, if a separate classifier can be established according to different cancer types, the classification results can be better.

In an updated study, I found additional issues with this classifier. More important is the study from Nonneville and Reddel, who believe that the definition of ALT samples on which this model is based is not rigorous. (de Nonneville and Reddel, 2021) Since the notion that loss of ATRX/DAXX is essentially equivalent to the presence of ALT activity may only apply to specific types of tumors, it is not accurate to treat ATRX/DAXX as ALT globally. They believe that the ratio of ALT associated with ATRX/DAXXtrunc was overestimated in this study and misclassified ALT tumors when these mutations were absent. Therefore, although this classifier classified ATRX/DAXXtrunc and TERTmod, it could not perfectly distinguish ALT/telomerase. (De Nonneville and Reddel, 2021; Feuerbach, 2021)

However, the performance of the ALT probability score proposed with this classifier was also tested with the C-circle assay data as target variable in the new ROC curve analysis. The performance is also considered as robust. (Figure 24) (Feuerbach, 2021)

This work was completed in 2017, when only partial data was available for PCAWG. I believe that as more data becomes available and more accurate ground truth for ALT/Telomerase can be determined, this method will achieve more accurate results.

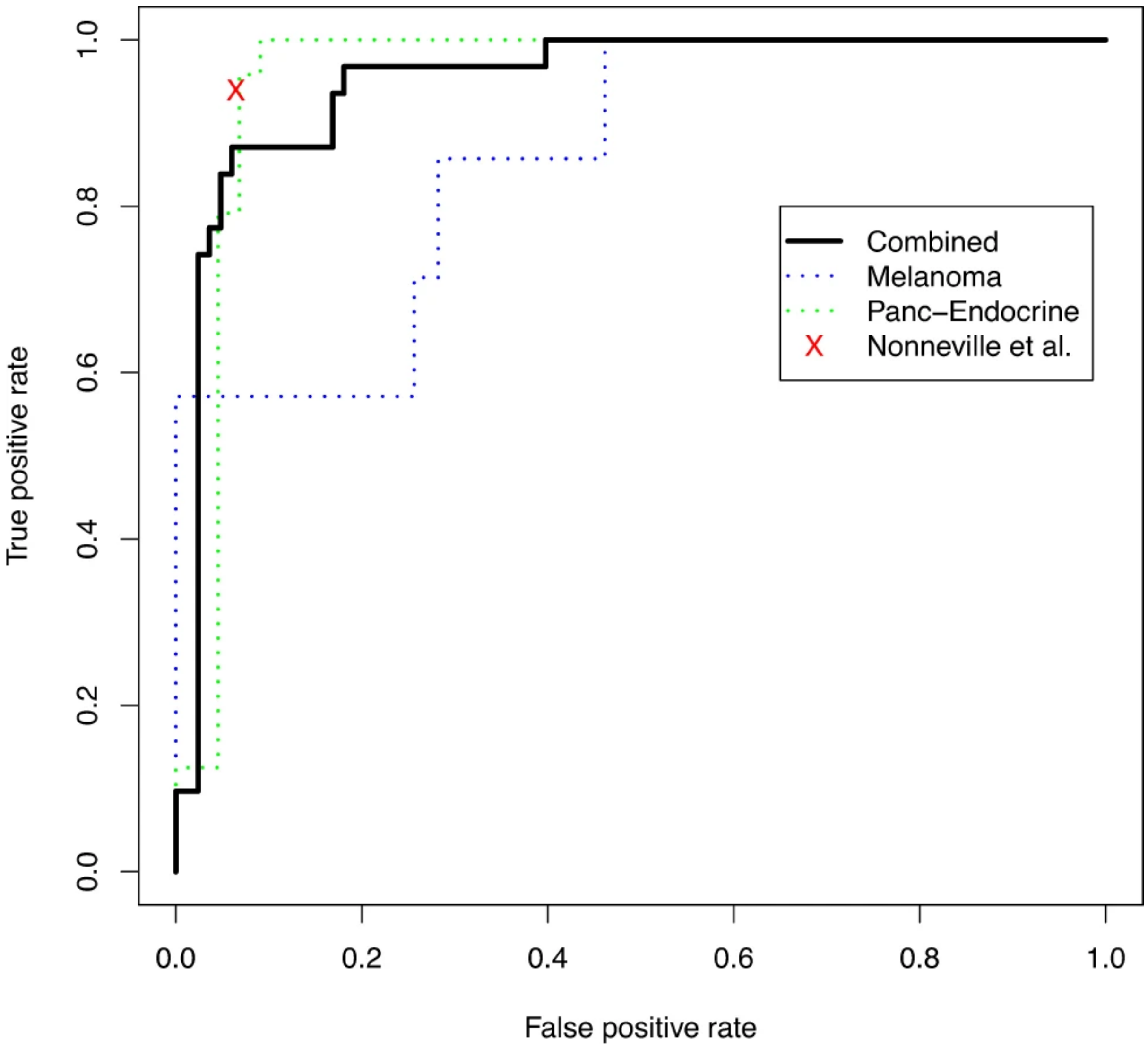


Figure 24: The curve of the true positive rate (sensitivity) versus the false positive rate (1 – specificity) for all possible thresholds on the ALT probability score.

The predicted class label is the C-circle status reported from Lee et al. (Lee et al., 2018), while the red cross depicts the performance of the classifier proposed in the Matters Arising article for which classification results are reported for only one threshold. (reprinted with original captions from Feuerbach, L. 'Formal reply to "Alternative lengthening of telomeres is not synonymous with mutations in ATRX/DAXX"', Nature communications, 12(1), pp. 1–3. 2021)

5 Visualization

Visualization is an integral part of data science nowadays. PCAWG data is collected from large numbers of patients and contains different types of data. It is very valuable for letting biologists and bioinformaticians quickly browse and query information from the database. Based on our solid understanding of the data, I implemented two powerful visualization systems: the TumorPrint and The GenomeTornadoPlot.

5.1 tumorPrint

5.1.1 Background

The motivation of TumorPrint is to quickly query and visualize gene mutation information. The TumorPrint has the following functions:

- showing the cohort distribution of gene mutations (both mono-allelic and bi-allelic inactivations; different levels of copy number amplifications)
- showing the gene variants and gene expression in every patient
- showing the entropy and CaSINo score of the gene
- showing the detail mutation information in each PCAWG cohort or meta-cohort based on histology

TumorPrint provides not only the mutations in genes and patients like Oncoprint, which is a widely used cancer variation visulaztion package, but also the detail information of variation types and RNA sequencing data. It helps users get intuitive insights of correlation between different mutation types and expression data. (Gu et al., 2016)

5.1.2 Basic Layout

The TumorPrint basical layout is shown in Figure 25.

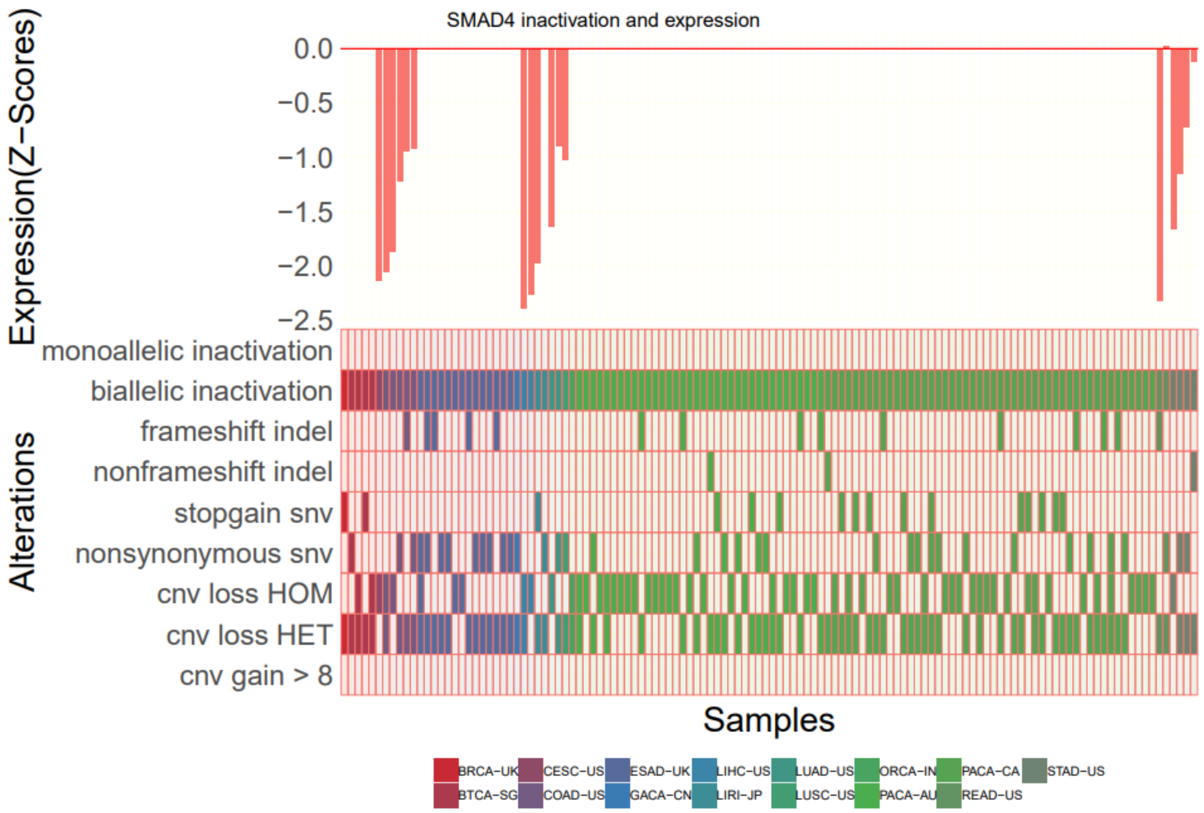


Figure 25: Schematic of TumorPrint plot of the mutation and expression in gene SMAD4.

The upper panel shows the expression (z-score, y-axis) of each sample (x-axis). The bottom panel shows the mutation types (including CNV loss, CNV gain and point mutations, y-axis) of each sample(x-axis). The colors represent the cohort origin of these patients. In this plot, the patients without genome mutations are neglected.

5.1.3 Implementation and Performance

The TumorPrint tool is coded in Python 2.7 and R 3.5.1 and can run in the DKFZ server by called with shell script. The parameters are simple, including gene or gene pair names, types of mutations, and file paths.

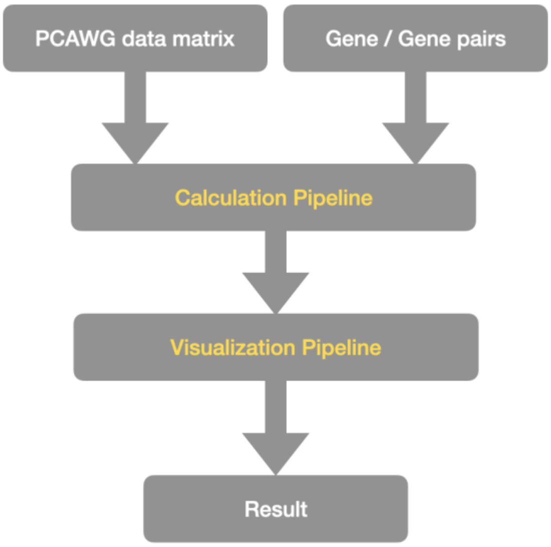


Figure 26: The workflow of the TumorPrint pipeline.

The input files include a PCAWG data matrix which is previously calculated and saved in cluster and configure shell script with the gene or gene pair name. The calculation pipeline is coded in Python. It extracts genome variations and RNA expression levels from the PCAWG data matrix for the given genes or gene pairs. It also transforms the raw mutation information into categories including mono-allelic inactivations, bi-allelic inactivations, and different levels of amplification. The visualization pipeline generates plots from the output of the calculation pipeline. It matches the z-scores of expression data and mutation categories for each individual patient and sorts them by expressions within each cohort. The result includes plots and a text file with stats of mutations.

For any gene or gene pair, the TumorPrint pipeline can generate the plots in less than 120 seconds with 1G RAM on DKFZ-ODCF server. The workflow of TumorPrint pipeline is shown in figure 26.

5.1.4 Applications and Examples

5.1.4.1 Single Gene Tumorprint

The single-gene TumorPrint is used to illustrate the bi-allelic inactivation, mono-allelic inactivation, amplification and expression levels from a single gene in cohorts. It also reports the distribution of genes in each cohort by calculating the entropy and generating a pie plot. The method of entropy calculation is the same as section

3.2.1. A few example are shown in Figure 27-30.

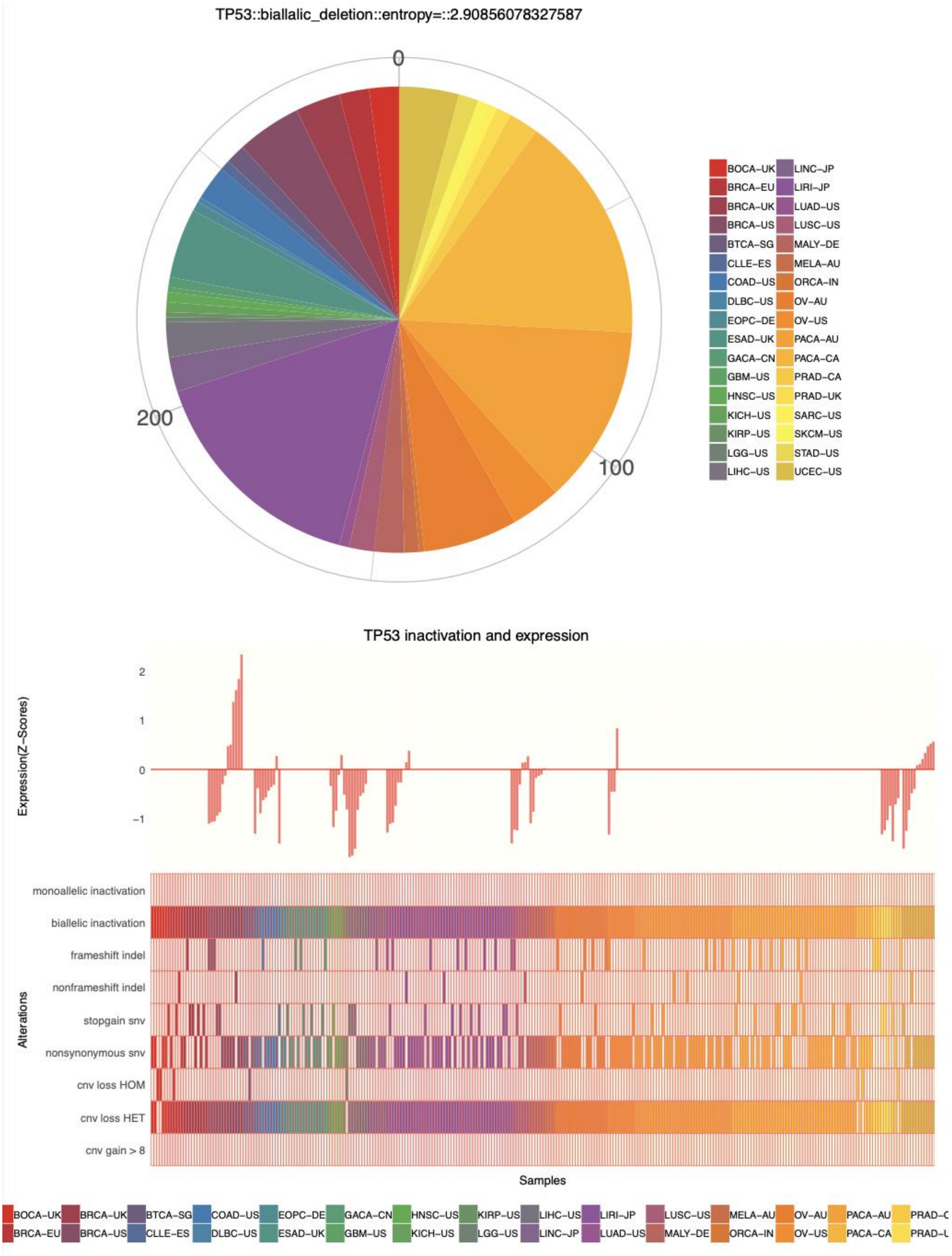
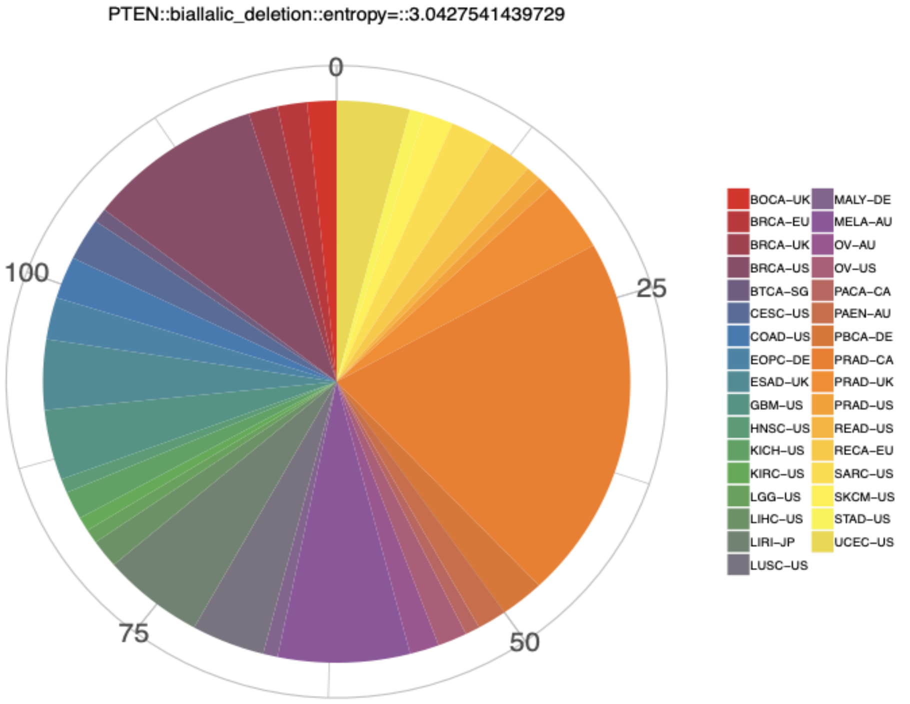


Figure 27: TumorPrint for bi-allelic inactivation of TP53.

TP53 is one of the most frequently inactivated TSG in multiple types of cancers. The bi-allelic inactivation cases are mainly resulted by the heterozygous deletion and nonsynonymous SNVs. The pie plot represents the number of patients with TP53 biallelic inactivations in each cohort. The barplot represent the z score of gene expression level of patient within cohort.



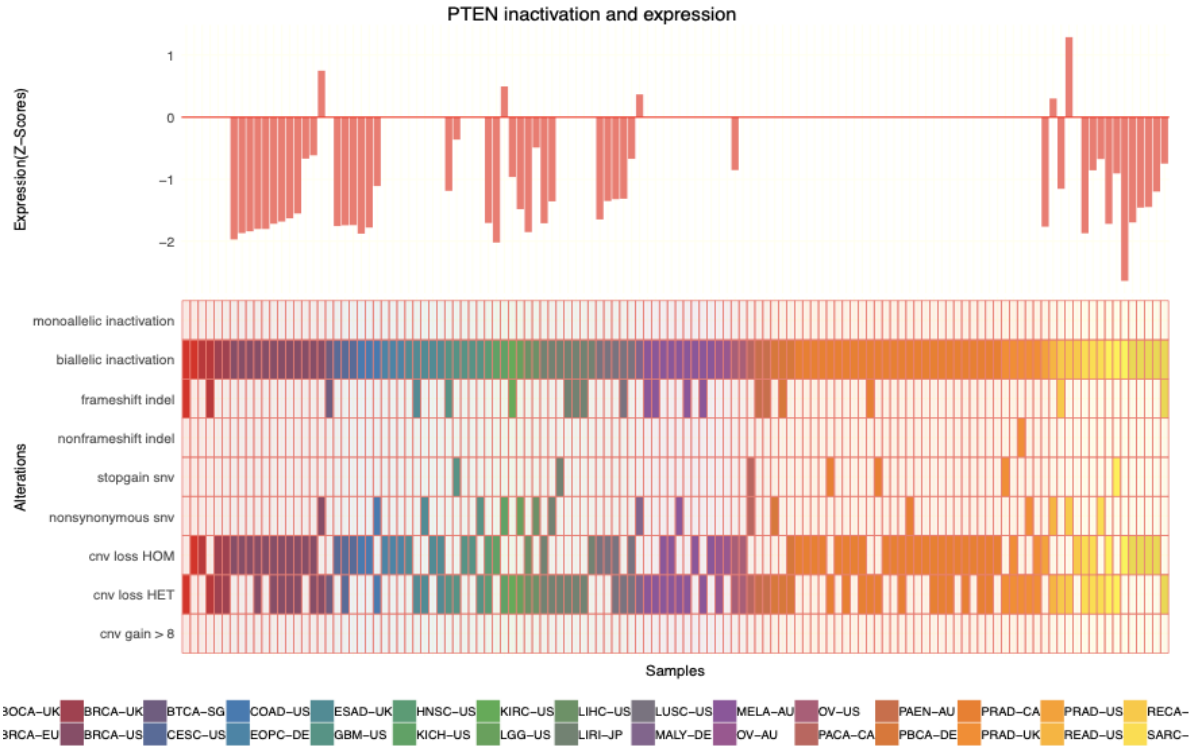


Figure 28: TumorPrint for bi-allelic inactivation of PTEN.

PTEN is one of the most frequently inactivated TSG in multiple types of cancers. The bi-allelic inactivation cases are resulted by both heterozygous deletion/nonsynonymous SNVs and homozygous deletions. The pie plot represents the number of patients with PTEN biallelic inactivations in each cohort. The barplot represent the z score of gene expression level of patient within cohort.

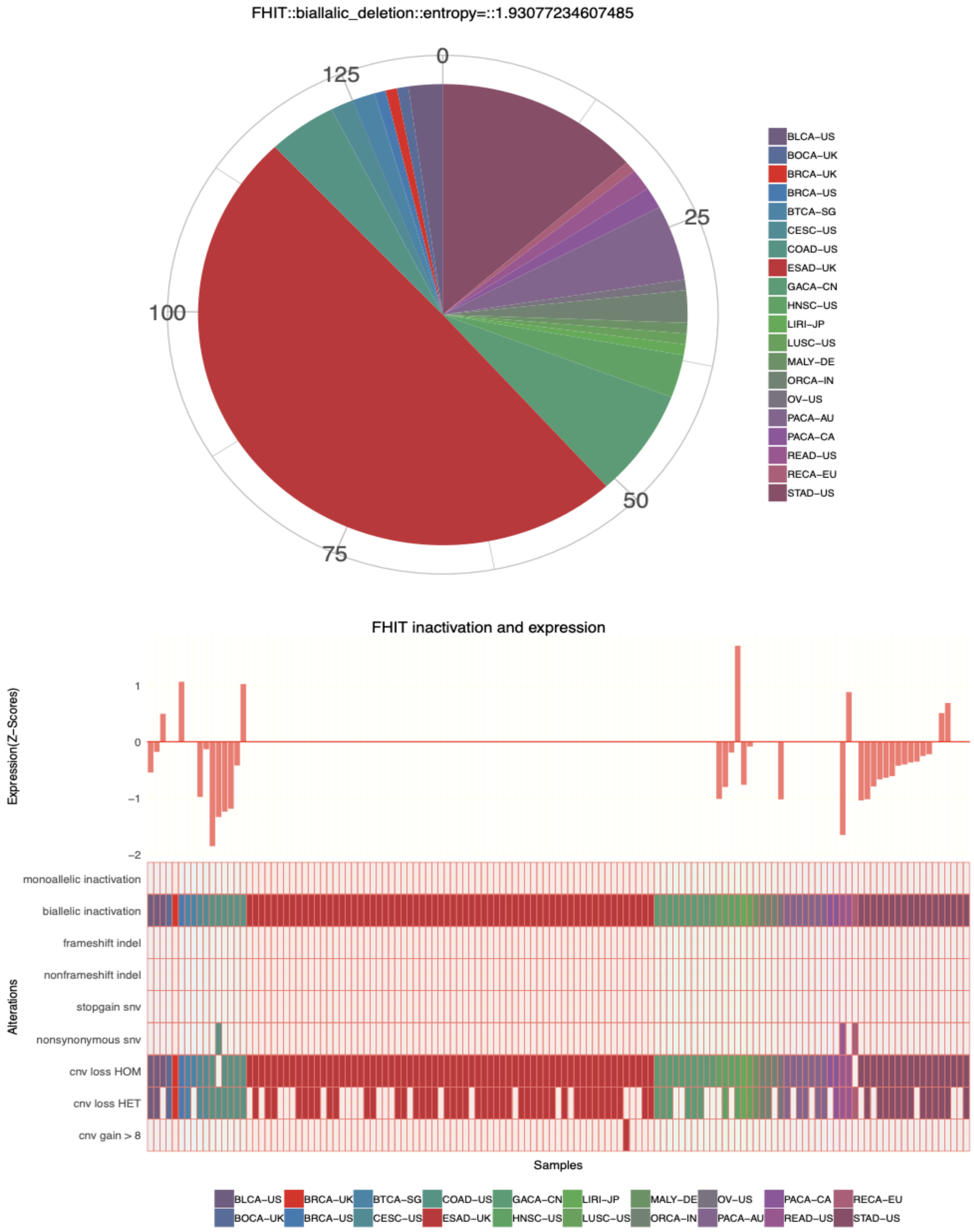


Figure 29: TumorPrint for FHIT.

FHIT is the top candidate in focality score analysis for focal CNV deletions. Homozygous deletions are commonly observed in FHIT in multiple cancer types, especially Esophageal Adenocarcinoma patients. The pie plot represents the *number of patients with FHIT biallelic inactivations in each cohort. The barplot represent the z score of gene expression level of patient within cohort.*

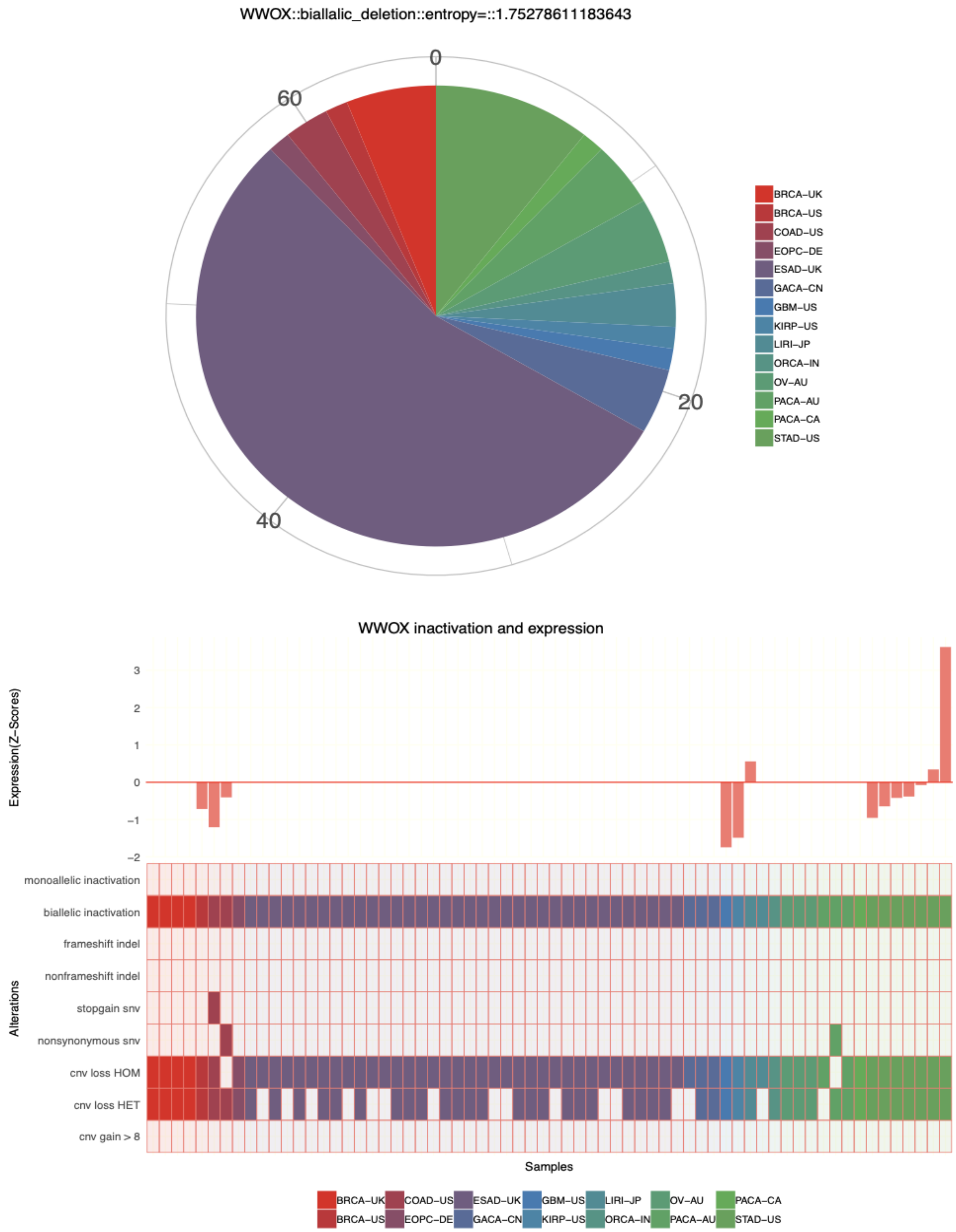


Figure 30: TumorPrint for WWOX.

WWOX is the second candidate in focality score analysis for focal CNV deletions. Similar to FHIT, WWOX is also frequently deleted in both allele in Esophageal Adenocarcinoma. The pie plot represents the number of patients with WWOX *biallelic inactivations in each cohort. The barplot represent the z score of gene expression level of patient within cohort.*

5.1.4.2 Bi-gene TumorPrint

The Bi-gene TumorPrint can be used for show co-mutations or mutual exclusive of a gene pair in cohorts. I used this tool for visualize the synthetic lethality analysis for cancer genes.

In bi-gene TumorPrint, plus represents amplification and minus represents inactivation. In detail:

+++: amplification with copy number greater than 8

++: amplification with copy number greater than 4 and not greater than 8

+: amplification with copy number not greater than 4

- - - : bi-allelic inactivation (homozygous deletion or heterozygous deletion with functional SNV in the same location)

- -: potential bi-allelic inactivation (heterozygous deletion with functional SNV or more than one functional SNVs but not in the same location)

-: mono-allelic inactivation (single heterozygous deletion or functional SNV)

Figure 31 to Figure 35 provide a few examples of bi-gene TumorPrint. The barplots represent the expression z scores of both genes. The calculation of z scores is presented in previous sections. In the alteration panel, each row represents one gene and each color represents one cohort. The shade of color represents the levels of alterations where a darker color stands for more convincing deletion or stronger amplification.

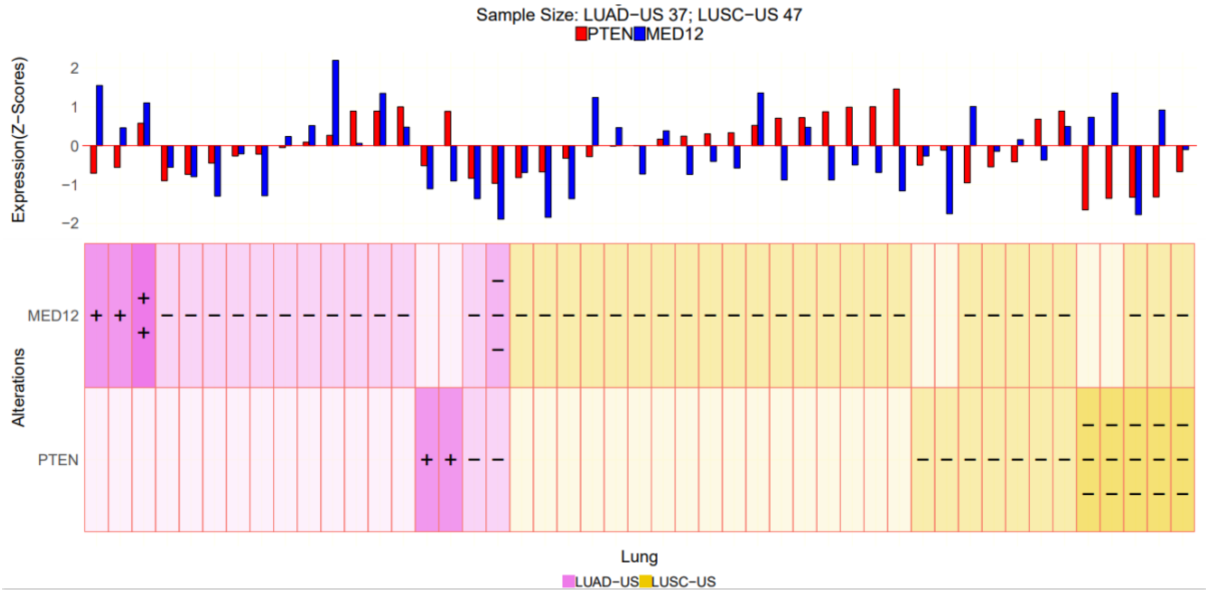


Figure 31: Bi-gene TumorPrint for MED12 and PTEN in lung cancer.

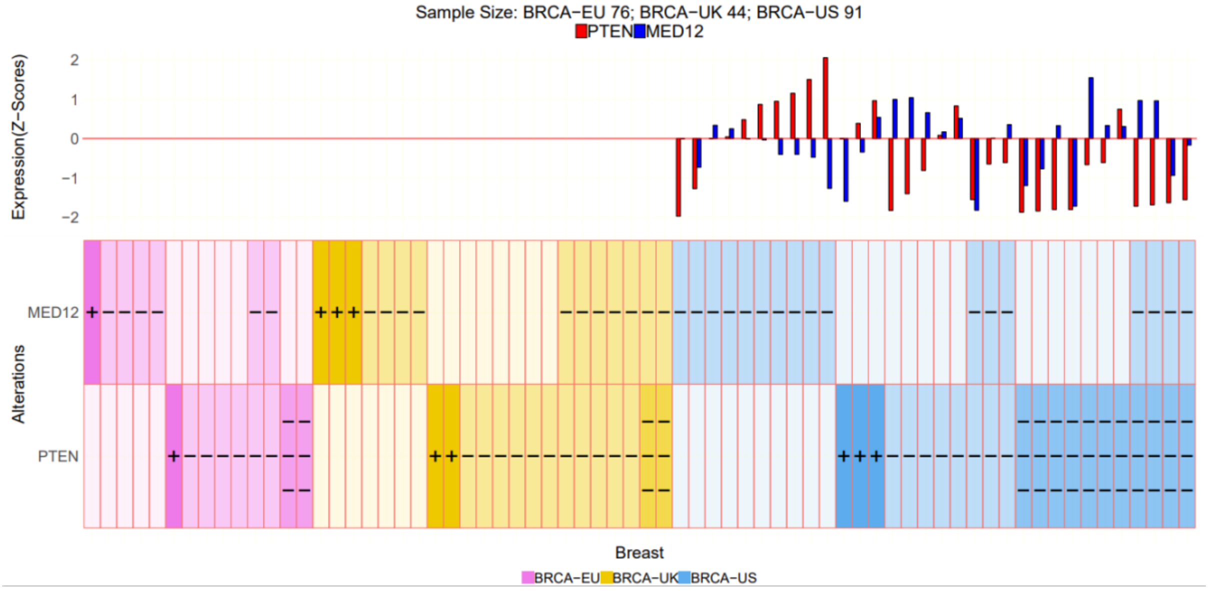


Figure 32: Bi-gene TumorPrint for MED12 and PTEN in breast cancer.

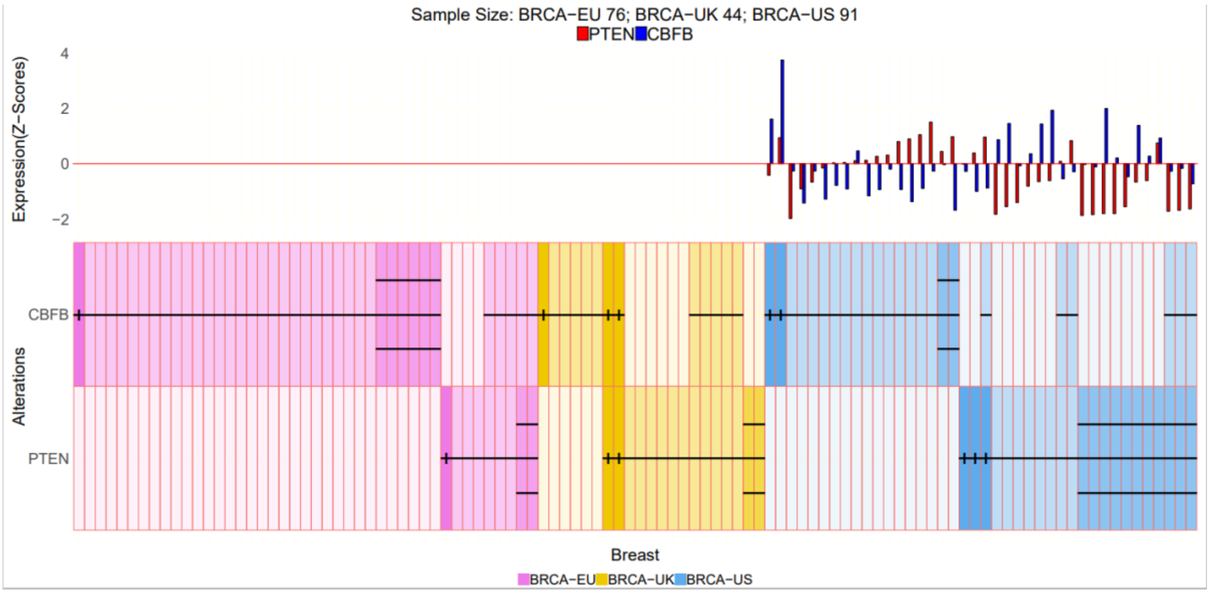


Figure 33: Bi-gene TumorPrint for PTEN and CBFB in breast cancer.

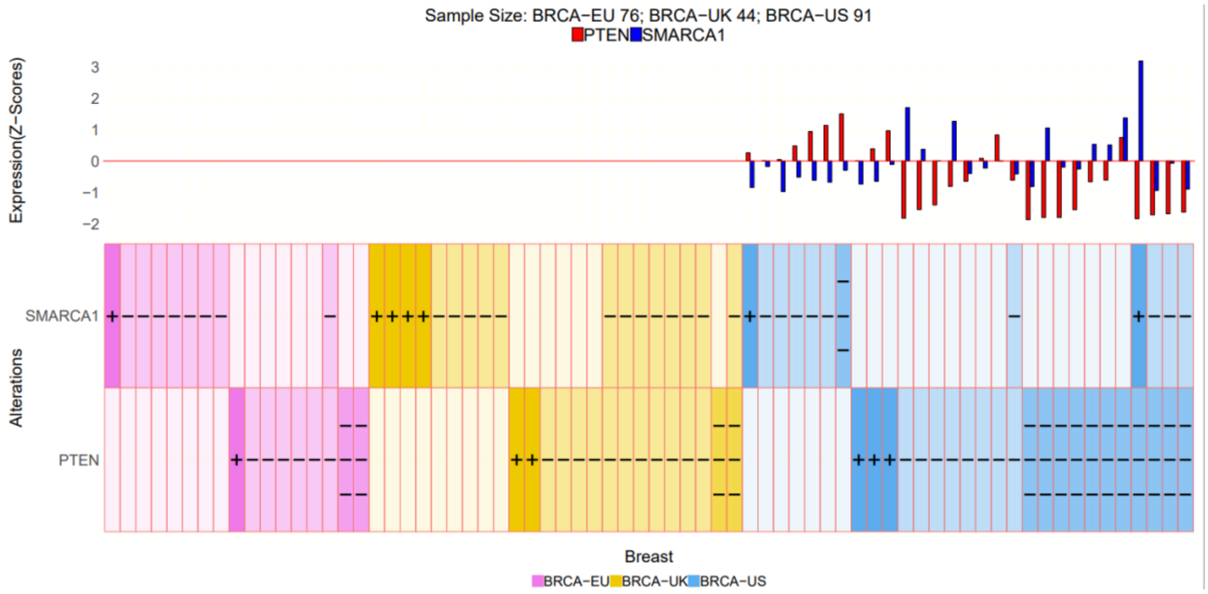


Figure 34: Bi-gene TumorPrint for SMARCA1 and PTEN in breast cancer.

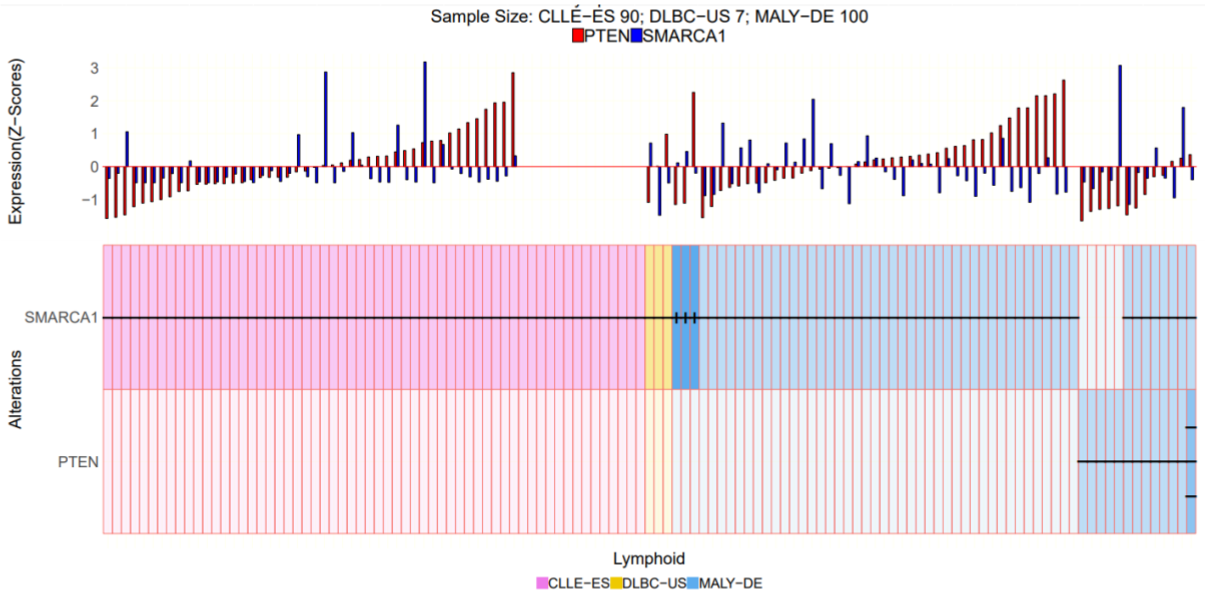


Figure 35: Bi-gene TumorPrint for SMARCA1 and PTEN in lymphoid cancer.

5.1.5 Discussion

TumorPrint is a very useful visualization tool for genome variations and expressions in any genes of interests in large cohort. With this tool, the users can very fast query and understand the overview of alterations in a particular gene.

The bi-gene TumorPrint can compare two genes which provides an convenient way to look into interactions between cancer genes. The phenomenon like co-deletions and mutual exclusivities can be shown in the figures clearly. It can also be used to understand the correlation between genomic variations and expressions.

The TumorPrint is very user-friendly. It works without many complicated parameters and heavy computational load. The wet-lab researchers can easily query interested genes and understand their abbreviations in PCAWG data to support their experiment discoveries or get shortcuts for their work.

5.2 Genome Tornado Plot

5.2.1 Background

Gains-of-function and losses-of-function of genes caused by CNVs play a pivotal role in the process of cell carcinogenesis. (Zhang et al., 2016) In the analysis of cancer genome data, many cancer-related genes and CNV associations have been found. (Zhang et al., 2016) Enrichment analysis of CNVs provides a reliable way to explore cancer-driver genes. However, because the lengths of CNVs vary, many CNV events cover a large number of gene regions and even entire chromosome arms are amplificated or deleted. It is therefore difficult to distinguish convincing cancer driver mutated genes from neighboring genes that are functionally not associated with cancer. The study of focal CNV addresses this issue. In this study, I defined Focal CNV as 1Mb-3Mb according to some publications. (Bierkens et al., 2013; Bignell et al., 2010) Amplifications and especially deletions of focal CNV are associated with many well-known cancer-associated genes, such as PTEN, CDNK2A, and RB1. (Garnis et al., 2006; Leary et al., 2008) These variations are commonly found in many cancer types such as breast, lung, and colon cancers. (Bierkens et al., 2013; Garnis et al., 2006)

In order to be able to quantitatively analyze and visualize the enrichment status of focal CNVs near genes of interest, I designed and implemented an R package "GenomeTornadoPlot". Nowadays, large-scale international research collaborations such as TCGA and ICGC provided massive cohort-level genomic data on multiple genomics analysis. (Tate et al., 2019; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020) "GenomeTornadoPlot" can not only analyze specific genes based on these public databases and provide statistically powerful variation information but also visualize experimental data from users and provide intuitive Focal CNV distribution information. In addition to visualizing single genes and calculating focality scores, this package can also perform a comparative analysis of paired genes on the same chromosome to find the most likely potential candidates of cancer drivers among neighboring genes. (Hong et al., 2022)

5.2.2 Implementation

5.2.2.1 Data Input

The input data of the GenomeTornadoPlot package is CNV information. The standard input format is an extended BED-like text file. The file requires strictly defined column names that include chromosome identifier (Chromosome), start position (Start), end position (End), score (Score), cohort-of-origin (Cohort), and patient ID (PID). The score column is generally used to store ploidy information, but can also be other user-defined data. The input data can not only consist of pre-extracted CNV information that only impact the gene of interest, but the users can also input genome-wide CNV data and the name of the target gene. A gene model embedded in the GenomeTornadoPlot package will automatically find the nearby coordinates according to the gene and then process the data. (Hong et al., 2022)

Users can input CNV data generated by themselves, or directly analyze target genes with ICGC-PCAWG data. The PCWAG data has been uploaded on the GitHub server, including 2976 samples from all 46 cancer cohorts. (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020) <https://github.com/chenhong-dkfz/GenomeTornadoPlot-files>. The data are saved in RData format and have been processed into the format that GenomeTornadoPlot can read directly. Users can download the data for the whole genome or by a specified chromosome. (Hong et al., 2022)

5.2.2.2 Visualization

The name of GenomeTornadoPlot was inspired by the shape of the tornado that occurs when the CNV segments in the cohort are arranged by length next to the chromosome. In the plots, chromosomes are displayed as ideograms, with CNV event spans displayed next to them. By default, CNVs will be sorted by lengths in ascending order. Through different parameter settings, users can also select sorting methods according to cohort, ploidy, or CNV type (amplification or deletion). GenomeTornadoPlot can also adjust the display positions of chromosomes and CNVs to meet different needs. In the "twin plot" mode, for genes with complex functions, the amplifications and deletions can be displayed on both sides of the chromosome ideograms respectively. For neighboring genes located on the same chromosome, CNVs that affect two genes separately will also be displayed on each side of the chromosome. If users want to understand the co-deletion or co-amplification of two genes, they can also set the parameter to "mixed plot" and at this point, all shared CNVs and individually matched CNVs for each gene are displayed separately with different colors. In addition, for relatively short genes or CNVs, the GenomeTornadoPlot package also provides zoom-in and rotation functions, which can display subtle variations more clearly. The entropy value and focality score of CNVs will also be displayed on the graph. GenomeTornadoPlot also allows users to save the output in different file formats. Generated plots can be saved as R objects, JPEG files, or vector graphics. (Hong et al., 2022)

5.2.2.3 Focality Score and Entropy

5.2.2.3 Focality Score and Entropy

In order to be able to quantify and compare the enrichment of focal CNVs between different genes, I defined the “focality score”. This value can provide a reference when screening potential cancer driver genes. “The default focality score is defined as:

$$S = \sum_{i=1}^{m} \quad \left(log(L_{max} - L_i)\right),$$

where *m* is the total number of focal variation events and the capping value L_{max} is the length of the longest event that is defined as focal CNVs in order to exclude large events such as chromosomal arm losses.” (Hong et al., 2022)

In shorter gene regions, focal CNVs may affect adjacent genes. In order to reduce this effect, based on the focality score, I implemented the edge score. “It is defined as

$$score_{edge} = (2*S_{gene} - S_{neighbour\_1} - S_{neighbour\_2})/2$$

where *Neighbour_1* and *Neighbour_2* are neighboring genes of the target gene. If the target gene is at the edge of the chromosome, the only neighbor gene counts as both Neighbor 1 and 2. In short, the edge score is the average difference between the standard focality score of one gene with its neighbors.” (Hong et al., 2022)

Edge score calculates the relative enrichment status of the gene in the local focal CNV by calculating the average value of the focality score difference between the gene of interest and the neighboring genes on both sides. (Hong et al., 2022)

In the current version, GenomeTornadoPlot also provides the possibility for users to define their own focality score according to their requirements. The users can input their self-defined focality score calculated in advance through the score column of the input data. User-defined scores will be displayed on the graph. (Hong et al., 2022)

GenomeTornadoPlot plots were used to process CNV data at the cohort level. Therefore, the distribution of CNV among different cancer sub-types has also become an interesting question. The GenomeTornadoPlot package calculates the Shannon entropy value to quantitatively evaluate the distribution trend of focal CNV events among different cancer cohorts-of-origin. (Hong et al., 2022)

The Shannon entropy is defined as:

$$H = -\sum_{i=1}^{m} \quad p_i log_2 p_i$$

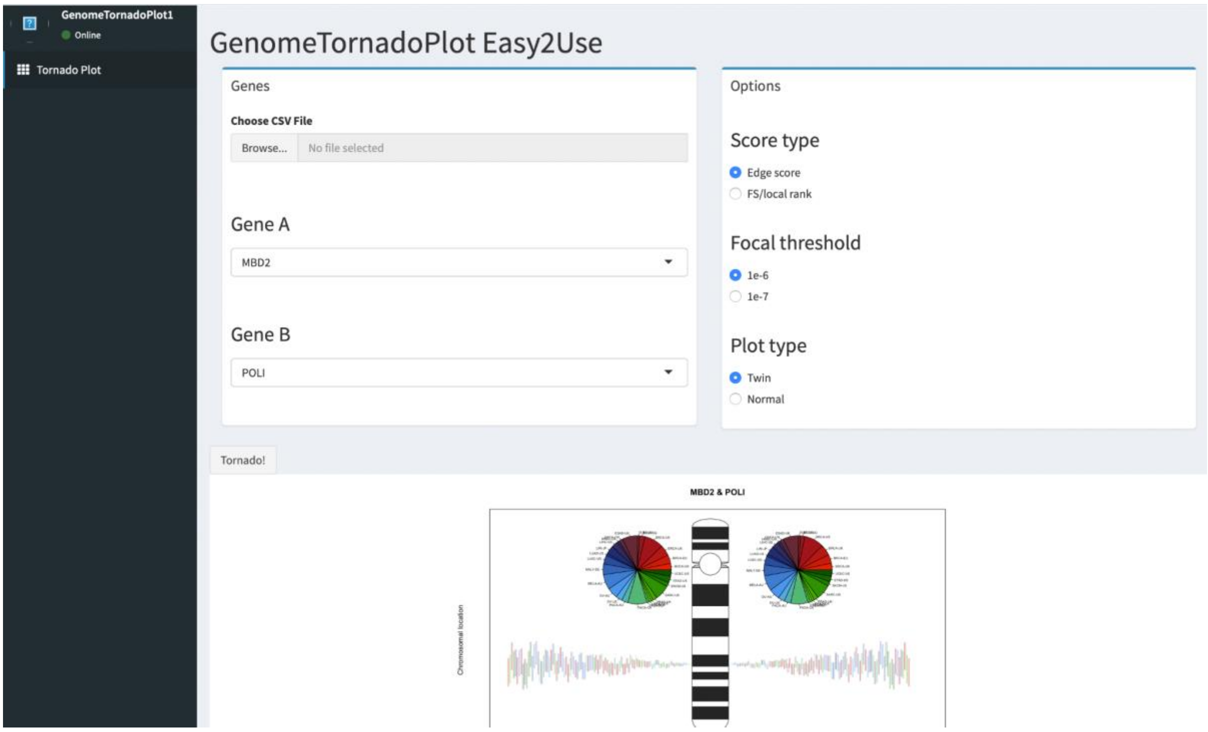
where p_i is the portion of patients of the i-th cohort in total patients from the cohorts.

The closer the Shannon entropy value is to 1, the more CNV events tend to be distributed in a specific cohort. In contrast, a large Shannon entropy value corresponds to an even distribution of CNV events across different patient groups. (Hong et al., 2022)

5.2.2.4 ShinyApp

The GeomeTornadoPlot is an R console-based software so it may not convenient for users who are not familiar with R programming. To solve this problem, the GenomeTornadoPlot package provides a user-friendly ShinyApp graphic user interface. (Figure 36) The users can load the file and set the parameters with a mouse click. The plots and quantitive information will be shown in the app and can be easily exported to JPEG and vector images by clicking the download button.

(Hong et al., 2022)



GenomeTornadoPlot Easy2Use

Genes

Choose CSV File: No file selected

Gene A: MBD2

Gene B: POLI

Options

Score type
● Edge score
○ FS/local rank

Focal threshold
● 1e-6
○ 1e-7

Plot type
● Twin
○ Normal

Figure 36: The shiny user interface of GenomeTornadoPlot.

The parameters including score types, length threshold of focal CNVs and plot types can be given by users in the user-interface. The gene list of Gene A and B will be generated automatically from the input CSV file and user can select the genes of interest from the drop-down menu. Users can click "Tornado!" button and generate the plot. The plot can be downloaded by right click.

5.2.3 Download, Installation, and Usage

5.2.3.1 Download

The GenomeTornadoPlot package is available on https://github.com/chenhong-dkfz/GenomeTornadoPlot, and the data for the test are available on https://github.com/chenhong-dkfz/GenomeTornadoPlot-files.

5.2.3.2 Installation

Prior to installing GenomeTornadoPlot, the following dependencies are required:

Table 16: R requirements for GenomeTornadoPlot

ggplot2	data.table	devtools
gridExtra	tiff	grid
entropy	shiny	shinydashboard
GenomicRanges	quantsmooth	IRanges

The GenomeTornadoPlot can be installed with the following steps

1 In the Git repository click on "Clone or Download".

2 Copy the HTTPS link.

3 Open a terminal and type or paste:

```
> git clone https://github.com/chenhong-dkfz/GenomeTornadoPlot
```

4 Open the folder GenomeTornadoPlot and open the “GenomeTornadoPlot.Rproj” file in RStudio.

5 In the RStudio console, install the package with function:

```
> devtools::install()
```

To successfully install the GenomeTornadoPlot package, R with a version higher than 3.5.0 is necessary.

5.2.3.3 Usage

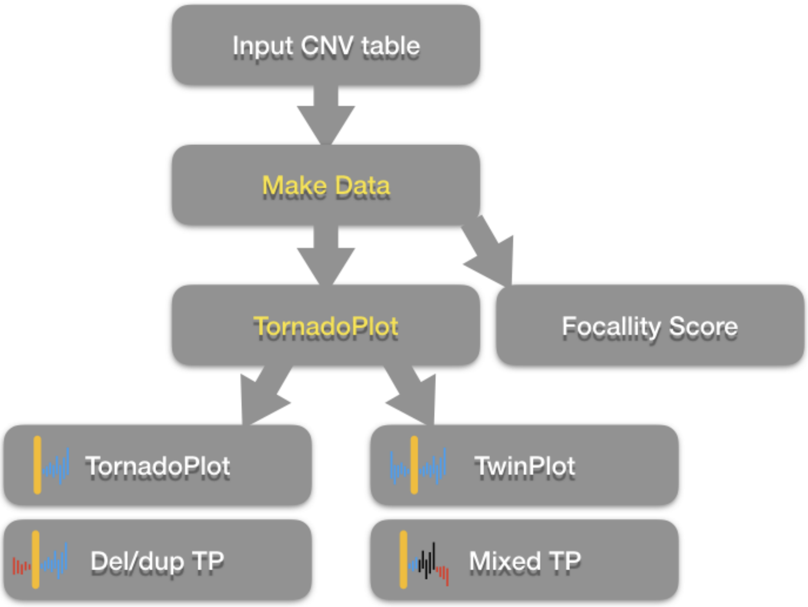


Figure 37: Workflow of the GenomeTornadoPlot package.

The MakeData function generates the intermedia data and calculates the focality score. The TornadoPlot function generates different types of tornado plots.

The workflow of GenomeTornadoPlot is shown in Figure 37. For the input data, users can prepare BED-like data and import it to the R session. (the real PCAWG data for test can be downloaded from https://github.com/chenhong-dkfz/GenomeTornadoPlot-files) In R, it should be a data frame and look like as following:

Table 17: An example of input data format of the GenomeTornadoPlot package.

Chromosome	Start	End	Score	Gene	Cohort	PID
17	6	18318423	3	DOC2B	BLCA-US	0c7aca3f
17	12499	14755572	3	DOC2B	BLCA-US	2b142863
17	827	22199998	1	DOC2B	BLCA-US	301d6ce3
17	6	10573886	3	DOC2B	BLCA-US	418a3dec
17	12499	521774	4	DOC2B	BLCA-US	448fe471
17	833	10272085	1	DOC2B	BLCA-US	8c619cbc
17	1868	5317402	3	DOC2B	BLCA-US	94108975
17	2800	19995288	3	DOC2B	BLCA-US	973d0577
17	833	22199998	1	DOC2B	BLCA-US	acc629cb

The Score column records copy numbers of each CNV event as default. It is important to make sure that column names of the data frame are exactly as in the example. Please pay attention to the first capital letter of each column name because of the case sensitivity of R.

There are two main functions in this package - the MakeData and TornadoPlot functions.

The MakeData function is used to transform the input data frame into an intermediate R objective and calculate entropies and focality scores.

```
> MakeData(CNV, gene_name_1, gene_name_2, score.type, max.length, score.method, cohort_thredshold, gene_score_1, gene_score_2)
```

The parameters are defined as following:

CNV: the input data frame with six columns: Chromosome, Start, End, Score, Gene, Cohort, PID

gene_name_1: the name of the first gene.

gene_name_2: the name of the second gene.

score.type: if the value is "del", calculate focality score of deletions. If the value is "amp", calculate focality score of amplifications.

max.length:if the value is "normal", calculate standard focality score.If the value is "edge", calculate the edge score.

score.method: if the value is "normal", calculate the standard focality score. If the value is "edge", calculate the edge score.

cohort_threshold: the names of cohorts whose event frequencies are below this value in all patients will not be shown in the plot. (default 5%)

gene_score_1: if the value is given by the user, use this input value as focality score of 1st gene in visualization. (optional)

gene_score_2: if the value is given by the user, use this input value as focality score of 2nd gene in visualization. (optional)

The tornadoPlot function generates different types of tornado-shaped plots according to the requirements of users. The input of the function is the R objective which is generated by MakeData function and the output can be either an R list objective that contains plots or graphic files in jpeg, tiff, or eps formats.

```
> TornadoPlots(object, pids, title, legend.type, path, format, color, color.method,
SaveAsObject, multi_panel, orient, zoomed, drop.low.amp, font.size.factor)
```

object: R object generated by MakeData() function.

legend: could be set to “pie”(default) or "barplot" (optional).

color: a vector of CNV colors, optional.

color.method: how to color the CNVs. It could be “cohort”(default) or “ploidy”(optional).

sort.method: how to sort the CNVs. It could be “length”(defult), "cohort" or "ploidy" (optional).

SaveAsObject: if TRUE, returns an rastergrob object. if FALSE the function only saves the plot.

format: if SaveAsObject is FALSE, the packge will save the plots in files. if this value is "tiff", the plot will be saved as a tiff image. if this value is "eps", the plot will be saved as an EPS vector image.

path: if SaveAsObject is FALSE, the packge will save the plots in files. the image will be saved in the path in disk.

multi_panel: if TRUE, a multiple panel plot will be displayed.

zoomed: the value should be "global", "region" or "gene". It indicates how the plot will be zoomed in.

orient: if the value is "v", vertically arranged plots will be displayed. if the value is "h", horizontally arranged plots will be displayed.

drop.low.amp: if the value is TRUE, the amplifications with CN<5 will be not shown in the plots.

font.size.factor: rescale of fonts shown in the plots.

5.2.3.4 Applications and Examples

With the GenomeTornadoPlot packages, the users can visualize and analyze different types of cancer CNVs in patients. I analyzed some genes of interest in PCAWG data. Here I raise a few examples of visualizations in Figure 38-42.

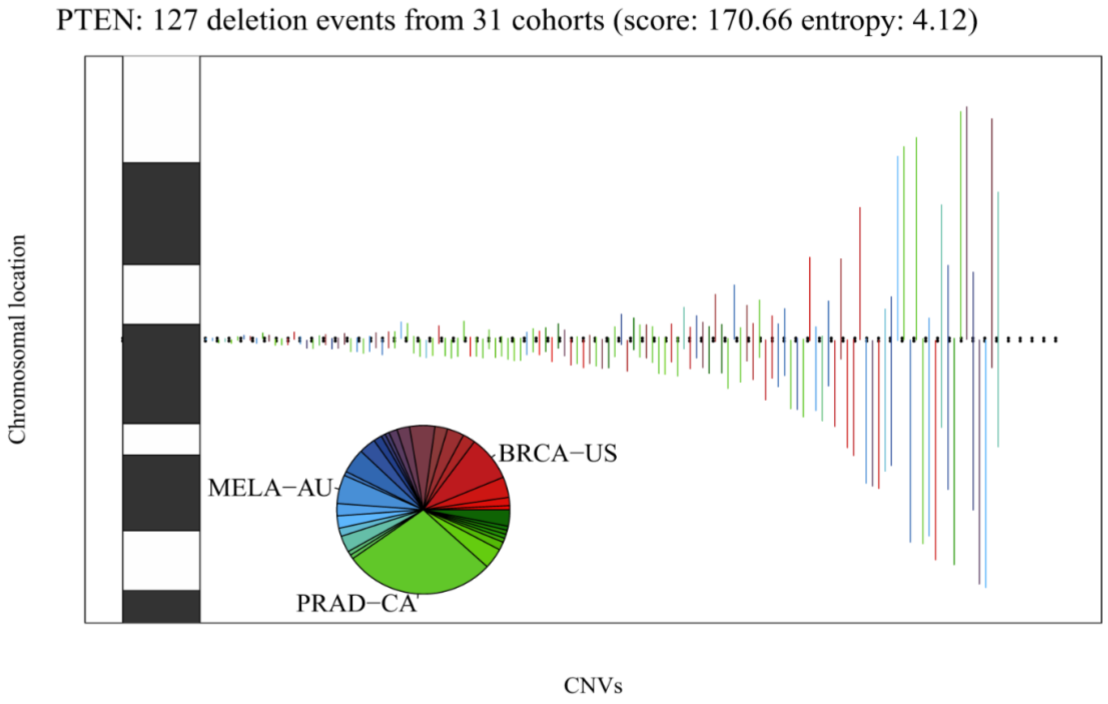


Figure 38: CNVs of PTEN throughout cohorts.

PTEN plays a key role as a tumor suppressor gene in many cancer entities (Chu and Tarnawski 2004). The tornado shape illustrates how the focal deletions are enriched in PTEN locus. The color of events and the respective pie chart shows the origin of these deletions. It is shown in the pie plot, that Breast cancer, prostate cancer, and melanoma are highlighted as cohorts with most focally deleted PTEN. (reprinted from Hong et al., 2022, GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis, Bioinformatics, Volume 38, Issue 7, Pages 2036–2038, 2022*)*

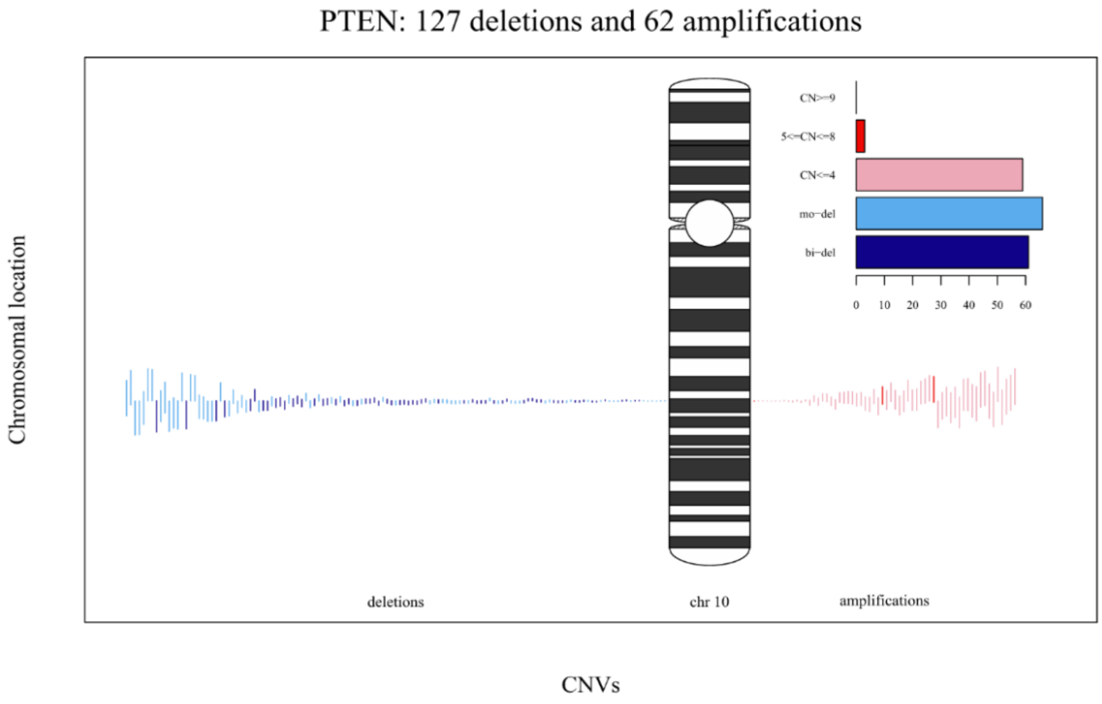


Figure 39: Different levels of deletions and amplifications of PTEN are shown in the tornado plot.

The deletions are in red shades on left side and the amplifications are in blue shades on the right side. As a typical TSG, PTEN has strong focal bi-allelic deletions signals in many cancer patients which is also identified in previous chapters of this thesis. (reprinted with original captions from Hong et al., 2022, GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis, Bioinformatics, Volume 38, Issue 7, Pages 2036–2038, 2022*)*

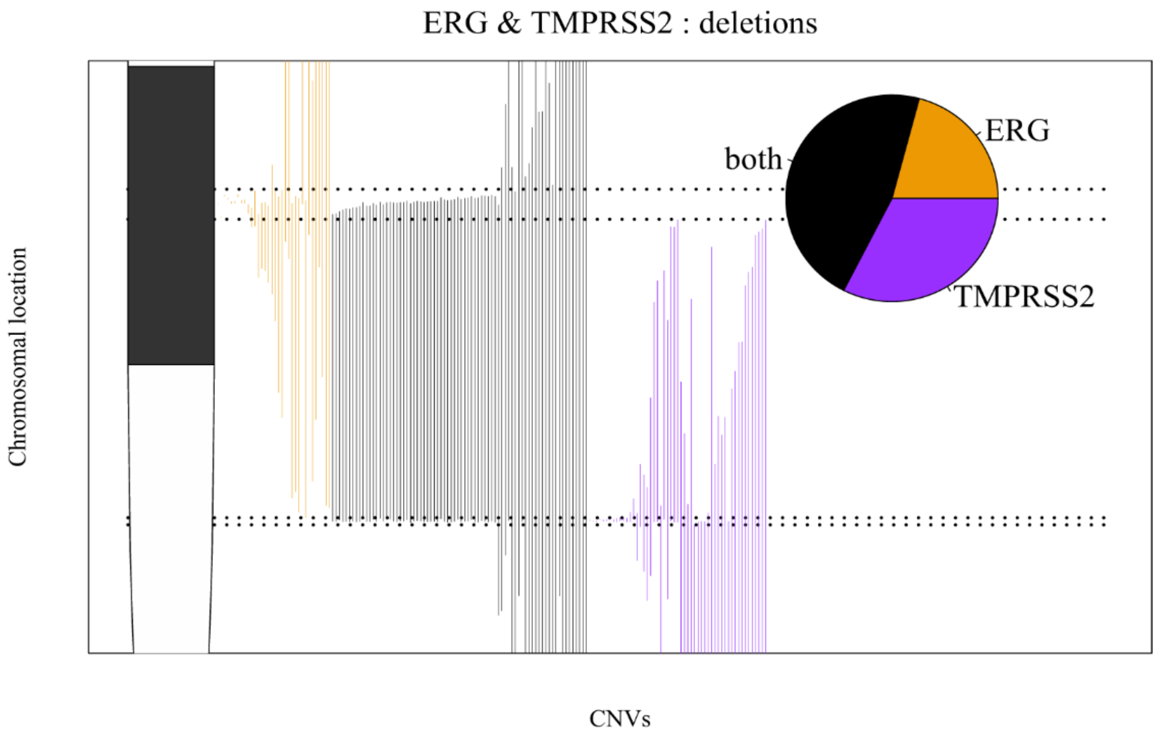


Figure 40: Zoomed-in Mixed Plot Example.

Co-deletion patterns in ERG and TMPRSS2 are shown: ERG is a well-known proto-oncogene. The promoter of TMPRSS2 and the gene body of ERG are located close on chromosome 21 and are frequently fused by the genomic deletion in prostate cancer. (Liu et al., 2001; Weischenfeldt et al., 2013) Gene fusions lead to oncogenic upregulation of ERGs and result in an opportunity for cancer development. In this figure, the co-mutation pattern dominates the single-locus event, implying a co-effect rather than the function of two separate driver mutations. Interestingly, the CNV deletions of ERG and TMPRSS2 are found very similar lengths among the patients. It may have resulted from the positive selection pressure for gain-of-function events. (reprinted with original captions from Hong et al., 2022, GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis, Bioinformatics, Volume 38, Issue 7, Pages 2036–2038, 2022*)*

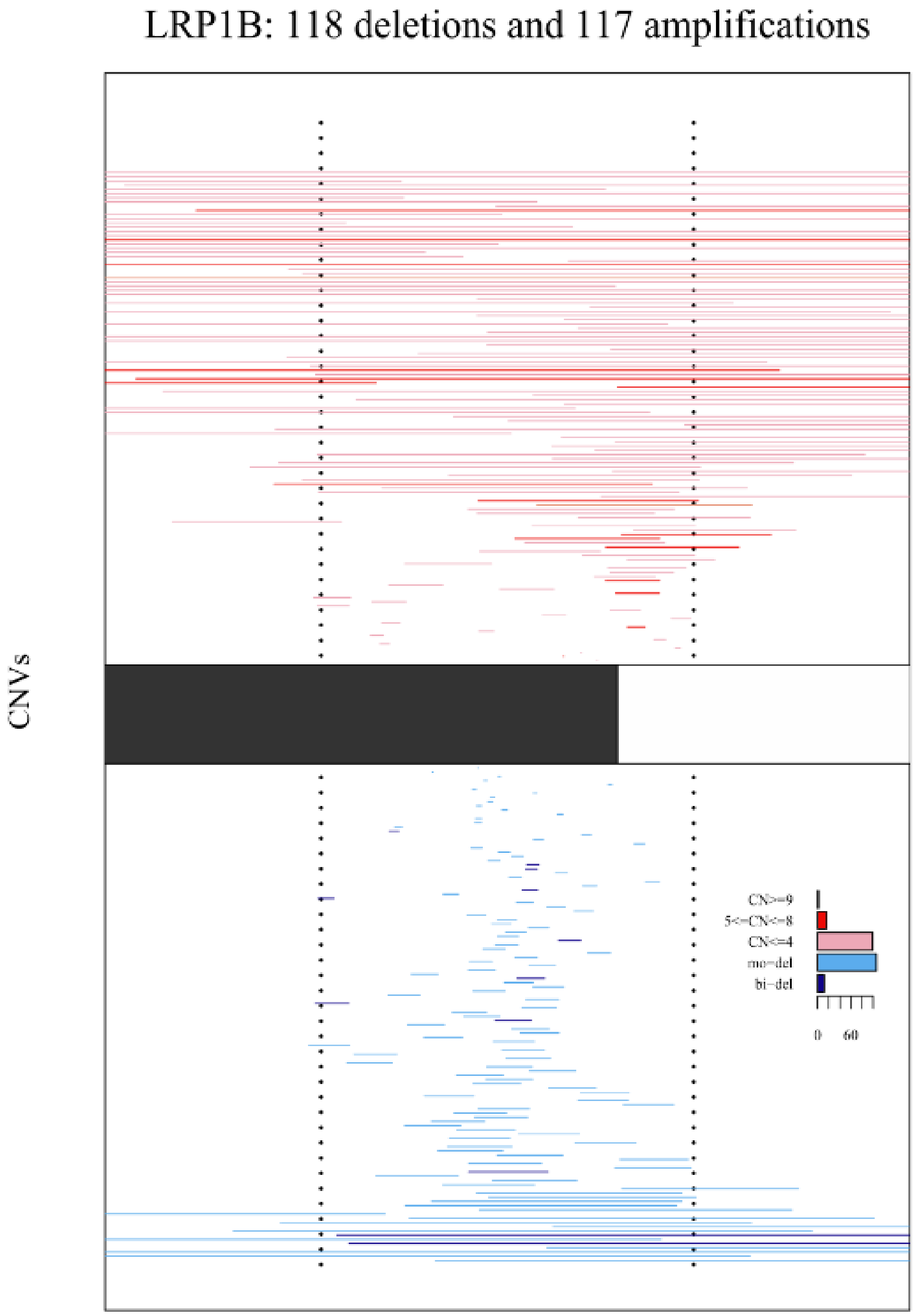


Figure 41: CNVs of LRP1B throughout cohort.

GenomeTornadoPlot provides strong functions for the alignment of the plot. This zoomed-in and vertically arranged plot clearly shows intragenic deletions and amplification in the gene region of LPR1B. Known as a tumor suppressor, LRP1B is functionally related to the clearance of extracellular ligands and signal transduction (Liu et al. 2001). The dashed lines indicate the gene boundaries. (reprinted with original captions from Hong et al., 2022, GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis, Bioinformatics, Volume 38, Issue 7, Pages 2036–2038, 2022*)*

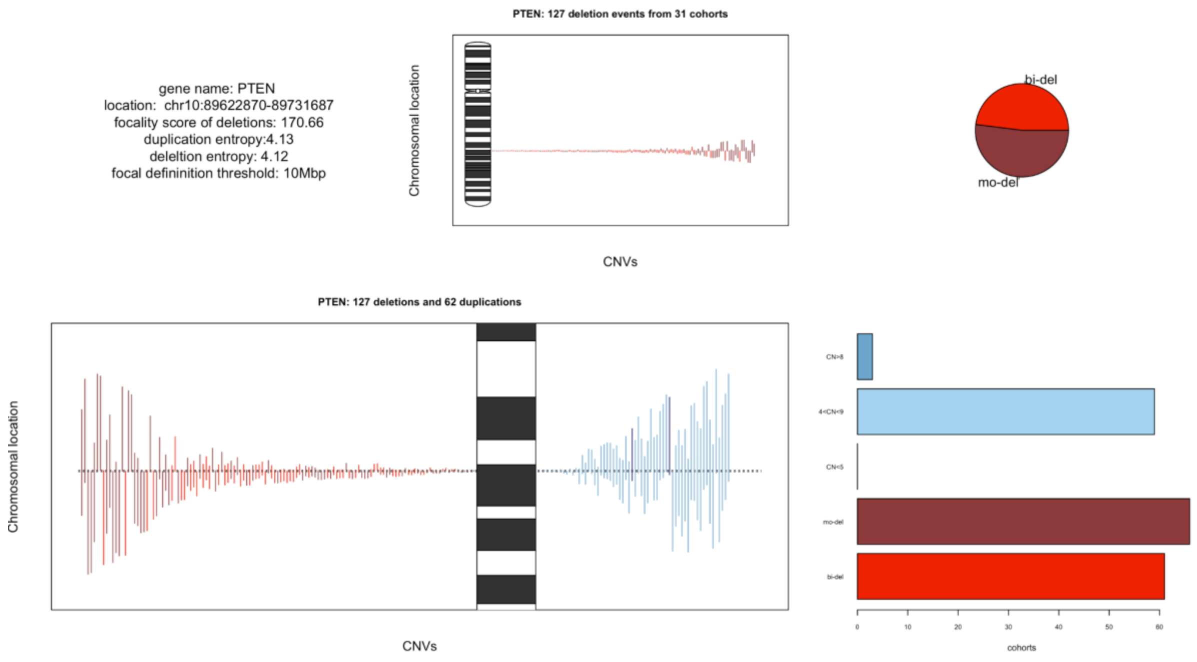


Figure 42: Example of the multi-panel plot.

The complex illustration has five panels. They are the information panel, simple genome tornado plot panel, pie plot panel, twin plot panel, and bar plot panel (in clockwise order). The multi-panel plot provides gene locations, focality scores, duplication/deletion entropies, and focal CNV definitions in the information panel. Tornado-shaped plots, pie plots and bar plots are generated automatically for different tasks.

5.2.3.5 Discussion

The GenomeTornadoPlot is a useful tool for visualizing and quantitative analysis of focal CNVs. It is helpful to identify cancer-related genes and explore the selection-driven accumulation of focal CNVs in cancer.

This tool uses BED-like files as input and can output in different formats to facilitate the different needs of users. Whether it's an R object or a high-definition vector illustration, users can easily set parameters to achieve the goals. In both scientific and aesthetic views, GenomeTornadoPlot is a high-quality R package.

Of course, the current version of GenomeTornadoPlot also has certain limitations. First, since this package is powerful, it requires users to set a lot of parameters. To take care of aesthetics, the users may have to fine-tune certain parameters. Second, due to limited data, the current version of GenomeTornadoPlot only provides basic focality score calculation methods. More sophisticated calculation methods are needed, which beyond the scope of the current study.

6. Conclusion

In this study, I integrated NGS data from more than 2900 patients from 48 different cohorts and implemented a pipeline to convert genomic variation information and transcriptome expression information into a new data structure. Based on this data structure, I tried to answer many questions about cancer from different perspectives.

Firstly, I performed an analysis of bi-allelic inactivation in cancer genomes. Through the statistical analysis, similar to the expected conclusions, I found that the frequency of bi-allelic inactivation was significantly higher in tumor suppressor genes than in other genes. Since bi-allelic inactivation is less susceptible to noise interference than mono-allelic inactivation, I applied bi-allelic inactivation to analysis such as driver gene and synthetic lethality partner screening. I identified some potential tumor suppressor genes that were not listed in the COSMIC Cancer Gene Census database. These interesting genes include CSMD1, WWOX, CCSER1 and MACROD2.

I have also applied the bi-allelic inactivation analysis approach beyond coding genes, such as to lncRNAs. Through bi-allelic inactivation analysis, I provide auxiliary arguments for the establishment of the Cancer LncRNA database. Some LncRNAs are found with more bi-allelic inactivations than control group, such as RP11-624C23.1, RP11-436D23.1, ERICH1-AS1, RP11-317N12.1, and CDKN2B-AS1. Their functions are not clear but RP11-624C23.1, RP11-436D23.1, ERICH1-AS1, RP11-317N12.1, and CDKN2B-AS1 are listed in the CLC. I also analyzed the amplifications of LncRNAs and found the top 10 candidates with the highest amplification rates are mostly involved in the CLC list (PVT, CCAT1, and PCAT1) or cancer susceptibility candidates (CASC11, CASC8, CASC21, CASC19). Although these arguments did not explain in principle why these lncRNAs are indeed responsible for cancer development, significant statistical trends are still obtained that bi-allelic inactivations and amplifications are linked to their experiment results.

In screening for synthetic lethality partners of PTEN, I combined bi-allelic inactivation and hypergeometric distribution tests to screen for these genes as potential drug targets. My colleagues assisted me in gene function annotation to filter unrealistic candidates from my result, and I finally got the following 10 most reliable candidate genes: MED12, BAP1, CBFB, GPR98, WNK3, SMARCA1, ZMYM3, CYSLTR2, NCOR1, and STS. This list of genes will be put into the laboratory for further validation to select the most promising drug targets for PTEN-deficient cancer patients.

In addition to bi-allelic inactivation analysis, I also used statistical methods for functional SNVs to identify the codons and nucleotide positions with high-frequency mutations and tried to find potential cancer driver mutation hotspots in the pan-cancer database. I applied the CaSINo, a statistical method for identification of cancer-related non-coding mutations based on gene mutation frequencies and individual background mutation rates. By analyzing all patient data, I found many potential driver mutation sites, including mutation sites on BRAF(7:140453136), KRAS(12:25398284), JAK2(9:5073770), and PIK3CA(3:178952085), and codons on KRAS:p.G12, BRAF:p.V600, V28, JAK2:p.V617, V468, M545, IDH1:p.R132. Of course, this analysis method has many limitations, the function of many mutation sites is not clear, and the low mutation rate may also lead to a large number of false positives. However, as the first step of mutation hotspot screening, the CaSINo pipeline is undoubtedly fast and interpretive. Like the application of other computational biology methods in the project, the results obtained by the CaSINo still need to be confirmed by further validation such as gene function annotation and laboratory work.

In order to find promoter mutations that alter gene expression at two sites like TERT C228T and C250T, I implemented a pipeline and applied it to the PCAWG data. After filtering through the expression filter, amplification filter, CGC filter, etc., I finally selected specific loci of 551 non-recurrent SNVs, four of the genes have recurrent mutation positions linked to the high expression: TERT, PIM1, BCL2, and EBF1. My colleague Irina Glas conducted a follow-up analysis and finally substituted more accurate results into the laboratory for validation.

In the study of focal CNV, I designed a pipeline to quantitatively analyze the CNV data of PCAWG. In collaboration with Dr. Lars Feuerbach, I designed the pipeline to calculate the focality score. This algorithm can quickly generate the score of how a gene is affected by focal CNVs. Through this algorithm, I not only provided evidence for some known cancer genes, such as FHIT, LRP1B, and PTEN but also potential cancer-related genes such as ARHGEF9, WWOX, and DMD.

Besides the exploration of potential cancer driver genes, I have also implemented a machine learning pipeline and performed it on the PCAWG data for the studies of telomere maintenance mechanisms. In order to be able to distinguish patients with ALT and Telomerase, I designed a random forest-based classifier to classify by the features of TTTGGG, TTCGGG, TGAGGG, TCAGGG, TTGGGG, and breakpoint count, telomere insertion count, and telomere content tumor/control log2 ratio. This classifier had an average area under the curve of 0.96, a sensitivity of 0.72, and a specificity of 0.98. To distinguish ATRX/DAXXtrunc from TERTmod samples, this performance is quite good. Although subsequent studies have shown that ATRX/DAXXtrunc do not fully represent ALT, this method still provides promising ideas for research in this direction.

In the last part of this thesis, based on my understanding of the content and application scenarios of the PCAWG data, I developed two data visualization tools. One of them is TumorPrint. With this tool, users can quickly display various types of mutations and corresponding expression levels of any gene or gene pairs. This tool also includes functions for calculating CaSINo scores and entropies. Users can simultaneously qualitatively and quantitatively observe different levels of variations such as bi-allelic inactivations and functional SNVs in any gene in the overall PCAWG data. Another powerful tool is GenomeTornadoPlot. This R-based package is used to visualize the distribution of focal CNVs on chromosomes. It not only displays the CNV events in cohorts, but also allows users to customize parameters, enabling dual-gene comparisons, image scaling, and multiple panel visualizations. The aforementioned focality scores can also be calculated by this package.

In summary, I designed and implemented several methods to analyze different levels of genomic variations and expression changes and provided a series of answers to the question of potential cancer drivers and telomere maintenance mechanism studies. Meanwhile, I created visualization tools to interpret and display the PCAWG data. Parts of the work are already published. (Carlevaro-Fita et al., 2020; Hong et al., 2022; Sieverling et al., 2020; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020) During the study, I found that more open questions to solve in the future and these methods are still to be improved. For example, the frequency-based cancer genome mutation analysis are still lack in data and golden standard of cancer-related genes. The correlation of SNVs or CNVs and cancers are still not clear. The computational methods of CaSINo and GenomeTornadoPlot can also be improved. More concerns about biological functions should be added into the computational approaches. In future, along with more state-of-art methods performed on data, more insights of gene mutations and cancers will be discovered.

References

A Patil, Pallavi, et al. "Loss of Expression of a Novel Chromatin Remodeler SMARCA1 in Soft Tissue Sarcoma." *Journal of Cytology & Histology*, vol. 09, no. 06, 2018. *DOI.org (Crossref)*, https://doi.org/10.4172/2157-7099.1000524.

Alioto, Tyler S., et al. "A Comprehensive Assessment of Somatic Mutation Detection in Cancer Using Whole-Genome Sequencing." *Nature Communications*, vol. 6, no. 1, Dec. 2015, p. 10001. *DOI.org (Crossref)*, https://doi.org/10.1038/ncomms10001.

Ando, Koichi, et al. "Comparative Efficacy and Safety of Lorlatinib and Alectinib for ALK-Rearrangement Positive Advanced Non-Small Cell Lung Cancer in Asian and Non-Asian Patients: A Systematic Review and Network Meta-Analysis." *Cancers*, vol. 13, no. 15, July 2021, p. 3704. *DOI.org (Crossref)*, https://doi.org/10.3390/cancers13153704.

Aqeilan, Rami I., et al. "Targeted Deletion of *Wwox* Reveals a Tumor Suppressor Function." *Proceedings of the National Academy of Sciences*, vol. 104, no. 10, Mar. 2007, pp. 3949–54. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.0609783104.

Arantes, Lidia Maria Rebolho Batista, et al. "TERT Promoter Mutation C228T Increases Risk for Tumor Recurrence and Death in Head and Neck Cancer Patients." *Frontiers in Oncology*, vol. 10, July 2020, p. 1275. *DOI.org (Crossref)*, https://doi.org/10.3389/fonc.2020.01275.

Arun, Gayatri, et al. "MALAT1 Long Non-Coding RNA: Functional Implications." *Non-Coding RNA*, vol. 6, no. 2, June 2020, p. 22. *DOI.org (Crossref)*, https://doi.org/10.3390/ncrna6020022.

Avrillon, Virginie, and Maurice Pérol. “Alectinib for Treatment of ALK-Positive Non-Small-Cell Lung Cancer.” *Future Oncology (London, England)*, vol. 13, no. 4, Feb. 2017, pp. 321–35. *PubMed*, https://doi.org/10.2217/fon-2016-0386.

Baldeyron, Céline, et al. “TIPIN Depletion Leads to Apoptosis in Breast Cancer Cells.” *Molecular Oncology*, vol. 9, no. 8, Oct. 2015, pp. 1580–98. *DOI.org (Crossref)*, https://doi.org/10.1016/j.molonc.2015.04.010.

Baryła, Izabela, et al. “Alteration of WWOX in Human Cancer, a Clinical View.” *Experimental Biology and Medicine*, vol. 240, no. 3, Mar. 2015, pp. 305–14. *DOI.org (Crossref)*, https://doi.org/10.1177/1535370214561953.

Bawa, Pushpinder Singh, et al. “A Novel Molecular Mechanism for a Long Non-Coding RNA PCAT92 Implicated in Prostate Cancer.” *Oncotarget*, vol. 9, no. 65, Aug. 2018, pp. 32419–34. *DOI.org (Crossref)*, https://doi.org/10.18632/oncotarget.25940.

Bell, Robert J. A., et al. “Understanding TERT Promoter Mutations: A Common Path to Immortality.” *Molecular Cancer Research*, vol. 14, no. 4, Apr. 2016, pp. 315–23. *DOI.org (Crossref)*, https://doi.org/10.1158/1541-7786.MCR-16-0003.

Bernard, Elsa, et al. “Implications of TP53 Allelic State for Genome Stability, Clinical Presentation and Outcomes in Myelodysplastic Syndromes.” *Nature Medicine*, vol. 26, no. 10, Oct. 2020, pp. 1549–56. *DOI.org (Crossref)*, https://doi.org/10.1038/s41591-020-1008-z.

Beroukhim, Rameen, et al. “The Landscape of Somatic Copy-Number Alteration across Human Cancers.” *Nature*, vol. 463, no. 7283, Feb. 2010, pp. 899–905. *DOI.org (Crossref)*, https://doi.org/10.1038/nature08822.

Bierkens, Mariska, et al. “Focal Aberrations Indicate *EYA2* and *Hsa-MiR-375* as Oncogene and Tumor Suppressor in Cervical Carcinogenesis.” *Genes, Chromosomes and Cancer*, vol. 52, no. 1, Jan. 2013, pp. 56–68. *DOI.org (Crossref)*, https://doi.org/10.1002/gcc.22006.

Bignell, Graham R., et al. “Signatures of Mutation and Selection in the Cancer Genome.” *Nature*, vol. 463, no. 7283, Feb. 2010, pp. 893–98. *DOI.org (Crossref)*, https://doi.org/10.1038/nature08768.

Braden, Amy, et al. “Breast Cancer Biomarkers: Risk Assessment, Diagnosis, Prognosis, Prediction of Treatment Efficacy and Toxicity, and Recurrence.” *Current Pharmaceutical Design*, vol. 20, no. 30, Aug. 2014, pp. 4879–98. *DOI.org (Crossref)*, https://doi.org/10.2174/1381612819666131125145517.

Brancaleoni, V., et al. “X-Chromosomal Inactivation Directly Influences the Phenotypic Manifestation of X-Linked Protoporphyria: Variable Heterozygous Expression in X-Linked Protoporphyria.” *Clinical Genetics*, vol. 89, no. 1, Jan. 2016, pp. 20–26. *DOI.org (Crossref)*, https://doi.org/10.1111/cge.12562.

Bratslavsky, Gennady, et al. “Novel Synthetic Lethality (SL) Anti-Cancer Drug Target in Urothelial Bladder Cancer (UCB) Based on *MTAP* Genomic Loss: Incidence and Correlations in Standard of Care (SOC).” *Journal of Clinical Oncology*, vol. 39, no. 6_suppl, Feb. 2021, pp. 485–485. *DOI.org (Crossref)*, https://doi.org/10.1200/JCO.2021.39.6_suppl.485.

Buseman, C. M., et al. “Is Telomerase a Viable Target in Cancer?” *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 730, no. 1–2, Feb. 2012, pp. 90–97. *DOI.org (Crossref)*, https://doi.org/10.1016/j.mrfmmm.2011.07.006.

Calin, George A., and Carlo M. Croce. “Chronic Lymphocytic Leukemia: Interplay between Noncoding RNAs and Protein-Coding Genes.” *Blood*, vol. 114, no. 23, Nov. 2009, pp. 4761–70. *DOI.org (Crossref)*, https://doi.org/10.1182/blood-2009-07-192740.

Capasso, Mario, et al. “Transcription Factors Involved in Tumorigenesis Are Over-Represented in Mutated Active DNA-Binding Sites in Neuroblastoma.” *Cancer Research*, vol. 80, no. 3, Feb. 2020, pp. 382–93. *DOI.org (Crossref)*, https://doi.org/10.1158/0008-5472.CAN-19-2883.

Carlevaro-Fita, Joana, et al. “Cancer LncRNA Census Reveals Evidence for Deep Functional Conservation of Long Noncoding RNAs in Tumorigenesis.” *Communications Biology*, vol. 3, no. 1, Feb. 2020, p. 56. *DOI.org (Crossref)*, https://doi.org/10.1038/s42003-019-0741-7.

Chakraborty, Abhijit, et al. “Knock-down of the TIM/TIPIN Complex Promotes Apoptosis in Melanoma Cells.” *Oncotarget*, vol. 11, no. 20, May 2020, pp. 1846–61. *DOI.org (Crossref)*, https://doi.org/10.18632/oncotarget.27572.

Chang, Hyeyoun, et al. “The Mechanisms Underlying PTEN Loss in Human Tumors Suggest Potential Therapeutic Opportunities.” *Biomolecules*, vol. 9, no. 11, Nov. 2019, p. 713. *DOI.org (Crossref)*, https://doi.org/10.3390/biom9110713.

Chavan, S. S., et al. “Bi-Allelic Inactivation Is More Prevalent at Relapse in Multiple Myeloma, Identifying RB1 as an Independent Prognostic Marker.” *Blood Cancer Journal*, vol. 7, no. 2, Feb. 2017, pp. e535–e535. *DOI.org (Crossref)*, https://doi.org/10.1038/bcj.2017.12.

Chen, Qiongyun, et al. “Long Non-Coding RNA ERICH3-AS1 Is an Unfavorable Prognostic Factor for Gastric Cancer.” *PeerJ*, vol. 8, Jan. 2020, p. e8050. *DOI.org (Crossref)*, https://doi.org/10.7717/peerj.8050.

Cheng, Jiqiu, et al. “Pan-Cancer Analysis of Homozygous Deletions in Primary Tumours Uncovers Rare Tumour Suppressors.” *Nature Communications*, vol. 8, no. 1, Oct. 2017, p. 1221. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-017-01355-0.

Comfort, Nathaniel. “Genetics: We Are the 98%.” *Nature*, vol. 520, no. 7549, Apr. 2015, pp. 615–16. *DOI.org (Crossref)*, https://doi.org/10.1038/520615a.

Conomos, Dimitri, et al. “Variant Repeats Are Interspersed throughout the Telomeres and Recruit Nuclear Receptors in ALT Cells.” *Journal of Cell Biology*, vol. 199, no. 6, Dec. 2012, pp. 893–906. *DOI.org (Crossref)*, https://doi.org/10.1083/jcb.201207189.

Daśko, Mateusz, et al. "Recent Progress in the Development of Steroid Sulphatase Inhibitors – Examples of the Novel and Most Promising Compounds from the Last Decade." *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 35, no. 1, Jan. 2020, pp. 1163–84. *DOI.org (Crossref)*, https://doi.org/10.1080/14756366.2020.1758692.

De Nonneville, Alexandre, and Roger R. Reddel. "Alternative Lengthening of Telomeres Is Not Synonymous with Mutations in ATRX/DAXX." *Nature Communications*, vol. 12, no. 1, Mar. 2021, p. 1552. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-021-21794-0.

Dees, Nathan D., et al. "MuSiC: Identifying Mutational Significance in Cancer Genomes." *Genome Research*, vol. 22, no. 8, Aug. 2012, pp. 1589–98. *DOI.org (Crossref)*, https://doi.org/10.1101/gr.134635.111.

Demir, E., et al. "An Ontology for Collaborative Construction and Analysis of Cellular Pathways." *Bioinformatics*, vol. 20, no. 3, Feb. 2004, pp. 349–56. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btg416.

Ding, Yufeng, et al. "Chromatin Remodeling ATPase BRG1 and PTEN Are Synthetic Lethal in Prostate Cancer." *Journal of Clinical Investigation*, vol. 129, no. 2, Jan. 2019, pp. 759–73. *DOI.org (Crossref)*, https://doi.org/10.1172/JCI123557.

Durkin, Sandra G., and Thomas W. Glover. "Chromosome Fragile Sites." *Annual Review of Genetics*, vol. 41, no. 1, Dec. 2007, pp. 169–92. *DOI.org (Crossref)*, https://doi.org/10.1146/annurev.genet.41.042007.165900.

Elliott, Kerryn, and Erik Larsson. "Non-Coding Driver Mutations in Human Cancer." *Nature Reviews Cancer*, vol. 21, no. 8, Aug. 2021, pp. 500–09. *DOI.org (Crossref)*, https://doi.org/10.1038/s41568-021-00371-z.

"Erratum: Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*, vol. 70, no. 4, July 2020, pp. 313–313. *DOI.org (Crossref)*, https://doi.org/10.3322/caac.21609.

Ertay, Ayse, et al. “WDHD1 Is Essential for the Survival of PTEN-Inactive Triple-Negative Breast Cancer.” *Cell Death & Disease*, vol. 11, no. 11, Nov. 2020, p. 1001. *DOI.org (Crossref)*, https://doi.org/10.1038/s41419-020-03210-5.

Escudero-Esparza, Astrid, et al. “Complement Inhibitor CSMD1 Acts as Tumor Suppressor in Human Breast Cancer.” *Oncotarget*, vol. 7, no. 47, Nov. 2016, pp. 76920–33. *DOI.org (Crossref)*, https://doi.org/10.18632/oncotarget.12729.

Farmer, Hannah, et al. “Targeting the DNA Repair Defect in BRCA Mutant Cells as a Therapeutic Strategy.” *Nature*, vol. 434, no. 7035, Apr. 2005, pp. 917–21. *DOI.org (Crossref)*, https://doi.org/10.1038/nature03445.

Farnham, Peggy J. “Insights from Genomic Profiling of Transcription Factors.” *Nature Reviews Genetics*, vol. 10, no. 9, Sept. 2009, pp. 605–16. *DOI.org (Crossref)*, https://doi.org/10.1038/nrg2636.

Feuerbach, Lars. “Formal Reply to ‘Alternative Lengthening of Telomeres Is Not Synonymous with Mutations in ATRX/DAXX.’” *Nature Communications*, vol. 12, no. 1, Mar. 2021, p. 1551. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-021-21796-y.

Flynt, Erin, et al. “Prognosis, Biology, and Targeting of TP53 Dysregulation in Multiple Myeloma.” *Cells*, vol. 9, no. 2, Jan. 2020, p. 287. *DOI.org (Crossref)*, https://doi.org/10.3390/cells9020287.

Fozzatti, Laura, et al. “Oncogenic Actions of the Nuclear Receptor Corepressor (NCOR1) in a Mouse Model of Thyroid Cancer.” *PLoS ONE*, edited by Moray Campbell, vol. 8, no. 6, June 2013, p. e67954. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pone.0067954.

Fusco, Nicola, et al. “PTEN Alterations and Their Role in Cancer Management: Are We Making Headway on Precision Medicine?” *Genes*, vol. 11, no. 7, June 2020, p. 719. *DOI.org (Crossref)*, https://doi.org/10.3390/genes11070719.

Gallolu Kankanamalage, Sachith, et al. “WNK Pathways in Cancer Signaling Networks.” *Cell Communication and Signaling*, vol. 16, no. 1, Dec. 2018, p. 72. *DOI.org (Crossref)*, https://doi.org/10.1186/s12964-018-0287-1.

Gan, Kok A., et al. “Identification of Single Nucleotide Non-Coding Driver Mutations in Cancer.” *Frontiers in Genetics*, vol. 9, Feb. 2018, p. 16. *DOI.org (Crossref)*, https://doi.org/10.3389/fgene.2018.00016.

Garnis, Cathie, et al. “High Resolution Analysis of Non-Small Cell Lung Cancer Cell Lines by Whole Genome Tiling Path Array CGH.” *International Journal of Cancer*, vol. 118, no. 6, Mar. 2006, pp. 1556–64. *DOI.org (Crossref)*, https://doi.org/10.1002/ijc.21491.

Gaspar, Tiago Bordeira, et al. “Telomere Maintenance Mechanisms in Cancer.” *Genes*, vol. 9, no. 5, May 2018, p. 241. *DOI.org (Crossref)*, https://doi.org/10.3390/genes9050241.

Gioia, Romain, et al. “LncRNAs Downregulated in Childhood Acute Lymphoblastic Leukemia Modulate Apoptosis, Cell Migration, and DNA Damage Response.” *Oncotarget*, vol. 8, no. 46, Oct. 2017, pp. 80645–50. *DOI.org (Crossref)*, https://doi.org/10.18632/oncotarget.20817.

Girgis, Andrew H. *Clear Cell Renal Cell Carcinoma with Biallelic Inactivation of CDKN2A/B on 9p21 Have Distinct Gene Expression Signature and Are Associated with Poor Prognosis*. preprint, Cancer Biology, 29 May 2017. *DOI.org (Crossref)*, https://doi.org/10.1101/143180.

Gonzalez-Perez, Abel, and Nuria Lopez-Bigas. “Functional Impact Bias Reveals Cancer Drivers.” *Nucleic Acids Research*, vol. 40, no. 21, Nov. 2012, pp. e169–e169. *DOI.org (Crossref)*, https://doi.org/10.1093/nar/gks743.

Gordon, Louisa G., et al. “Estimating the Costs of Genomic Sequencing in Cancer Control.” *BMC Health Services Research*, vol. 20, no. 1, Dec. 2020, p. 492. *DOI.org (Crossref)*, https://doi.org/10.1186/s12913-020-05318-y.

Grant, Charles E., et al. "FIMO: Scanning for Occurrences of a given Motif." *Bioinformatics*, vol. 27, no. 7, Apr. 2011, pp. 1017–18. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btr064.

Greider, Carol W., and Elizabeth H. Blackburn. "Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts." *Cell*, vol. 43, no. 2, Dec. 1985, pp. 405–13. *DOI.org (Crossref)*, https://doi.org/10.1016/0092-8674(85)90170-9.

Gu, Zuguang, et al. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." *Bioinformatics*, vol. 32, no. 18, Sept. 2016, pp. 2847–49. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btw313.

Guichard, Cécile, et al. "Integrated Analysis of Somatic Mutations and Focal Copy-Number Changes Identifies Key Genes and Pathways in Hepatocellular Carcinoma." *Nature Genetics*, vol. 44, no. 6, June 2012, pp. 694–98. *DOI.org (Crossref)*, https://doi.org/10.1038/ng.2256.

Guterres, Adam N., and Jessie Villanueva. "Targeting Telomerase for Cancer Therapy." *Oncogene*, vol. 39, no. 36, Sept. 2020, pp. 5811–24. *DOI.org (Crossref)*, https://doi.org/10.1038/s41388-020-01405-w.

Gutschner, Tony, et al. "The Noncoding RNA *MALAT1* Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells." *Cancer Research*, vol. 73, no. 3, Feb. 2013, pp. 1180–89. *DOI.org (Crossref)*, https://doi.org/10.1158/0008-5472.CAN-12-2850.

Halfon, Marc S. "Silencers, Enhancers, and the Multifunctional Regulatory Genome." *Trends in Genetics: TIG*, vol. 36, no. 3, Mar. 2020, pp. 149–51. *PubMed*, https://doi.org/10.1016/j.tig.2019.12.005.

Hamarsheh, Shaima'a, et al. "Immune Modulatory Effects of Oncogenic KRAS in Cancer." *Nature Communications*, vol. 11, no. 1, Oct. 2020, p. 5439. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-020-19288-6.

Hartwell, Leland H., et al. "Integrating Genetic Approaches into the Discovery of Anticancer Drugs." *Science*, vol. 278, no. 5340, Nov. 1997, pp. 1064–68. *DOI.org (Crossref)*, https://doi.org/10.1126/science.278.5340.1064.

Heaphy, Christopher M., Roeland F. De Wilde, et al. "Altered Telomeres in Tumors with *ATRX* and *DAXX* Mutations." *Science*, vol. 333, no. 6041, July 2011, pp. 425–425. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1207313.

Heaphy, Christopher M., Wenya Linda Bi, et al. "Telomere Length Alterations and ATRX/DAXX Loss in Pituitary Adenomas." *Modern Pathology*, vol. 33, no. 8, Aug. 2020, pp. 1475–81. *DOI.org (Crossref)*, https://doi.org/10.1038/s41379-020-0523-2.

Hofree, Matan, et al. "Challenges in Identifying Cancer Genes by Analysis of Exome Sequencing Data." *Nature Communications*, vol. 7, no. 1, July 2016, p. 12096. *DOI.org (Crossref)*, https://doi.org/10.1038/ncomms12096.

Hong, Chen, et al. "GenomeTornadoPlot: A Novel R Package for CNV Visualization and Focality Analysis." *Bioinformatics*, edited by Can Alkan, vol. 38, no. 7, Mar. 2022, pp. 2036–38. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btac037.

---. *Integrative Mutation Analysis of Noncoding Functional Elements in Multiple Cancer Types.* Saarland University, 16 Aug. 2016.

Horn, Susanne, et al. "*TERT* Promoter Mutations in Familial and Sporadic Melanoma." *Science*, vol. 339, no. 6122, Feb. 2013, pp. 959–61. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1230062.

Hou, Helei, et al. "The Role of MDM2 Amplification and Overexpression in Therapeutic Resistance of Malignant Tumors." *Cancer Cell International*, vol. 19, no. 1, Dec. 2019, p. 216. *DOI.org (Crossref)*, https://doi.org/10.1186/s12935-019-0937-4.

Hu, Nan, et al. "Integrative Genomics Analysis of Genes with Biallelic Loss and Its Relation to the Expression of MRNA and Micro-RNA in Esophageal Squamous Cell Carcinoma." *BMC*

Genomics, vol. 16, no. 1, Dec. 2015, p. 732. *DOI.org (Crossref)*, https://doi.org/10.1186/s12864-015-1919-0.

Huang, Alan, et al. “Synthetic Lethality as an Engine for Cancer Drug Target Discovery.” *Nature Reviews Drug Discovery*, vol. 19, no. 1, Jan. 2020, pp. 23–38. *DOI.org (Crossref)*, https://doi.org/10.1038/s41573-019-0046-z.

Huang, Franklin W., et al. “Highly Recurrent *TERT* Promoter Mutations in Human Melanoma.” *Science*, vol. 339, no. 6122, Feb. 2013, pp. 957–59. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1229259.

Ismail, Heba M. S., et al. “Multiple Patterns of *FHIT* Gene Homozygous Deletion in Egyptian Breast Cancer Patients.” *International Journal of Breast Cancer*, vol. 2011, 2011, pp. 1–9. *DOI.org (Crossref)*, https://doi.org/10.4061/2011/325947.

Jafri, Mohammad A., et al. “Roles of Telomeres and Telomerase in Cancer, and Advances in Telomerase-Targeted Therapies.” *Genome Medicine*, vol. 8, no. 1, Dec. 2016, p. 69. *DOI.org (Crossref)*, https://doi.org/10.1186/s13073-016-0324-x.

Jäger, Natalie, et al. “Hypermutation of the Inactive X Chromosome Is a Frequent Event in Cancer.” *Cell*, vol. 155, no. 3, Oct. 2013, pp. 567–81. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cell.2013.09.042.

Jensen, David E., et al. “BAP1: A Novel Ubiquitin Hydrolase Which Binds to the BRCA1 RING Finger and Enhances BRCA1-Mediated Cell Growth Suppression.” *Oncogene*, vol. 16, no. 9, Mar. 1998, pp. 1097–112. *DOI.org (Crossref)*, https://doi.org/10.1038/sj.onc.1201861.

Kim, Taewan, and Carlo M. Croce. “Long Noncoding RNAs: Undeciphered Cellular Codes Encrypting Keys of Colorectal Cancer Pathogenesis.” *Cancer Letters*, vol. 417, Mar. 2018, pp. 89–95. *DOI.org (Crossref)*, https://doi.org/10.1016/j.canlet.2017.12.033.

Knudson, Alfred G. "Mutation and Cancer: Statistical Study of Retinoblastoma." *Proceedings of the National Academy of Sciences*, vol. 68, no. 4, Apr. 1971, pp. 820–23. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.68.4.820.

Kryukov, Gregory V., et al. "*MTAP* Deletion Confers Enhanced Dependency on the PRMT5 Arginine Methyltransferase in Cancer Cells." *Science*, vol. 351, no. 6278, Mar. 2016, pp. 1214–18. *DOI.org (Crossref)*, https://doi.org/10.1126/science.aad5214.

Kumar, Rajesh, et al. "HumCFS: A Database of Fragile Sites in Human Chromosomes." *BMC Genomics*, vol. 19, no. S9, Apr. 2019, p. 985. *DOI.org (Crossref)*, https://doi.org/10.1186/s12864-018-5330-5.

Kumar, Sushant, et al. "Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences." *Cell*, vol. 180, no. 5, Mar. 2020, pp. 915-927.e16. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cell.2020.01.032.

Kurose, K., et al. "Biallelic Inactivating Mutations and an Occult Germline Mutation of PTEN in Primary Cervical Carcinomas." *Genes, Chromosomes & Cancer*, vol. 29, no. 2, Oct. 2000, pp. 166–72.

Lai, Ju-Geng, et al. "Zebrafish WNK Lysine Deficient Protein Kinase 1 (Wnk1) Affects Angiogenesis Associated with VEGF Signaling." *PLoS ONE*, edited by Ramani Ramchandran, vol. 9, no. 8, Aug. 2014, p. e106129. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pone.0106129.

Lancho, Olga, and Daniel Herranz. "The MYC Enhancer-Ome: Long-Range Transcriptional Regulation of MYC in Cancer." *Trends in Cancer*, vol. 4, no. 12, Dec. 2018, pp. 810–22. *DOI.org (Crossref)*, https://doi.org/10.1016/j.trecan.2018.10.003.

Lawrence, Michael S., et al. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature*, vol. 499, no. 7457, July 2013, pp. 214–18. *DOI.org (Crossref)*, https://doi.org/10.1038/nature12213.

Leary, Rebecca J., et al. “Integrated Analysis of Homozygous Deletions, Focal Amplifications, and Sequence Alterations in Breast and Colorectal Cancers.” *Proceedings of the National Academy of Sciences*, vol. 105, no. 42, Oct. 2008, pp. 16224–29. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.0808041105.

Lee, Michael, Mark Hills, et al. “Telomere Extension by Telomerase and ALT Generates Variant Repeats by Mechanistically Distinct Processes.” *Nucleic Acids Research*, vol. 42, no. 3, Feb. 2014, pp. 1733–46. *DOI.org (Crossref)*, https://doi.org/10.1093/nar/gkt1117.

Lee, Michael, Erdahl T. Teber, et al. “Telomere Sequence Content Can Be Used to Determine ALT Activity in Tumours.” *Nucleic Acids Research*, vol. 46, no. 10, June 2018, pp. 4903–18. *DOI.org (Crossref)*, https://doi.org/10.1093/nar/gky297.

Lee, Sung Hak, et al. “Whole-Exome Sequencing Identified Mutational Profiles of High-Grade Colon Adenomas.” *Oncotarget*, vol. 8, no. 4, Jan. 2017, pp. 6579–88. *PubMed*, https://doi.org/10.18632/oncotarget.14172.

Le Tallec, Benoît, et al. “Common Fragile Site Profiling in Epithelial and Erythroid Cells Reveals That Most Recurrent Cancer Deletions Lie in Fragile Sites Hosting Large Genes.” *Cell Reports*, vol. 4, no. 3, Aug. 2013, pp. 420–28. *DOI.org (Crossref)*, https://doi.org/10.1016/j.celrep.2013.07.003.

Leung, Justin W. C., et al. “ZMYM3 Regulates BRCA1 Localization at Damaged Chromatin to Promote DNA Repair.” *Genes & Development*, vol. 31, no. 3, Feb. 2017, pp. 260–74. *DOI.org (Crossref)*, https://doi.org/10.1101/gad.292516.116.

Li, Yanan, et al. “The Emerging Role of ISWI Chromatin Remodeling Complexes in Cancer.” *Journal of Experimental & Clinical Cancer Research*, vol. 40, no. 1, Dec. 2021, p. 346. *DOI.org (Crossref)*, https://doi.org/10.1186/s13046-021-02151-x.

Lin, Anqi, et al. “Effect of NCOR1 Mutations on Immune Microenvironment and Efficacy of Immune Checkpoint Inhibitors in Patient with Bladder Cancer.” *Frontiers in Immunology*, vol. 12, Mar. 2021, p. 630773. *DOI.org (Crossref)*, https://doi.org/10.3389/fimmu.2021.630773.

Liu, Chun-Xiang, et al. "The Putative Tumor Suppressor LRP1B, a Novel Member of the Low Density Lipoprotein (LDL) Receptor Family, Exhibits Both Overlapping and Distinct Properties with the LDL Receptor-Related Protein." *Journal of Biological Chemistry*, vol. 276, no. 31, Aug. 2001, pp. 28889–96. *DOI.org (Crossref)*, https://doi.org/10.1074/jbc.M102727200.

Liu, Tiantian, et al. "The Activating TERT Promoter Mutation C228T Is Recurrent in Subsets of Adrenal Tumors." *Endocrine-Related Cancer*, vol. 21, no. 3, June 2014, pp. 427–34. *DOI.org (Crossref)*, https://doi.org/10.1530/ERC-14-0016.

Liu, Yongbo, et al. "Telomerase Reverse Transcriptase (TERT) Is a Therapeutic Target of Oleanane Triterpenoid CDDO-Me in Prostate Cancer." *Molecules*, vol. 17, no. 12, Dec. 2012, pp. 14795–809. *DOI.org (Crossref)*, https://doi.org/10.3390/molecules171214795.

Luo, Jian, et al. "Long Non-Coding RNAs: A Rising Biotarget in Colorectal Cancer." *Oncotarget*, vol. 8, no. 13, Mar. 2017, pp. 22187–202. *DOI.org (Crossref)*, https://doi.org/10.18632/oncotarget.14728.

Malcikova, Jitka, et al. "Monoallelic and Biallelic Inactivation of TP53 Gene in Chronic Lymphocytic Leukemia: Selection, Impact on Survival, and Response to DNA Damage." *Blood*, vol. 114, no. 26, Dec. 2009, pp. 5307–14. *DOI.org (Crossref)*, https://doi.org/10.1182/blood-2009-07-234708.

Malik, Navdeep, et al. "The Transcription Factor CBFB Suppresses Breast Cancer through Orchestrating Translation and Transcription." *Nature Communications*, vol. 10, no. 1, May 2019, p. 2071. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-019-10102-6.

McNamara, Keely M., et al. "Phase Two Steroid Metabolism and Its Roles in Breast and Prostate Cancer Patients." *Frontiers in Endocrinology*, vol. 4, 2013. *DOI.org (Crossref)*, https://doi.org/10.3389/fendo.2013.00116.

Mello, Ramon, et al. "EGFR and EML4-ALK Updated Therapies in Non-Small Cell Lung Cancer." *Recent Patents on Anti-Cancer Drug Discovery*, vol. 11, no. 4, Nov. 2016, pp. 393–400. *DOI.org (Crossref)*, https://doi.org/10.2174/1574892811666160803090944.

Melnikova, L. S., et al. "The Functions and Mechanisms of Action of Insulators in the Genomes of Higher Eukaryotes." *Acta Naturae*, vol. 12, no. 4, Dec. 2020, pp. 15–33. *DOI.org (Crossref)*, https://doi.org/10.32607/actanaturae.11144.

Milella, Michele, et al. "PTEN: Multiple Functions in Human Malignant Tumors." *Frontiers in Oncology*, vol. 5, Feb. 2015. *DOI.org (Crossref)*, https://doi.org/10.3389/fonc.2015.00024.

Molinari, Francesca, and Milo Frattini. "Functions and Regulation of the PTEN Gene in Colorectal Cancer." *Frontiers in Oncology*, vol. 3, 2014. *DOI.org (Crossref)*, https://doi.org/10.3389/fonc.2013.00326.

Möller, Inga, et al. "Activating Cysteinyl Leukotriene Receptor 2 (CYSLTR2) Mutations in Blue Nevi." *Modern Pathology*, vol. 30, no. 3, Mar. 2017, pp. 350–56. *DOI.org (Crossref)*, https://doi.org/10.1038/modpathol.2016.201.

Moniz, Sónia, and Peter Jordan. "Emerging Roles for WNK Kinases in Cancer." *Cellular and Molecular Life Sciences*, vol. 67, no. 8, Apr. 2010, pp. 1265–76. *DOI.org (Crossref)*, https://doi.org/10.1007/s00018-010-0261-6.

Mora, Antonio, et al. "In the Loop: Promoter–Enhancer Interactions and Bioinformatics." *Briefings in Bioinformatics*, Nov. 2015, p. bbv097. *DOI.org (Crossref)*, https://doi.org/10.1093/bib/bbv097.

Moroi, K., and T. Sato. “Comparison between Procaine and Isocarboxazid Metabolism in Vitro by a Liver Microsomal Amidase-Esterase.” *Biochemical Pharmacology*, vol. 24, no. 16, Aug. 1975, pp. 1517–21. *PubMed*, https://doi.org/10.1016/0006-2952(75)90029-5.

Nell, Rogier J., et al. “Involvement of Mutant and Wild-Type CYSLTR2 in the Development and Progression of Uveal Nevi and Melanoma.” *BMC Cancer*, vol. 21, no. 1, Dec. 2021, p. 164. *DOI.org (Crossref)*, https://doi.org/10.1186/s12885-021-07865-x.

Nijman, Sebastian M. B. “Synthetic Lethality: General Principles, Utility and Detection Using Genetic Screens in Human Cells.” *FEBS Letters*, vol. 585, no. 1, Jan. 2011, pp. 1–6. *DOI.org (Crossref)*, https://doi.org/10.1016/j.febslet.2010.11.024.

Noblejas-López, María Del Mar, et al. “Evaluation of Transcriptionally Regulated Genes Identifies NCOR1 in Hormone Receptor Negative Breast Tumors and Lung Adenocarcinomas as a Potential Tumor Suppressor Gene.” *PLOS ONE*, edited by Irina U. Agoulnik, vol. 13, no. 11, Nov. 2018, p. e0207776. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pone.0207776.

O’Hayre, Morgan, et al. “The Emerging Mutational Landscape of G Proteins and G-Protein-Coupled Receptors in Cancer.” *Nature Reviews Cancer*, vol. 13, no. 6, June 2013, pp. 412–24. *DOI.org (Crossref)*, https://doi.org/10.1038/nrc3521.

Olivier, M., et al. “TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use.” *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 1, Jan. 2010, pp. a001008–a001008. *DOI.org (Crossref)*, https://doi.org/10.1101/cshperspect.a001008.

Patel, Kalpesh, et al. “FAM190A Deficiency Creates a Cell Division Defect.” *The American Journal of Pathology*, vol. 183, no. 1, July 2013, pp. 296–303. *DOI.org (Crossref)*, https://doi.org/10.1016/j.ajpath.2013.03.020.

Plöthner, Marika, et al. “Cost Analysis of Whole Genome Sequencing in German Clinical Practice.” *The European Journal of Health Economics*, vol. 18, no. 5, June 2017, pp. 623–33. *DOI.org (Crossref)*, https://doi.org/10.1007/s10198-016-0815-0.

Pompili, Luca, et al. “Diagnosis and Treatment of ALT Tumors: Is Trabectedin a New Therapeutic Option?” *Journal of Experimental & Clinical Cancer Research*, vol. 36, no. 1, Dec. 2017, p. 189. *DOI.org (Crossref)*, https://doi.org/10.1186/s13046-017-0657-3.

Pospiech, Karolina, et al. “WWOX Tumor Suppressor Gene in Breast Cancer, a Historical Perspective and Future Directions.” *Frontiers in Oncology*, vol. 8, Aug. 2018, p. 345. *DOI.org (Crossref)*, https://doi.org/10.3389/fonc.2018.00345.

Powter, Branka, et al. “Human TERT Promoter Mutations as a Prognostic Biomarker in Glioma.” *Journal of Cancer Research and Clinical Oncology*, vol. 147, no. 4, Apr. 2021, pp. 1007–17. *DOI.org (Crossref)*, https://doi.org/10.1007/s00432-021-03536-3.

Quddus, MuhammadBilal, et al. “Chromosomal Aberrations in Renal Cell Carcinoma: An Overview with Implications for Clinical Practice.” *Urology Annals*, vol. 11, no. 1, 2019, p. 6. *DOI.org (Crossref)*, https://doi.org/10.4103/UA.UA_32_18.

Rahman, Nazneen, and Richard H. Scott. “Cancer Genes Associated with Phenotypes in Monoallelic and Biallelic Mutation Carriers: New Lessons from Old Players.” *Human Molecular Genetics*, vol. 16, no. R1, Apr. 2007, pp. R60–66. *DOI.org (Crossref)*, https://doi.org/10.1093/hmg/ddm026.

Ramnarine, Varune Rohan, et al. “The Evolution of Long Noncoding RNA Acceptance in Prostate Cancer Initiation, Progression, and Its Clinical Utility in Disease Management.” *European Urology*, vol. 76, no. 5, Nov. 2019, pp. 546–59. *DOI.org (Crossref)*, https://doi.org/10.1016/j.eururo.2019.07.040.

Recagni, Marta, et al. “The Role of Alternative Lengthening of Telomeres Mechanism in Cancer: Translational and Therapeutic Implications.” *Cancers*, vol. 12, no. 4, Apr. 2020, p. 949. *DOI.org (Crossref)*, https://doi.org/10.3390/cancers12040949.

Reddy, E. Premkumar, et al. “A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene.” *Nature*, vol. 300, no. 5888, Nov. 1982, pp. 149–52. *DOI.org (Crossref)*, https://doi.org/10.1038/300149a0.

Reid, Alison H. M., et al. “Novel, Gross Chromosomal Alterations Involving PTEN Cooperate with Allelic Loss in Prostate Cancer.” *Modern Pathology*, vol. 25, no. 6, June 2012, pp. 902–10. *DOI.org (Crossref)*, https://doi.org/10.1038/modpathol.2011.207.

Reimand, Jüri, and Gary D. Bader. “Systematic Analysis of Somatic Mutations in Phosphorylation Signaling Predicts Novel Cancer Drivers.” *Molecular Systems Biology*, vol. 9, no. 1, Jan. 2013, p. 637. *DOI.org (Crossref)*, https://doi.org/10.1038/msb.2012.68.

Review of Breast Cancer Pathologigcal Image Processing.

Rheinbay, Esther, et al. “Recurrent and Functional Regulatory Mutations in Breast Cancer.” *Nature*, vol. 547, no. 7661, July 2017, pp. 55–60. *DOI.org (Crossref)*, https://doi.org/10.1038/nature22992.

Roy, Ananda L., and Dinah S. Singer. “Core Promoters in Transcription: Old Problem, New Insights.” *Trends in Biochemical Sciences*, vol. 40, no. 3, Mar. 2015, pp. 165–71. *DOI.org (Crossref)*, https://doi.org/10.1016/j.tibs.2015.01.007.

Sabarinathan, Radhakrishnan, et al. *The Whole-Genome Panorama of Cancer Drivers*. preprint, Cancer Biology, 20 Sept. 2017. *DOI.org (Crossref)*, https://doi.org/10.1101/190330.

Shimizu, Yasuomi, et al. “Steroid Sulfatase Promotes Invasion through Epithelial-mesenchymal Transition and Predicts the Progression of Bladder Cancer.” *Experimental and Therapeutic Medicine*, Sept. 2018. *DOI.org (Crossref)*, https://doi.org/10.3892/etm.2018.6787.

Sieverling, Lina, et al. “Genomic Footprints of Activated Telomere Maintenance Mechanisms in Cancer.” *Nature Communications*, vol. 11, no. 1, Feb. 2020, p. 733. *DOI.org (Crossref)*, https://doi.org/10.1038/s41467-019-13824-9.

Smith, Ngaio C., and Jacqueline M. Matthews. “Mechanisms of DNA-Binding Specificity and Functional Gene Regulation by Transcription Factors.” *Current Opinion in Structural Biology*, vol. 38, June 2016, pp. 68–74. *DOI.org (Crossref)*, https://doi.org/10.1016/j.sbi.2016.05.006.

Solimini, Nicole L., et al. “Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential.” *Science*, vol. 337, no. 6090, July 2012, pp. 104–09. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1219580.

Sondka, Zbyslaw, et al. “The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers.” *Nature Reviews Cancer*, vol. 18, no. 11, Nov. 2018, pp. 696–705. *DOI.org (Crossref)*, https://doi.org/10.1038/s41568-018-0060-1.

Speck, Nancy A. “Core Binding Factor and Its Role in Normal Hematopoietic Development:” *Current Opinion in Hematology*, vol. 8, no. 4, July 2001, pp. 192–96. *DOI.org (Crossref)*, https://doi.org/10.1097/00062752-200107000-00002.

Sriram, Krishna, Cristina Salmerón, et al. “GPCRs in Pancreatic Adenocarcinoma: Contributors to Tumour Biology and Novel Therapeutic Targets.” *British Journal of Pharmacology*, vol. 177, no. 11, June 2020, pp. 2434–55. *DOI.org (Crossref)*, https://doi.org/10.1111/bph.15028.

Sriram, Krishna, Kevin Moyung, et al. “GPCRs Show Widespread Differential MRNA Expression and Frequent Mutation and Copy Number Variation in Solid Tumors.” *PLOS Biology*, edited by Nancy Hynes, vol. 17, no. 11, Nov. 2019, p. e3000434. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pbio.3000434.

Tamborero, David, et al. “OncodriveCLUST: Exploiting the Positional Clustering of Somatic Mutations to Identify Cancer Genes.” *Bioinformatics*, vol. 29, no. 18, Sept. 2013, pp. 2238–44. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btt395.

Tang, Lu, et al. “NCOR1 May Be a Potential Biomarker of a Novel Molecular Subtype of Prostate Cancer.” *FEBS Open Bio*, vol. 10, no. 12, Dec. 2020, pp. 2678–86. *DOI.org (Crossref)*, https://doi.org/10.1002/2211-5463.13004.

Tang, Qingfeng, et al. “Three Circulating Long Non-Coding RNAs Act as Biomarkers for Predicting NSCLC.” *Cellular Physiology and Biochemistry*, vol. 37, no. 3, 2015, pp. 1002–09. *DOI.org (Crossref)*, https://doi.org/10.1159/000430226.

Tate, John G., et al. “COSMIC: The Catalogue Of Somatic Mutations In Cancer.” *Nucleic Acids Research*, vol. 47, no. D1, Jan. 2019, pp. D941–47. *PubMed*, https://doi.org/10.1093/nar/gky1015.

Thanendrarajan, Sharmilan, et al. “The Level of Deletion 17p and Bi-Allelic Inactivation of *TP53* Has a Significant Impact on Clinical Outcome in Multiple Myeloma.” *Haematologica*, vol. 102, no. 9, Sept. 2017, pp. e364–67. *DOI.org (Crossref)*, https://doi.org/10.3324/haematol.2017.168872.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, et al. “Pan-Cancer Analysis of Whole Genomes.” *Nature*, vol. 578, no. 7793, Feb. 2020, pp. 82–93. *DOI.org (Crossref)*, https://doi.org/10.1038/s41586-020-1969-6.

Tokheim, Collin J., et al. “Evaluating the Evaluation of Cancer Driver Genes.” *Proceedings of the National Academy of Sciences*, vol. 113, no. 50, Dec. 2016, pp. 14330–35. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.1616440113.

Tomasini, Pascale, et al. “Alectinib in the Treatment of ALK-Positive Metastatic Non-Small Cell Lung Cancer: Clinical Trial Evidence and Experience with a Focus on Brain Metastases.”

Therapeutic Advances in Respiratory Disease, vol. 13, Jan. 2019, p. 175346661983190. *DOI.org (Crossref)*, https://doi.org/10.1177/1753466619831906.

Tomczak, Katarzyna, et al. "Review The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge." *Współczesna Onkologia*, vol. 1A, 2015, pp. 68–77. *DOI.org (Crossref)*, https://doi.org/10.5114/wo.2014.47136.

Tseng, Chia-Chun, et al. "Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases." *International Journal of Molecular Sciences*, vol. 22, no. 8, Apr. 2021, p. 4187. *DOI.org (Crossref)*, https://doi.org/10.3390/ijms22084187.

Van Gent, Dik C., et al. "Chromosomal Stability and the DNA Double-Stranded Break Connection." *Nature Reviews Genetics*, vol. 2, no. 3, Mar. 2001, pp. 196–206. *DOI.org (Crossref)*, https://doi.org/10.1038/35056049.

Varley, Helen, et al. "Molecular Characterization of Inter-Telomere and Intra-Telomere Mutations in Human ALT Cells." *Nature Genetics*, vol. 30, no. 3, Mar. 2002, pp. 301–05. *DOI.org (Crossref)*, https://doi.org/10.1038/ng834.

Veigl, Martina L., et al. "Biallelic Inactivation of *HMLH* 1 by Epigenetic Gene Silencing, a Novel Mechanism Causing Human MSI Cancers." *Proceedings of the National Academy of Sciences*, vol. 95, no. 15, July 1998, pp. 8698–702. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.95.15.8698.

Vorontsov, Ilya E., et al. "Negative Selection Maintains Transcription Factor Binding Motifs in Human Cancer." *BMC Genomics*, vol. 17, no. S2, June 2016, p. 395. *DOI.org (Crossref)*, https://doi.org/10.1186/s12864-016-2728-9.

Wang, Wen-Jie, et al. "Long Non-Coding RNA CASC19 Is Associated with the Progression and Prognosis of Advanced Gastric Cancer." *Aging*, vol. 11, no. 15, Aug. 2019, pp. 5829–47. *DOI.org (Crossref)*, https://doi.org/10.18632/aging.102190.

Wang, Xing, et al. “CSMD1 Suppresses Cancer Progression by Inhibiting Proliferation, Epithelial-Mesenchymal Transition, Chemotherapy-Resistance and Inducing Immunosuppression in Esophageal Squamous Cell Carcinoma.” *Experimental Cell Research*, vol. 417, no. 2, Aug. 2022, p. 113220. *DOI.org (Crossref)*, https://doi.org/10.1016/j.yexcr.2022.113220.

Wang, Yingyi, et al. “High Cancer Susceptibility Candidate 8 Expression Is Associated With Poor Prognosis of Pancreatic Adenocarcinoma: Validated Analysis Based on Four Cancer Databases.” *Frontiers in Cell and Developmental Biology*, vol. 8, June 2020, p. 392. *DOI.org (Crossref)*, https://doi.org/10.3389/fcell.2020.00392.

Weischenfeldt, Joachim, et al. “Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer.” *Cancer Cell*, vol. 23, no. 2, Feb. 2013, pp. 159–70. *DOI.org (Crossref)*, https://doi.org/10.1016/j.ccr.2013.01.002.

Wu, Siqi, et al. “A Novel Micropeptide Encoded by Y-Linked LINC00278 Links Cigarette Smoking and AR Signaling in Male Esophageal Squamous Cell Carcinoma.” *Cancer Research*, vol. 80, no. 13, July 2020, pp. 2790–803. *DOI.org (Crossref)*, https://doi.org/10.1158/0008-5472.CAN-19-3440.

Yang, Qifeng, et al. “Two-Hit Inactivation of FHIT by Loss of Heterozygosity and Hypermethylation in Breast Cancer.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, vol. 8, no. 9, Sept. 2002, pp. 2890–93.

Yao, Kunhou, et al. “Correlation Between CASC8, SMAD7 Polymorphisms and the Susceptibility to Colorectal Cancer: An Updated Meta-Analysis Based on GWAS Results.” *Medicine*, vol. 94, no. 46, Nov. 2015, p. e1884. *DOI.org (Crossref)*, https://doi.org/10.1097/MD.0000000000001884.

Žaja, R., et al. “Comparative Analysis of MACROD1, MACROD2 and TARG1 Expression, Localisation and Interactome.” *Scientific Reports*, vol. 10, no. 1, May 2020, p. 8286. *DOI.org (Crossref)*, https://doi.org/10.1038/s41598-020-64623-y.

Zeng, Hanlin, et al. “Bi-Allelic Loss of CDKN2A Initiates Melanoma Invasion via BRN2 Activation.” *Cancer Cell*, vol. 34, no. 1, July 2018, pp. 56-68.e9. *DOI.org (Crossref)*, https://doi.org/10.1016/j.ccell.2018.05.014.

Zhang, Feng, et al. “Copy Number Variation in Human Health, Disease, and Evolution.” *Annual Review of Genomics and Human Genetics*, vol. 10, no. 1, Sept. 2009, pp. 451–81. *DOI.org (Crossref)*, https://doi.org/10.1146/annurev.genom.9.081307.164217.

Zhang, Jia-Min, and Lee Zou. “Alternative Lengthening of Telomeres: From Molecular Mechanisms to Therapeutic Outlooks.” *Cell & Bioscience*, vol. 10, no. 1, Dec. 2020, p. 30. *DOI.org (Crossref)*, https://doi.org/10.1186/s13578-020-00391-6.

Zhang, Liangcai, et al. “Identification of Recurrent Focal Copy Number Variations and Their Putative Targeted Driver Genes in Ovarian Cancer.” *BMC Bioinformatics*, vol. 17, no. 1, Dec. 2016, p. 222. *DOI.org (Crossref)*, https://doi.org/10.1186/s12859-016-1085-7.

Zhang, Shengjie, et al. “The Emerging Role of Mediator Complex Subunit 12 in Tumorigenesis and Response to Chemotherapeutics.” *Cancer*, vol. 126, no. 5, Mar. 2020, pp. 939–48. *DOI.org (Crossref)*, https://doi.org/10.1002/cncr.32672.

Zhang, X., et al. “Telomere Shortening and Apoptosis in Telomerase-Inhibited Human Tumor Cells.” *Genes & Development*, vol. 13, no. 18, Sept. 1999, pp. 2388–99. *DOI.org (Crossref)*, https://doi.org/10.1101/gad.13.18.2388.

Zhang, Xiaoyang, and Matthew Meyerson. “Illuminating the Noncoding Genome in Cancer.” *Nature Cancer*, vol. 1, no. 9, Sept. 2020, pp. 864–72. *DOI.org (Crossref)*, https://doi.org/10.1038/s43018-020-00114-3.

Publication List

Rheinbay, E., Nielsen, M.M., Abascal, F. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 578, 102–111 (2020). https://doi.org/10.1038/s41586-020-1965-x

Sieverling, L., Hong, C., Koser, S.D. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. Nat Commun 11, 733 (2020). https://doi.org/10.1038/s41467-019-13824-9

Carlevaro-Fita, J., Lanzós, A., Feuerbach, L. et al. Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun Biol 3, 56 (2020). https://doi.org/10.1038/s42003-019-0741-7

Chen Hong, Robin Thiele, Lars Feuerbach, GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis, Bioinformatics, Volume 38, Issue 7, 1 April 2022, Pages 2036–2038, https://doi.org/10.1093/bioinformatics/btac037

Declaration of own contribution

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgments.
I further specify this by using the pronoun ‘we’ when others are substantially involved in the work and ‘I’ for those parts that are purely my own work.
I further state that no substantial part of my dissertation has already been submitted or is being concurrently submitted for any such degree, diploma or other qualification at the Heidelberg University or any other University or similar institution.
For the first paper, I have been responsible for data processing, SNV/CNV data analysis and bi-allelic inactivation analysis.
For the second paper, I have been responsible for structure variation data processing, machine learning method implementation and part of manuscript writing.
For the third paper, I have been responsible for data analysis and bi-allelic inactivations.
For the fourth paper, I have been responsible for package implementation, data preprocessing and test, and manuscript writing.

Chen Hong
July 2023

Supplement

Supplemented_Table 1: Bi-allelic inactivation and entropy of CGC genes

gene_name	chr	gene_start	gene_end	entropy	sum	role
CDKN2A	9	21967751	21995300	2.21260178249442	304	TSG
TP53	17	7565097	7590856	2.29055453558936	239	oncogene, TSG, fusion
FHIT	3	59735036	61237133	1.70019938189912	132	TSG, fusion
SMAD4	18	48494410	48611415	1.26108178914847	107	TSG
PTEN	10	89622870	89731687	2.37614595831218	102	TSG
PTPRD	9	8314246	10612723	2.07352799380496	63	TSG
MAP2K4	17	11924141	12047147	1.91298755126949	37	oncogene, TSG
MLLT3	9	20341663	20622542	1.78307381513079	35	fusion
RB1	13	48877887	49056122	2.24619948328564	34	TSG
LRP1B	2	140988992	142889270	2.42736868324076	31	TSG
VHL	3	10182692	10193904	0.142505867392738	31	TSG
DCC	18	49866542	51057784	1.3068894077018	28	
FAT1	4	187508937	187647876	2.26891294638815	23	TSG
MUC16	19	8959520	9092018	1.75172691992395	23	oncogene
FAS	10	90750414	90775542	2.26038344273221	22	TSG
PBRM1	3	52579368	52719933	1.15312285025421	21	TSG
KRAS	12	25357723	25403870	0.425848449238581	18	oncogene
NF1	17	29421945	29709134	1.79810550262427	18	TSG, fusion
TGFBR2	3	30647994	30735634	1.20008732276787	17	TSG
APC	5	112043195	112181936	2.07944154167984	16	TSG
ARHGEF10	8	1772142	1906807	2.04673853269455	16	TSG
BRCA2	13	32889611	32973805	1.89892678933633	15	TSG
NCOR1	17	15932471	16121499	2.04493117484959	14	TSG

RUNX1	21	36160098	37376965	1.25276296849537	14	oncogene, TSG, fusion
CYSLTR2	13	49280951	49283498	1.8446214763655	13	oncogene
ERBB4	2	212240446	213403565	1.23426786607908	12	oncogene, TSG
NFIB	9	14081842	14398982	1.863679987341	12	fusion
NUTM2D	10	89117425	89130452	1.79175946922805	12	fusion
SETD2	3	47057919	47205457	1.23426786607908	12	TSG
ARID1A	1	27022524	27108595	1.59416669911802	11	TSG, fusion
AXIN1	16	337440	402673	0.304636097349238	11	TSG
ATM	11	108093211	108239829	1.47080847632211	10	TSG
MAP3K1	5	56111401	56191979	1.6094379124341	10	oncogene, TSG
ROBO2	3	75955846	77699115	1.6094379124341	10	TSG
B2M	15	45003675	45011075	1.58109375017182	9	TSG
CBFB	16	67063019	67134961	1.00271826451752	9	TSG, fusion
ITGAV	2	187454792	187545628	1.67698777432242	9	
JAK2	9	4985033	5128183	1.67698777432242	9	oncogene, fusion
NBEA	13	35516424	36247159	1.52295506753132	9	
PRDM16	1	2985732	3355185	1.52295506753132	9	oncogene, fusion
FOXP1	3	71003844	71633140	1.49417513828931	8	oncogene, fusion
JAK1	1	65298912	65432187	0.974314752869349	8	oncogene, TSG
ARID1B	6	157099063	157531913	1.74786809746676	7	TSG
CNTNAP2	7	145813453	148118090	1.94591014905531	7	TSG
CYP2C8	10	96796530	96829254	1.15374194270109	7	
IL6ST	5	55230923	55290821	0.955699891112534	7	oncogene
KAT6B	10	76585340	76792380	1.5498260458782	7	TSG, fusion
MEN1	11	64570982	64578766	0	7	TSG
TMPRSS2	21	42836478	42903043	0.955699891112534	7	fusion
TSC1	9	135766735	135820020	1.47507631105469	7	TSG

ZNF521	18	22641890	22932154	1.07899220787758	7	oncogene, fusion
BAP1	3	52435029	52444366	1.56071040904141	6	TSG
CBFA2T3	16	88941266	89043612	1.56071040904141	6	TSG, fusion
CDH1	16	68771128	68869451	1.242453324894	6	TSG
CYLD	16	50775961	50835846	1.242453324894	6	TSG
EED	11	85955586	85989855	1.32966134885476	6	TSG
GPHN	14	66974125	67648520	1.56071040904141	6	fusion
KMT2C	7	151832010	152133090	1.56071040904141	6	TSG
LATS1	6	149979289	150039392	1.32966134885476	6	TSG
MTOR	1	11166592	11322564	1.242453324894	6	oncogene
NCOR2	12	124808961	125052135	1.79175946922805	6	TSG
NRG1	8	31496902	32622548	1.79175946922805	6	TSG, fusion
PER1	17	8043790	8059824	1.79175946922805	6	TSG, fusion
SETBP1	18	42260138	42648475	1.32966134885476	6	oncogene, fusion
YWHAE	17	1247566	1303672	1.56071040904141	6	TSG, fusion
ZBTB16	11	113930315	114121398	0.867563228481461	6	TSG, fusion
ZFHX3	16	72816784	73093597	1.32966134885476	6	TSG
ZNRF3	22	29279580	29453475	1.56071040904141	6	TSG
ASXL2	2	25956622	26101385	1.05492016798614	5	TSG
CAMTA1	1	6845384	7829766	1.33217904021012	5	TSG, fusion
CASP8	2	202098166	202152434	1.33217904021012	5	TSG
CDKN2C	1	51426417	51440305	1.33217904021012	5	TSG
CIC	19	42772689	42799949	0.950270539233235	5	oncogene, TSG, fusion
CTNNB1	3	41236328	41301587	1.33217904021012	5	oncogene, fusion
FAT4	4	126237554	126414087	1.33217904021012	5	TSG
FBXW7	4	153242410	153457253	1.6094379124341	5	TSG
GOPC	6	117639374	117923691	1.6094379124341	5	fusion

GPC5	13	92050929	93519490	1.6094379124341	5	TSG
KEAP1	19	10596796	10614417	0.950270539233235	5	TSG
N4BP2	4	40058446	40159872	0.950270539233235	5	TSG
QKI	6	163835032	163999628	1.05492016798614	5	oncogene, TSG
RABEP1	17	5185558	5289129	1.33217904021012	5	fusion
RNF43	17	56429861	56494956	0.500402423538188	5	TSG
ROS1	6	117609463	117747018	1.33217904021012	5	oncogene, fusion
SPECC1	17	19912657	20222339	1.6094379124341	5	fusion
STK11	19	1189406	1228428	1.05492016798614	5	TSG
USP6	17	5019733	5078329	1.6094379124341	5	oncogene, fusion
VAV1	19	6772725	6857377	0.950270539233235	5	fusion
ARID2	12	46123448	46301823	1.38629436111989	4	TSG
CASP3	4	185548850	185570663	1.03972077083992	4	TSG
CD274	9	5450503	5470566	1.03972077083992	4	TSG, fusion
DICER1	14	95552565	95624347	1.03972077083992	4	TSG
EPHA3	3	89156674	89531284	0.562335144618808	4	
FANCA	16	89803957	89883065	1.03972077083992	4	TSG
FSTL3	19	676392	683385	1.38629436111989	4	oncogene, fusion
GAS7	17	9813926	10101868	1.38629436111989	4	fusion
KCNJ5	11	128761251	128790930	1.38629436111989	4	oncogene
NUTM1	15	34635516	34649938	1.38629436111989	4	oncogene, fusion
PAX7	1	18957500	19075360	1.03972077083992	4	fusion
PDCD1LG2	9	5510545	5571282	1.38629436111989	4	oncogene, fusion
PICALM	11	85668727	85780924	1.03972077083992	4	fusion
PIK3CA	3	178865902	178957881	1.38629436111989	4	oncogene
PRDM1	6	106534195	106557814	1.03972077083992	4	TSG
RALGDS	9	135973107	136039301	0.693147180559945	4	fusion
RHOH	4	40192673	40248587	0.562335144618808	4	TSG, fusion

SKI	1	2160134	2241558	1.38629436111989	4	oncogene
SMARCA4	19	11071598	11176071	1.38629436111989	4	TSG
TCF3	19	1609291	1652604	1.03972077083992	4	oncogene, TSG, fusion
TCF7L2	10	114710009	114927437	1.03972077083992	4	oncogene, fusion
TNC	9	117782806	117880536	1.03972077083992	4	oncogene
A1CF	10	52559169	52645435	0	3	oncogene
ACSL6	5	131142683	131347936	1.09861228866811	3	fusion
ACVR2A	2	148602086	148688393	0.636514168294813	3	TSG
BRD4	19	15347647	15443356	0.636514168294813	3	oncogene, fusion
CDH11	16	64977656	65160015	0.636514168294813	3	TSG, fusion
CDKN1B	12	12867992	12875305	0.636514168294813	3	TSG
CTCF	16	67596310	67673086	0	3	TSG
CTNNA2	2	79412357	80875905	0.636514168294813	3	oncogene
DNM2	19	10828755	10944164	0.636514168294813	3	TSG
ELK4	1	205577071	205601090	0	3	oncogene, fusion
ESR1	6	151977826	152450754	1.09861228866811	3	oncogene, TSG, fusion
FLCN	17	17115526	17140502	1.09861228866811	3	TSG
FOXO1	13	41129804	41240734	1.09861228866811	3	oncogene, TSG, fusion
FUBP1	1	78409740	78444794	0	3	oncogene
ID3	1	23884409	23886285	0	3	TSG
JUN	1	59246465	59249785	1.09861228866811	3	oncogene
LARP4B	10	855484	977564	0.636514168294813	3	TSG
MAF	16	79619740	79634611	1.09861228866811	3	oncogene, fusion
MAML2	11	95709762	96076344	0.636514168294813	3	oncogene, fusion
MAP2K2	19	4090319	4124126	1.09861228866811	3	oncogene
MDS2	1	23907985	23967058	0.636514168294813	3	fusion

MLLT1	19	6212966	6279959	1.09861228866811	3	fusion
MLLT10	10	21823094	22032559	1.09861228866811	3	oncogene, fusion
MN1	22	28144265	28197486	1.09861228866811	3	oncogene, fusion
NF2	22	29999545	30094587	1.09861228866811	3	TSG
NR4A3	9	102584137	102629173	1.09861228866811	3	oncogene, fusion
PAX5	9	36833272	37034103	0.636514168294813	3	oncogene, TSG, fusion
POLE	12	133200348	133263951	1.09861228866811	3	TSG
PRDM2	1	14026693	14151574	1.09861228866811	3	TSG
RHOA	3	49396578	49450431	1.09861228866811	3	oncogene, TSG
RNF213	17	78234665	78372586	0.636514168294813	3	fusion
SH3GL1	19	4360367	4400544	1.09861228866811	3	oncogene, fusion
SPEN	1	16174359	16266955	0.636514168294813	3	TSG
SRGAP3	3	9022275	9404737	1.09861228866811	3	fusion
SS18	18	23596578	23671181	1.09861228866811	3	fusion
SUFU	10	104263744	104393292	1.09861228866811	3	TSG
TET1	10	70320413	70454239	0.636514168294813	3	oncogene, TSG, fusion
TET2	4	106067032	106200973	0.636514168294813	3	TSG
TNFRSF14	1	2487078	2496821	0	3	TSG
TSC2	16	2097466	2138716	0.636514168294813	3	TSG
TSHR	14	81421333	81612646	0.636514168294813	3	oncogene
WNK2	9	95947198	96082854	0.636514168294813	3	TSG
XPC	3	14186647	14220283	1.09861228866811	3	TSG
ABL1	9	133589333	133763062	0.693147180559945	2	oncogene, fusion
ANK1	8	41510739	41754280	0.693147180559945	2	
ARHGAP26	5	142149949	142608576	0.693147180559945	2	TSG, fusion
ARHGAP5	14	32545320	32628934	0.693147180559945	2	oncogene
ARHGEF10L	1	17866330	18024369	0	2	TSG

ARHGEF12	11	120207787	120360645	0.693147180559945	2	TSG, fusion
ASPSCR1	17	79934683	79975282	0.693147180559945	2	fusion
BAZ1A	14	35221937	35344853	0	2	TSG
BCL11A	2	60678302	60780702	0.693147180559945	2	oncogene, fusion
CALR	19	13049392	13055303	0.693147180559945	2	oncogene
CARD11	7	2945775	3083579	0.693147180559945	2	oncogene
CBLB	3	105374305	105588396	0.693147180559945	2	TSG
CCDC6	10	61548521	61666414	0.693147180559945	2	TSG, fusion
CCR4	3	32993066	32997841	0.693147180559945	2	oncogene
CD209	19	7804879	7812464	0.693147180559945	2	
CHEK2	22	29083731	29138410	0.693147180559945	2	TSG
CLTCL1	22	19166986	19279239	0.693147180559945	2	TSG, fusion
COL3A1	2	189839046	189877472	0.693147180559945	2	fusion
CREB3L1	11	46299212	46342972	0.693147180559945	2	TSG, fusion
CREBBP	16	3775055	3930727	0.693147180559945	2	oncogene, TSG, fusion
CSF1R	5	149432854	149492935	0.693147180559945	2	oncogene
CSMD3	8	113235157	114449328	0.693147180559945	2	TSG
CUL3	2	225334867	225450110	0.693147180559945	2	TSG
CUX1	7	101458959	101927249	0.693147180559945	2	oncogene, TSG
DAXX	6	33286335	33297046	0	2	oncogene, TSG
EPHA7	6	93949738	94129265	0.693147180559945	2	
ERCC2	19	45853095	45874176	0.693147180559945	2	TSG
ERG	21	39751949	40033704	0.693147180559945	2	oncogene, fusion
ETV6	12	11802788	12048336	0	2	TSG, fusion
EWSR1	22	29663998	29696515	0.693147180559945	2	oncogene, fusion
FAT3	11	92085262	92629618	0	2	
FBXO11	2	48016455	48132932	0.693147180559945	2	TSG

FGFR1OP	6	167412670	167466201	0.693147180559945	2	fusion
FGFR3	4	1795034	1810599	0.693147180559945	2	oncogene, fusion
FIP1L1	4	54243810	55161439	0	2	fusion
FLT3	13	28577411	28674729	0.693147180559945	2	oncogene
FNBP1	9	132649466	132805473	0.693147180559945	2	fusion
FOXO3	6	108881038	109005977	0.693147180559945	2	oncogene, TSG, fusion
GATA3	10	8095567	8117161	0.693147180559945	2	oncogene, TSG
GNA11	19	3094408	3124002	0.693147180559945	2	oncogene
HSP90AA1	14	102547075	102606036	0.693147180559945	2	fusion
IKZF1	7	50343720	50472799	0.693147180559945	2	TSG, fusion
ISX	22	35462129	35483380	0.693147180559945	2	
KLF4	9	110247133	110252763	0.693147180559945	2	oncogene, TSG
KTN1	14	56025790	56168244	0.693147180559945	2	fusion
LEF1	4	108968701	109090112	0.693147180559945	2	oncogene, TSG
LEPROTL1	8	29952914	30034724	0.693147180559945	2	TSG
LPP	3	187871072	188608460	0.693147180559945	2	oncogene, fusion
LYL1	19	13209847	13213975	0	2	oncogene, fusion
MALT1	18	56338618	56417371	0	2	oncogene, fusion
MECOM	3	168801287	169381406	0	2	oncogene, fusion
MITF	3	69788586	70017488	0.693147180559945	2	oncogene
MLH1	3	37034823	37107380	0.693147180559945	2	TSG
MYC	8	128747680	128753674	0.693147180559945	2	oncogene, fusion
MYH11	16	15797029	15950890	0.693147180559945	2	fusion
NOTCH1	9	139388896	139440314	0.693147180559945	2	oncogene, TSG, fusion
NSD1	5	176560026	176727216	0	2	fusion
NTHL1	16	2089816	2097867	0.693147180559945	2	TSG
PCM1	8	17780349	17885478	0.693147180559945	2	fusion

PDE4DIP	1	144836157	145076186	0.693147180559945	2	fusion
PHOX2B	4	41746099	41750987	0	2	TSG
PIK3R1	5	67511548	67597649	0.693147180559945	2	TSG
PPP2R1A	19	52693292	52730687	0.693147180559945	2	TSG
PPP6C	9	127908852	127952218	0.693147180559945	2	TSG
PRKAR1A	17	66507921	66547460	0.693147180559945	2	oncogene, TSG, fusion
PRPF40B	12	49962001	50038449	0.693147180559945	2	
PTCH1	9	98205262	98279339	0	2	TSG
PTPN13	4	87515468	87736324	0.693147180559945	2	TSG
RAD17	5	68665120	68710628	0.693147180559945	2	TSG
RAD51B	14	68286496	69196935	0.693147180559945	2	TSG, fusion
RET	10	43572475	43625799	0.693147180559945	2	oncogene, fusion
RMI2	16	11343476	11445619	0	2	TSG, fusion
RSPO3	6	127439749	127518910	0	2	oncogene, fusion
SDHA	5	218356	256815	0.693147180559945	2	TSG
SLC45A3	1	205626979	205649587	0	2	fusion
SMAD2	18	45357922	45457515	0.693147180559945	2	TSG
SMARCB1	22	24129150	24176703	0.693147180559945	2	TSG
SOCS1	16	11348262	11350036	0	2	TSG
STAG1	3	136055077	136471220	0	2	TSG
SYK	9	93564069	93660831	0.693147180559945	2	oncogene, fusion
TBL1XR1	3	176737143	176915261	0.693147180559945	2	oncogene, TSG, fusion
TNFAIP3	6	138188351	138204449	0.693147180559945	2	TSG
TP63	3	189349205	189615068	0.693147180559945	2	oncogene, TSG
TRAF7	16	2205699	2228130	0.693147180559945	2	TSG
TRIP11	14	92432335	92507240	0.693147180559945	2	fusion
VTI1A	10	114206756	114578503	0.693147180559945	2	fusion

WIF1	12	65444406	65515346	0.693147180559945	2	TSG, fusion
ZEB1	10	31607424	31818742	0.693147180559945	2	oncogene
ZMYM2	13	20532810	20665968	0.693147180559945	2	fusion
ZNF331	19	54024235	54083523	0.693147180559945	2	TSG, fusion
ZNF429	19	21679484	21739072	0.693147180559945	2	
ACVR1	2	158592958	158732374	0	1	oncogene
AFF3	2	100162323	100759201	0	1	oncogene, fusion
AKT1	14	105235686	105262088	0	1	oncogene
ALK	2	29415640	30144432	0	1	oncogene, fusion
ATP1A1	1	116915290	116952883	0	1	oncogene, TSG
ATR	3	142168077	142297668	0	1	TSG
AXIN2	17	63524681	63557765	0	1	TSG
BAX	19	49458072	49465055	0	1	TSG
BCL11B	14	99635624	99737861	0	1	oncogene, TSG, fusion
BCL2	18	60790579	60987361	0	1	oncogene, fusion
BCL2L12	19	50168823	50177173	0	1	oncogene
BCL9L	11	118764584	118796317	0	1	oncogene, TSG
BCLAF1	6	136578001	136610989	0	1	
BCR	22	23521891	23660224	0	1	fusion
BIRC6	2	32582096	32843966	0	1	oncogene, fusion
BLM	15	91260558	91358859	0	1	TSG
BMP5	6	55618443	55740362	0	1	
BMPR1A	10	88516407	88692595	0	1	oncogene, TSG
BRD3	9	136895427	136933657	0	1	oncogene, fusion
BRIP1	17	59758627	59940882	0	1	TSG
C15orf65	15	55700746	55710962	0	1	fusion
CACNA1D	3	53528683	53847760	0	1	oncogene

CASP9	1	15817327	15853029	0	1	TSG
CCNE1	19	30302805	30315215	0	1	oncogene
CD74	5	149781200	149792492	0	1	oncogene, fusion
CD79A	19	42381190	42385439	0	1	oncogene
CDH10	5	24487209	24645087	0	1	TSG
CDH17	8	95139399	95229531	0	1	oncogene
CDKN1A	6	36644305	36655116	0	1	oncogene, TSG
CHD4	12	6679249	6716642	0	1	oncogene
CHST11	12	104849073	105155792	0	1	oncogene, fusion
CIITA	16	10971055	11026079	0	1	TSG, fusion
CLP 1	11	57416465	57429340	0	1	fusion
CLTC	17	57697219	57773671	0	1	TSG, fusion
CNBD1	8	87878670	88627447	0	1	
CNOT3	19	54641444	54659419	0	1	TSG
COL1A1	17	48260650	48278993	0	1	fusion
COL2A1	12	48366748	48398269	0	1	fusion
CPEB3	10	93806449	94050844	0	1	TSG
CREB1	2	208394461	208468155	0	1	oncogene, fusion
CREB3L2	7	137559725	137686813	0	1	oncogene, fusion
CRTC1	19	18794487	18893004	0	1	oncogene, fusion
CTNND1	11	57520715	57587018	0	1	
DEK	6	18224099	18265054	0	1	oncogene, fusion
DGCR8	22	20067755	20099400	0	1	oncogene
DNAJB1	19	14625582	14640582	0	1	fusion
DNMT3A	2	25455845	25565459	0	1	TSG
ECT2L	6	139117063	139225207	0	1	
EIF3E	8	109213445	109447562	0	1	TSG, fusion
EIF4A2	3	186500994	186507689	0	1	fusion

ELL	19	18553473	18632937	0	1	TSG, fusion
EP300	22	41487790	41576081	0	1	TSG, fusion
EPAS1	2	46520806	46613836	0	1	oncogene, TSG
ERBB2	17	37844167	37886679	0	1	oncogene, fusion
ERC1	12	1099675	1605099	0	1	fusion
ERCC3	2	128014866	128051752	0	1	TSG
ERCC4	16	14014014	14046202	0	1	TSG
EZH2	7	148504475	148581413	0	1	oncogene, TSG
FAM135B	8	139142266	139509065	0	1	
FANCD2	3	10068098	10143614	0	1	TSG
FANCF	11	22644079	22647387	0	1	TSG
FANCG	9	35073832	35080013	0	1	TSG
FBLN2	3	13573824	13679922	0	1	TSG
FCGR2B	1	161551101	161648444	0	1	oncogene, fusion
FCRL4	1	157543539	157567870	0	1	oncogene, fusion
FGFR4	5	176513887	176525145	0	1	oncogene
FLI1	11	128556430	128683162	0	1	oncogene, fusion
FLT4	5	180028506	180076624	0	1	oncogene
FOXR1	11	118842417	118852001	0	1	oncogene, fusion
GNAQ	9	80331003	80646374	0	1	oncogene
GRIN2A	16	9852376	10276611	0	1	TSG
GRM3	7	86273230	86494200	0	1	oncogene
HERPUD1	16	56965960	56977798	0	1	fusion
HIP1	7	75162621	75368280	0	1	oncogene, fusion
HIST1H3B	6	26031817	26032288	0	1	oncogene
HLA-A	6	29909037	29913661	0	1	fusion
HNF1A	12	121416346	121440315	0	1	TSG
HOOK3	8	42752075	42885682	0	1	fusion

HOXC11	12	54366910	54371427	0	1	oncogene, fusion
IL2	4	123372625	123377880	0	1	fusion
IL21R	16	27413483	27462115	0	1	fusion
IL7R	5	35852797	35879705	0	1	oncogene
ITK	5	156569944	156682201	0	1	fusion
JAK3	19	17935589	17958880	0	1	oncogene
KAT6A	8	41786997	41909508	0	1	oncogene, fusion
KDM5A	12	389295	498620	0	1	oncogene, fusion
KIF5B	10	32297938	32345359	0	1	fusion
KLF6	10	3818188	3827473	0	1	TSG
KLK2	19	51364824	51383823	0	1	fusion
KMT2D	12	49412758	49453557	0	1	oncogene, TSG
LATS2	13	21547171	21635686	0	1	TSG
LCP1	13	46700055	46786006	0	1	fusion
LIFR	5	38475065	38608456	0	1	fusion
LMO2	11	33880122	33913836	0	1	oncogene, fusion
LZTR1	22	21333751	21353327	0	1	TSG
MAFB	20	39314488	39317880	0	1	oncogene, fusion
MAX	14	65472892	65569413	0	1	TSG
MGMT	10	131265448	131566271	0	1	TSG
MNX1	7	156786745	156803345	0	1	fusion
MSI2	17	55333212	55762046	0	1	oncogene, fusion
MYB	6	135502453	135540311	0	1	oncogene, fusion
MYD88	3	38179969	38184513	0	1	oncogene
MYH9	22	36677327	36784063	0	1	TSG, fusion
MYO5A	15	52599480	52821247	0	1	fusion
NAB2	12	57482677	57489259	0	1	TSG, fusion
NBN	8	90945564	91015456	0	1	TSG

NCOA1	2	24714783	24993571	0	1	fusion
NFKB2	10	104153867	104162281	0	1	oncogene, TSG, fusion
NFKBIE	6	44225903	44233500	0	1	TSG
NIN	14	51186481	51297839	0	1	fusion
NKX2-1	14	36985602	36990354	0	1	oncogene, TSG
NRAS	1	115247090	115259515	0	1	oncogene
NUP214	9	134000948	134110057	0	1	fusion
NUP98	11	3692313	3819022	0	1	oncogene, fusion
NUTM2B	10	81462983	81474437	0	1	fusion
OLIG2	21	34398153	34401504	0	1	oncogene, fusion
PAFAH1B2	11	117014983	117047610	0	1	fusion
PATZ1	22	31721790	31742218	0	1	TSG, fusion
PDGFRA	4	55095264	55164414	0	1	oncogene, fusion
PDGFRB	5	149493400	149535435	0	1	oncogene, fusion
PMS1	2	190649107	190742355	0	1	
POLD1	19	50887461	50921273	0	1	TSG
POT1	7	124462440	124570037	0	1	TSG
PPARG	3	12328867	12475855	0	1	TSG, fusion
PRKACA	19	14202500	14228896	0	1	oncogene
PSIP1	9	15464064	15511017	0	1	oncogene, fusion
PTPRB	12	70910630	71031220	0	1	TSG
PTPRK	6	128289924	128841870	0	1	TSG, fusion
PTPRT	20	40701392	41818610	0	1	TSG
PWWP2A	5	159488808	159546430	0	1	fusion
RAC1	7	6414154	6443608	0	1	oncogene
RAD21	8	117858174	117887105	0	1	oncogene, TSG
RAF1	3	12625100	12705725	0	1	oncogene, fusion

RANBP2	2	109335937	109402267	0	1	TSG, fusion
REL	2	61108656	61158745	0	1	oncogene
RGS7	1	240931554	241520530	0	1	
RPL5	1	93297582	93307481	0	1	TSG
SDHD	11	111957627	112064528	0	1	TSG
SEPT9	17	75276651	75496678	0	1	fusion
SET	9	131445703	131458679	0	1	oncogene, fusion
SIRPA	20	1875154	1920543	0	1	TSG
SIX1	14	61110133	61124977	0	1	oncogene
SLC34A2	4	25656923	25680370	0	1	TSG, fusion
SMAD3	15	67356101	67487533	0	1	TSG
SNX29	16	12070594	12668146	0	1	fusion
SOX21	13	95361886	95364389	0	1	TSG
SRC	20	35973088	36034453	0	1	oncogene
STAT3	17	40465342	40540586	0	1	oncogene
STAT5B	17	40351186	40428725	0	1	oncogene, TSG, fusion
TAF15	17	34136459	34191619	0	1	oncogene, fusion
TAL1	1	47681962	47697892	0	1	oncogene, fusion
TCF12	15	57210821	57591479	0	1	fusion
TCL1A	14	96176304	96180533	0	1	oncogene, fusion
TEC	4	48137800	48271881	0	1	oncogene, fusion
TERT	5	1253262	1295184	0	1	oncogene, TSG
TFPT	19	54610320	54619055	0	1	fusion
TPM4	19	16177831	16213813	0	1	fusion
WRN	8	30891317	31031285	0	1	TSG
XPO1	2	61704984	61765761	0	1	oncogene

Supplemented_Table 2: Top 3 bi-allelic inactivation in every chromosome

Gene	Bi-allelic inactivation	CGC annotation	Chromosome	Gene_start	Gene_end
HSPG2	28	NA	1	22148738	22263790
ARID1A	20	TSG_fusion	1	27022524	27108595
PRDM16	13	oncogene_fusion	1	2985732	3355185
AC096579.7	53	NA	2	89130700	89165653
AC096579.13	49	NA	2	89109984	89161075
IGKV4-1	44	NA	2	89184913	89185669
FHIT	133	TSG_fusion	3	59735036	61237133
VHL	62	TSG	3	10182692	10193904
RP11-641C17.4	40	NA	3	60602555	60603812
FAT1	36	TSG	4	187508937	187647876
CCSER1	35	NA	4	91048686	92523064
DCHS2	24	NA	4	155153399	155412930
PCDHA1	40	NA	5	140165876	140391929
PCDHA2	40	NA	5	140174444	140391929
PCDHA3	40	NA	5	140180783	140391929
PARK2	25	NA	6	161768452	163148803
SYNE1	21	NA	6	152442819	152958936
MLLT4	16	NA	6	168227602	168372703
RP11-715L17.1	13	NA	7	61821869	61822186
THSD7A	13	NA	7	11409984	11871824
KMT2C	11	TSG	7	151832010	152133090
CSMD1	102	NA	8	2792875	4852494
DLGAP2	25	NA	8	1449532	1656642
UNC5D	21	NA	8	35092975	35654068
RP11-145E5.5	331	NA	9	21802635	22032985
CDKN2A	328	TSG	9	21967751	21995300
C9orf53	265	NA	9	21967137	21967738
PTEN	123	TSG	10	89622870	89731687

RP11-380G5.3	61	NA	10	89705259	89705781
MED6P1	60	NA	10	89807892	89809580
RP11-574M7.2	40	NA	11	50368213	50381487
ATM	18	TSG	11	108093211	108239829
HBG2	18	NA	11	5274420	5667019
KRAS	36	oncogene	12	25357723	25403870
NCOR2	9	TSG	12	124808961	125052135
ACVR1B	8	NA	12	52345451	52390862
RB1	46	TSG	13	48877887	49056122
BRCA2	24	TSG	13	32889611	32973805
LPAR6	21	NA	13	48963707	49018840
IGHJ6	65	NA	14	106329408	106329468
IGHJ5	54	NA	14	106330024	106330072
IGHJ2	42	NA	14	106331409	106331460
RYR3	18	NA	15	33603163	34158303
B2M	13	TSG	15	45003675	45011075
RP11-69H14.6	11	NA	15	22278010	22413497
WWOX	64	NA	16	78133310	79246564
RBFOX1	23	NA	16	6069095	7763340
AXIN1	20	TSG	16	337440	402673
TP53	470	oncogene_tsg_fusion	17	7565097	7590856
CTC-297N7.11	88	NA	17	10286461	10527203
RP11-799N11.1	80	NA	17	10286449	10441179
SMAD4	157	TSG	18	48494410	48611415
RP11-729L2.2	76	NA	18	48494389	48584514
MEX3C	47	NA	18	48700920	48744674
MUC16	43	oncogene	19	8959520	9092018
ABCA7	10	NA	19	1040102	1065571
ZNF728	10	NA	19	23158270	23185978
MACROD2	38	NA	20	13976015	16033842
MACROD2-AS1	17	NA	20	14864899	14910161

RP5-974N19.1	14	NA	20	14914489	14916009
RUNX1	15	oncogene_tsg_fusion	21	36160098	37376965
ANKRD30BP1	14	NA	21	14756570	14800094
FGF7P2	14	NA	21	14721586	14722969
IGLV3-1	59	NA	22	23222886	23223576
IGLL5	47	NA	22	23229960	23238287
IGLV3-10	28	NA	22	23154244	23154782

Supplemented_Table 3: Top 100 codons with highest CaSINo scores

gene	**codon**	**score**	**role**
FAM174B	FAM174B:p.S68	0.00718949293716475	NA
KRAS	KRAS:p.G12	0.00603122834678532	oncogene
TIPIN	TIPIN:p.R142	0.00557872475374628	NA
TIPIN	TIPIN:p.R41	0.00557872475374628	NA
BRAF	BRAF:p.V600	0.00515421412977671	oncogene, fusion
BRAF	BRAF:p.V28	0.00515421412977671	oncogene, fusion
JAK2	JAK2:p.V617	0.00435965029916284	oncogene, fusion
JAK2	JAK2:p.V468	0.00435965029916284	oncogene, fusion
GBP4	GBP4:p.M545	0.0043006233029069	NA
DSPP	DSPP:p.D673	0.00282263267249971	NA
HLA-DQA1	HLA-DQA1:p.A92	0.00252451047643133	NA
HLA-DQB1	HLA-DQB1:p.G121	0.00229541431993378	NA
RP1L1	RP1L1:p.T1327	0.00223969938866847	NA
IDH1	IDH1:p.R132	0.00211030292005979	oncogene
PIK3CA	PIK3CA:p.H1047	0.00207690802875385	oncogene
IGFN1	IGFN1:p.E2099	0.00204246380040236	NA
TPRXL	TPRXL:p.S219	0.00202385850681739	NA
HLA-DRB1	HLA-DRB1:p.S66	0.00185079665532673	NA
C2orf82	C2orf82:p.A10	0.00184432670543624	NA
DPP7	DPP7:p.L47	0.00184432670543624	NA
IGHV3-23	IGHV3-23:p.S76	0.00184261254987945	NA
AL390778.1	AL390778.1:p.A81	0.00181121893587705	NA
RP1L1	RP1L1:p.E1328	0.00178890289578971	NA
IGHV3-23	IGHV3-23:p.A69	0.00177690861158474	NA

AKAP17A	AKAP17A:p.R416	0.0017585200766246	NA
RARRES1	RARRES1:p.D42	0.00175597356036723	NA
SIRPB1	SIRPB1:p.L95	0.0017488424391218	NA
QRICH2	QRICH2:p.G634	0.00172987345714036	NA
HLA-DQA1	HLA-DQA1:p.M89	0.00171383483680916	NA
IGHV3-23	IGHV3-23:p.G72	0.00170889512211672	NA
ATP1A3	ATP1A3:p.H1151	0.00170070475958498	NA
GBP4	GBP4:p.E546	0.00169760940310457	NA
KRTAP9-1	KRTAP9-1:p.C162	0.00167076299295663	NA
HLA-DQA2	HLA-DQA2:p.Q241	0.00166051511514349	NA
IGHV3-23	IGHV3-23:p.S73	0.0016226730799879	NA
IGHV3-23	IGHV3-23:p.S71	0.00161826269820383	NA
ACTA1	ACTA1:p.M213	0.00155981261179879	NA
ACTA1	ACTA1:p.M301	0.00155981261179879	NA
C5orf60	C5orf60:p.D22	0.00155676466354221	NA
SIRPA	SIRPA:p.A57	0.00150193209751132	TSG
HLA-DQA1	HLA-DQA1:p.G84	0.00143892439514527	NA
PTCH1	PTCH1:p.E91	0.00143047027118658	TSG
PTCH1	PTCH1:p.E223	0.00143047027118658	TSG
PTCH1	PTCH1:p.E253	0.00143047027118658	TSG
PTCH1	PTCH1:p.E308	0.00143047027118658	TSG
PTCH1	PTCH1:p.E374	0.00143047027118658	TSG
PTCH1	PTCH1:p.E373	0.00143047027118658	TSG
HLA-DQA1	HLA-DQA1:p.I98	0.0014292600904128	NA
PABPC3	PABPC3:p.Y417	0.00141489296138706	NA
AL390778.1	AL390778.1:p.T27	0.00140587374091021	NA
MUC19	MUC19:p.G2758	0.00137987163154403	NA

USP8	USP8:p.R763	0.00133031932489563	oncogene
USP8	USP8:p.R657	0.00133031932489563	oncogene
USP8	USP8:p.N764	0.00132987313560467	oncogene
USP8	USP8:p.N658	0.00132987313560467	oncogene
AL390778.1	AL390778.1:p.H36	0.00132232102940341	NA
GBP4	GBP4:p.M542	0.00128304381927286	NA
USP8	USP8:p.L776	0.00128054997202534	oncogene
USP8	USP8:p.L670	0.00128054997202534	oncogene
OR51B2	OR51B2:p.R160	0.00127688803856289	NA
ICOSLG	ICOSLG:p.E418	0.00126145527771507	NA
HLA-DRB1	HLA-DRB1:p.A14	0.00126131098700082	NA
TP53	TP53:p.R141	0.00126122582184646	oncogene, TSG, fusion
TP53	TP53:p.R273	0.00126122582184646	oncogene, TSG, fusion
MMRN2	MMRN2:p.E421	0.00125856600919755	NA
KRTAP4-5	KRTAP4-5:p.R87	0.00122546896821448	NA
FAM8A1	FAM8A1:p.R220	0.00120983569185502	NA
ACTC1	ACTC1:p.T105	0.0011993639607043	NA
KRTAP1-1	KRTAP1-1:p.R38	0.00118722948103887	NA
C16orf3	C16orf3:p.S57	0.00118480007095189	NA
SPATA20	SPATA20:p.E209	0.00117769339831916	NA
SPATA20	SPATA20:p.E253	0.00117769339831916	NA
SPATA20	SPATA20:p.E269	0.00117769339831916	NA
DSPP	DSPP:p.N685	0.00117614573433925	NA
FGFR1	FGFR1:p.N544	0.00116896300243733	oncogene, fusion
FGFR1	FGFR1:p.N546	0.00116896300243733	oncogene, fusion
PLIN4	PLIN4:p.A784	0.00115820637238598	NA
ZNF132	ZNF132:p.R476	0.00114998267789045	NA

FAM186A	FAM186A:p.E1586	0.00114909206691322	NA
FGFR1	FGFR1:p.N536	0.0011397408718363	oncogene, fusion
FGFR1	FGFR1:p.N457	0.0011397408718363	oncogene, fusion
FGFR1	FGFR1:p.N577	0.0011397408718363	oncogene, fusion
FGFR1	FGFR1:p.N455	0.0011397408718363	oncogene, fusion
MUC22	MUC22:p.A1210	0.00113948998983416	NA
PIK3CA	PIK3CA:p.E545	0.00113805146308361	oncogene
IGLV3-1	IGLV3-1:p.S70	0.0011336021434307	NA
ACTR3C	ACTR3C:p.Q119	0.00111288209322971	NA
ACTR3C	ACTR3C:p.Q121	0.00111288209322971	NA
ACTR3C	ACTR3C:p.Q220	0.00111288209322971	NA
KRTAP4-5	KRTAP4-5:p.Q82	0.00110743403711935	NA
PABPC3	PABPC3:p.R278	0.00110581945966735	NA
SIRPA	SIRPA:p.R54	0.00109830827770534	TSG
FGFR1	FGFR1:p.K687	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K654	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K656	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K56	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K567	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K565	0.00109772944982884	oncogene, fusion
FGFR1	FGFR1:p.K646	0.00109772944982884	oncogene, fusion
FAM230A	FAM230A:p.V238	0.00108996835895924	NA

Supplemented_Table 4: Top 100 positions with highest CaSINo scores

gene	position	score	role
TIPIN	15:66641448:TIPIN	0.00557872475374628	NA
BRAF	7:140453136:BRAF	0.00513523465793539	oncogene, fusion
KRAS	12:25398284:KRAS	0.00477408641543645	oncogene
FAM174B	15:93198688:FAM174B	0.00438019804840099	NA
JAK2	9:5073770:JAK2	0.00435965029916284	oncogene, fusion
FAM174B	15:93198687:FAM174B	0.00280929488876375	NA
GBP4	1:89652088:GBP4	0.0021676901083092	NA
RP1L1	8:10467628:RP1L1	0.00216157424648286	NA
GBP4	1:89652090:GBP4	0.00213293319459771	NA
IGFN1	1:201180317:IGFN1	0.00204246380040236	NA
TPRXL	3:14106332:TPRXL	0.00202385850681739	NA
PIK3CA	3:178952085:PIK3CA	0.00199396549476417	oncogene
HLA-DRB1	6:32552060:HLA-DRB1	0.00185079665532673	NA
QRICH2	17:74288410:QRICH2	0.00172987345714036	NA
HLA-DQA1	6:32609271:HLA-DQA1	0.00171383483680916	NA
ATP1A3	19:42470962:ATP1A3	0.00170070475958498	NA
GBP4	1:89652087:GBP4	0.00169760940310457	NA
HLA-DQA1	6:32609278:HLA-DQA1	0.00167795226265418	NA
KRTAP9-1	17:39346622:KRTAP9-1	0.00167076299295663	NA
HLA-DQA2	6:32714125:HLA-DQA2	0.00166051511514349	NA
IDH1	2:209113112:IDH1	0.00163064204424891	oncogene
C5orf60	5:179071958:C5orf60	0.00155676466354221	NA
SIRPA	20:1895835:SIRPA	0.00150193209751132	TSG
DSPP	4:88535832:DSPP	0.00143677064554804	NA
HLA-DQA1	6:32609297:HLA-DQA1	0.0014292600904128	NA
AL390778.1	9:138151197:AL390778.1	0.00140587374091021	NA

DSPP	4:88535831:DSPP	0.00138586202695167	NA
RP1L1	8:10467626:RP1L1	0.00138482799134645	NA
MUC19	12:40876727:MUC19	0.00137987163154403	NA
USP8	15:50784950:USP8	0.00133031932489563	oncogene
USP8	15:50784955:USP8	0.00132987313560467	oncogene
AL390778.1	9:138151169:AL390778.1	0.00132232102940341	NA
GBP4	1:89652097:GBP4	0.00128304381927286	NA
ICOSLG	21:45649582:ICOSLG	0.00126145527771507	NA
KRAS	12:25398285:KRAS	0.00125714193134887	oncogene
KRTAP4-5	17:39305760:KRTAP4-5	0.00122546896821448	NA
FAM8A1	6:17601299:FAM8A1	0.00120983569185502	NA
AL390778.1	9:138151035:AL390778.1	0.00119181256857061	NA
HLA-DQB1	6:32632592:HLA-DQB1	0.00119171912174055	NA
DSPP	4:88535868:DSPP	0.00117614573433925	NA
PLIN4	19:4511580:PLIN4	0.00115820637238598	NA
FGFR1	8:38274849:FGFR1	0.0011397408718363	oncogene, fusion
HLA-DRB1	6:32557480:HLA-DRB1	0.00113446456753267	NA
C16orf3	16:90095582:C16orf3	0.00112688103413107	NA
PABPC3	13:25671168:PABPC3	0.00110581945966735	NA
HLA-DQB1	6:32632593:HLA-DQB1	0.00110369519819323	NA
SIRPA	20:1895826:SIRPA	0.00109830827770534	TSG
FGFR1	8:38272308:FGFR1	0.00109772944982884	oncogene, fusion
FAM230A	22:20708980:FAM230A	0.00108996835895924	NA
ROS1	6:117622184:ROS1	0.00106647945203623	oncogene, fusion
ROS1	6:117622188:ROS1	0.00106647945203623	oncogene, fusion
GZMH	14:25076877:GZMH	0.00102970151745312	NA
IGSF9B	11:133788992:IGSF9B	0.001025474553459	NA
KRT18	12:53343099:KRT18	0.0010215163119449	NA

OR4A16	11:55111365:OR4A16	0.00101140831992807	NA
FAM8A1	6:17601284:FAM8A1	0.0010080118555056	NA
ICOSLG	21:45649580:ICOSLG	0.00100137226680314	NA
HLA-DQB1	6:32632601:HLA-DQB1	0.000994302883255517	NA
FAM230A	22:20708977:FAM230A	0.000982030508409981	NA
PIK3CA	3:178936091:PIK3CA	0.00097431709779524	oncogene
SIRPA	20:1895820:SIRPA	0.000968439730501564	TSG
IGLV3-1	22:23223439:IGLV3-1	0.000966230549682632	NA
TP53	17:7577120:TP53	0.000962178969941832	oncogene, TSG, fusion
HLA-DQA1	6:32609254:HLA-DQA1	0.000952883236204546	NA
GZMB	14:25101589:GZMB	0.00093761420894176	NA
C6	5:41159289:C6	0.000937573899872721	NA
ICOSLG	21:45649487:ICOSLG	0.000936870601312229	NA
KRTAP4-5	17:39305769:KRTAP4-5	0.00093545660112285	NA
IGHV3-23	14:106725347:IGHV3-23	0.000935361027248267	NA
IGHV3-23	14:106725337:IGHV3-23	0.000928988816217323	NA
EMR2	19:14877816:EMR2	0.000928722212758647	NA
HLA-DRB1	6:32552050:HLA-DRB1	0.000927913629026232	NA
MUC17	7:100680117:MUC17	0.000926470214057068	NA
KRTAP4-5	17:39305774:KRTAP4-5	0.000921967681059945	NA
IGHV3-23	14:106725397:IGHV3-23	0.000921652479822614	NA
MACF1	1:39835746:MACF1	0.000915004413640158	NA
SIRPB1	20:1592153:SIRPB1	0.000907086436048057	NA
KRT18	12:53343303:KRT18	0.000904727800990854	NA
OR1L4	9:125486717:OR1L4	0.000903891723142766	NA
IGHV3-23	14:106725403:IGHV3-23	0.000901603361392744	NA
PER1	17:8047060:PER1	0.000896175593954048	TSG, fusion

IGHV3-23	14:106725325:IGHV3-23	0.000887888189987088	NA
IGHV3-23	14:106725328:IGHV3-23	0.000887888189987088	NA
HLA-DQA2	6:32713771:HLA-DQA2	0.000885988176415229	NA
ATXN1	6:16327903:ATXN1	0.000878752291386945	NA
GBP4	1:89652102:GBP4	0.000873474787735532	NA
HLA-DQB1	6:32632605:HLA-DQB1	0.000863269632587438	NA
HLA-DQB1	6:32632659:HLA-DQB1	0.000861073135132222	NA
FAM186A	12:50746414:FAM186A	0.000850298002092954	NA
HLA-DQA1	6:32609279:HLA-DQA1	0.000846558213777152	NA
IGHV3-23	14:106725335:IGHV3-23	0.000842766774088507	NA
SIRPB1	20:1592152:SIRPB1	0.000841756003073742	NA
IGHV3-23	14:106725346:IGHV3-23	0.000841547584336478	NA
IGHV3-23	14:106725341:IGHV3-23	0.000838356392304435	NA
IGHV3-23	14:106725344:IGHV3-23	0.000838356392304435	NA
HLA-DQB1	6:32632646:HLA-DQB1	0.000830449095787504	NA
IGHV3-23	14:106725324:IGHV3-23	0.000825027721797978	NA
CES1	16:55862791:CES1	0.000820958950910637	NA
MUC17	7:100680120:MUC17	0.000815413541418485	NA
KRTAP4-5	17:39305773:KRTAP4-5	0.000805871889484953	NA

Supplemented_Table 5: Top 200 genes with highest CNV focality peak scores (focal deletions below 10Mb)

chr	gene	score	length	cgc	tsg	peak	peak_value
X	ARHGEF9	1605.16695187	150579	none	none	TRUE	1507.1866407483
3	FHIT	983.081020351	1502097	cgc	tsg	TRUE	851.213332309
16	WWOX	702.885882163	1113254	none	none	TRUE	589.587174635
9	RP11-143M1.7	801.29063767	20871	none	none	FALSE	554.880504251
19	OR4G1P	499.095780377	936	none	none	FALSE	499.095780377
9	CBWD1	815.021236355	67938	none	none	TRUE	408.273869869
18	RP11-451L19.1	551.401327872	4318	none	none	FALSE	397.621266177
4	CCSER1	514.096053126	1474378	none	none	TRUE	396.0853319
20	MACROD2	419.008793511	2057827	none	none	TRUE	394.6318504121
2	LRP1B	468.266307734	1900278	cgc	tsg	TRUE	379.0055110565
8	RPL23AP53	758.121098959	19045	none	none	FALSE	368.14272893
9	PTPRD	523.842247213	2298477	none	none	TRUE	366.125660343
16	PIEZO1	545.981268777	69868	none	none	TRUE	362.291346399
9	RP11-145E5.5	816.881232864	230350	none	none	TRUE	358.31389795
5	PDE4D	513.872046231	1553082	none	none	TRUE	355.272712825
16	CDT1	539.040613098	6045	none	none	FALSE	351.376712115
X	DMD	471.26291887	2241764	none	none	TRUE	341.202312843
6	PARK2	462.806541116	1380351	none	none	TRUE	323.433735
13	TMCO3	432.504336804	59232	none	none	TRUE	306.599718089
13	DCUN1D2	426.548420828	35133	none	none	TRUE	300.643802113
8	CSMD1	529.889931497	2059619	none	none	TRUE	265.700602655
22	C22orf34	418.50503147	243014	none	none	TRUE	262.516957371

9	MTAP	641.440687282	129104	none	none	FALSE	262.475774891
18	RP11-683L23.2	268.163004759	552	none	none	FALSE	232.3764314713
3	LSAMP	266.159193365	2194860	none	none	TRUE	229.4422784713
15	HERC2P3	228.897147543	123564	none	none	TRUE	228.897147543
12	FAM138D	225.448348371	2277	none	none	TRUE	223.49362357013
22	TTC28	314.454388867	701849	none	none	TRUE	214.6523748092
22	CECR2	216.463585853	197013	none	none	TRUE	211.39076194082
10	OR6L1P	216.876148557	909	none	none	FALSE	210.02995199512
8	RP11-585F1.6	965.811976951	237	none	none	FALSE	207.690877992
2	AC096579.13	271.37816513	51091	none	none	FALSE	207.4096445568
X	HNRNPDP1	305.266947846	753	none	none	FALSE	207.2866367243
21	SLC19A1	279.541983631	50839	none	none	TRUE	206.4478502097
9	UBA52P6	634.144880311	381	none	none	FALSE	201.931506105
9	AL953854.2	333.346842128	18467	none	none	TRUE	201.021815016
11	ODF3	361.071249459	3523	none	none	TRUE	198.820180018
3	NAALADL2	214.090323188	1367065	none	none	TRUE	187.4105286467
9	DOCK8	409.961156319	250405	none	none	TRUE	185.756167513
17	KSR1	185.391112767	169791	none	none	TRUE	173.8951568951
7	IMMP2L	243.489201756	899463	none	none	TRUE	173.2268774301
10	PTEN	510.098596345	108817	cgc	tsg	TRUE	170.554919047
21	TMPRSS2	231.366671308	66565	cgc	none	TRUE	167.9189517437
6	PDCD2	167.409923677	9397	none	none	FALSE	167.409923677
16	RP11-488I20.3	173.579208248	27172	none	none	TRUE	166.13189229782
6	EYS	261.241127135	1987242	none	none	TRUE	159.304579183
16	RBFOX1	200.942476258	1694245	none	none	TRUE	155.5126300306

X	XG	425.162263031	64448	none	none	TRUE	155.307457767
6	RP3-416J7.5	408.588463793	3079	none	none	FALSE	154.734720889
18	GREB1L	160.870797931	283175	none	none	TRUE	151.8470848907
X	DIAPH2	309.864624657	920334	none	none	TRUE	149.115974971
18	RP11-296E23.1	157.494152837	70731	none	none	TRUE	148.4704397967
19	CTD-3113P16.7	447.688006667	102	none	none	FALSE	143.792661091
6	GMDS	252.706740882	621885	none	none	TRUE	143.394742457
5	HMGB1P47	144.913980084	588	none	none	FALSE	143.27997564304
3	RP11-451B8.1	142.516402752	62227	none	none	FALSE	142.516402752
11	DLG2	199.511344232	2172911	none	none	TRUE	142.4430271605
16	RP11-420N3.2	177.322492846	1536212	none	none	TRUE	137.5290720993
11	AP004550.1	160.578921934	9660	none	none	FALSE	136.3704186531
Y	RP1-85D24.3	164.25720647	198	none	none	FALSE	135.2928063879
8	RP11-1134I14.2	139.922941093	109289	none	none	TRUE	134.13573602066
13	PHF2P2	134.702569639	115963	none	none	TRUE	131.60873917259
2	AC093642.4	181.077287077	993	none	none	FALSE	130.8805214313
8	VN1R46P	136.428556477	849	none	none	FALSE	130.64135140466
13	RB1	353.963323174	178235	cgc	tsg	TRUE	129.762147722
3	ROBO2	296.547293581	1743269	none	none	TRUE	128.485921266
21	ERG	178.373648902	281755	cgc	none	TRUE	128.3158152687
9	EXD3	261.960615067	116366	none	none	TRUE	128.308590533
12	ULK1	306.203996284	28516	none	none	FALSE	127.891772029
12	IQSEC3	222.833677972	111695	none	none	FALSE	126.3452733072

3	RP11-641C17.4	388.16100518	1257	none	none	FALSE	125.127632455
22	CLCP1	129.602087307	387	none	none	TRUE	124.52926339482
18	ROCK1P1	222.767789549	13154	none	none	FALSE	123.1498753073
10	KSR1P1	126.805595512	239	none	none	TRUE	122.65180652804
6	PRIM2	176.497638756	333772	none	none	TRUE	122.0764966093
14	IGHD6-13	484.313822713	20	none	none	FALSE	121.528863126
11	BET1L	453.00743923	39644	none	none	TRUE	121.216744912
4	TEC	121.076620942	134081	none	none	TRUE	120.098390321093
2	AC068287.1	196.832234871	294	none	none	TRUE	117.9683753119
17	RPTOR	175.11615923	421552	none	none	TRUE	115.5194568333
11	RP11-574M7.2	174.151133084	13274	none	none	TRUE	113.8485284029
1	SMYD3	168.828101588	757972	none	none	TRUE	111.1278045425
12	ANKS1B	168.285637695	1258197	none	none	TRUE	110.885708083
17	RP11-285M22.2	121.539520972	1506	none	none	FALSE	110.0435651001
17	RP11-720N19.1	121.152889029	837	none	none	FALSE	109.6569331571
17	RP11-28G8.1	165.7002327	3992	none	none	FALSE	106.1035303033
10	RNLS	360.938563559	310666	none	none	TRUE	106.093878795
10	NRG3	227.039464608	1111865	none	none	TRUE	104.974074907
7	CNTNAP2	146.363051333	2304637	none	none	TRUE	102.6547085099
X	KRT8P17	200.273318969	1447	none	none	TRUE	102.2930078473
21	AP001464.4	128.482204169	17016	none	none	TRUE	101.6275091816
2	AC096579.7	303.15022477	34953	none	none	TRUE	99.499138889
7	AUTS2	135.278205248	1194149	none	none	TRUE	98.7023179108
2	PARD3B	147.166361176	1074370	none	none	TRUE	98.312868682
19	CTD-3113P16.9	303.895345576	86	none	none	FALSE	98.061381521

4	DCUN1D4	96.370966393	73837	none	none	TRUE	95.392735772093
3	SUMF1	218.274356832	766467	none	none	TRUE	95.069953452
1	AGBL4	151.829445983	1491058	none	none	TRUE	95.02113137
15	DNM1P46	153.257944392	16771	none	none	TRUE	93.3532921468
16	ANKRD26P1	99.821848828	99756	none	none	TRUE	92.37453287782
6	RP11-448N11.3	98.918254891	174	none	none	TRUE	91.92793286749
2	ERBB4	149.36804223	1163119	cgc	none	TRUE	91.480318762
X	TMLHE	298.214316712	179829	none	none	TRUE	90.466215497
20	RP5-974N19.1	132.829971717	1520	none	none	FALSE	90.1085850902
21	RUNX1	138.97752117	1216867	cgc	none	TRUE	87.8886529897
14	IGHJ6	453.736157896	60	none	none	FALSE	87.728897153
4	GRID2	195.762943719	1470157	none	none	TRUE	87.184487208
8	SGCZ	258.348368361	1148475	none	none	TRUE	86.904511457
7	VIPR2	99.7580533899	116783	none	none	FALSE	86.190588756
8	NRG1	230.202500332	1125646	cgc	none	TRUE	85.699922428
11	SHANK2	122.336235683	649662	none	none	TRUE	85.1648911784
18	SMAD4	266.445980456	117005	cgc	tsg	TRUE	85.135169866
9	RP11-149I2.5	608.909697422	1616	none	none	FALSE	84.1519741980001
6	RP11-143A22.1	90.402072003	24972	none	none	TRUE	83.41174997949
3	FOXP1	247.001202827	629296	cgc	none	TRUE	83.099995985
17	NF1	229.897396738	287189	cgc	none	TRUE	82.823798292
9	CDKN2A	765.193087911	27549	cgc	tsg	TRUE	82.094326516
16	PRDM7	504.26064704	35506	none	none	FALSE	81.939987395
11	CNTN5	147.13504555	1337933	none	none	TRUE	81.2366615591
2	ASTL	82.7501689186	14586	none	none	TRUE	81.18303974033

9	RP11-408N14.1	432.213374206	10682	none	none	FALSE	81.145366654
1	NEGR1	144.516531172	886794	none	none	TRUE	81.1015985073
9	TUBBP5	86.7570168857	27256	none	none	FALSE	80.40243187855
3	ROBO1	227.087943262	1170575	none	none	TRUE	80.006056761
4	AF146191.4	254.928331794	159936	none	none	TRUE	79.226149238
X	KDM6A	229.152943935	239090	cgc	none	TRUE	78.839247815
10	TUBB8	126.69754793	27275	none	none	TRUE	78.283556505
22	IGLC2	191.166818742	461	none	none	FALSE	78.065511512
15	CTD-2054N24.2	159.314799876	70125	none	none	TRUE	77.3373073331
8	RP11-1134I14.4	82.8719072037	255	none	none	FALSE	77.08470213136
14	RAD51B	195.284332059	910439	cgc	none	TRUE	77.011665085
13	GPC6	160.629490954	1180560	none	none	TRUE	76.3502984094
9	FOCAD	270.855102854	337646	none	none	TRUE	75.93133738
8	ADAM5	198.938961066	102787	none	none	TRUE	75.349395094
X	IL1RAPL1	199.907726589	1369324	none	none	TRUE	75.23974709
10	PCDH15	171.602659122	1825171	none	none	TRUE	74.7055985273
3	RP11-641C17.1	247.908434836	492	none	none	TRUE	74.275971406
9	C9orf53	683.098761395	601	none	none	FALSE	74.189063973
6	DUSP22	253.853742904	59725	none	none	FALSE	73.997222206
5	RP11-6N13.1	209.577654723	1009672	none	none	TRUE	73.480237561
3	CADM2	215.336866615	1115447	none	none	TRUE	71.175086178
Y	RP1-85D24.1	164.25720647	168	none	none	FALSE	70.4801990122
10	PRKG1	163.809540021	1307165	none	none	TRUE	70.2217671993
12	RP11-297L6.2	70.7190878032	470	none	none	FALSE	70.101171532163

1	RP11-107G24.3	122.156644431	678	none	none	FALSE	68.7869728727
3	PTMAP8	127.890757256	341	none	none	FALSE	68.1841980091
20	MACROD2-AS1	165.831205496	45262	none	none	TRUE	68.0243944927
2	AC131097.3	290.038352264	197359	none	none	TRUE	67.970202779
10	CTNNA3	156.809331353	1783651	none	none	TRUE	66.7164209035
9	CACNA1B	180.681669125	246835	none	none	TRUE	66.513037923
9	RP11-70L8.4	524.757723224	3016	none	none	FALSE	66.1903883099999
10	C10orf11	163.552758745	958927	none	none	TRUE	65.680725036
X	STS	290.48631554	135354	none	none	TRUE	65.119211226
14	OR4K4P	290.273219866	683	none	none	FALSE	64.575234353
17	CCDC144A	157.488096114	114916	none	none	FALSE	64.4348444786
17	KANSL1	178.7752674	195451	none	none	TRUE	64.071945147
16	IL9RP3	136.509488699	8985	none	none	FALSE	63.8165461372
11	RP11-179A16.1	113.790255055	838231	none	none	TRUE	63.4705789761
X	HDHD1	313.516551555	99270	none	none	TRUE	63.372294607
9	RP11-160N1.9	284.108013489	11807	none	none	TRUE	63.154579999
1	RP11-206L10.11	367.21929363	31838	none	none	TRUE	62.666714682
9	RP11-370B11.3	310.12120835	1141	none	none	FALSE	62.618036131
2	AC093642.3	241.935135672	6593	none	none	TRUE	60.857848595
1	RERE	312.887496676	465245	none	none	TRUE	60.458081326
5	CTD-2254N19.1	248.002711034	11771	none	none	TRUE	60.116736526
16	ANKRD11	607.332084325	222931	none	none	TRUE	59.911352804
16	CTD-2144E22.8	160.244070365	1100	none	none	FALSE	59.877187671

7	PTPRN2	143.862709051	1048730	none	none	TRUE	59.8672921705
11	OR5I1	64.2667945963	944	none	none	FALSE	59.5555224455
6	RP3-416J7.4	467.432726532	24018	none	none	TRUE	58.844262739
2	AC010983.1	95.4372612376	22659	none	none	TRUE	58.4299475147
4	LRBA	179.179987665	751285	none	none	TRUE	58.281903893
X	IL1RAPL2	196.756439107	1200826	none	none	TRUE	58.215478715
4	GALNTL6	216.55974076	1229305	none	none	TRUE	58.01729961
18	AP005901.1	66.901550282	22619	none	none	TRUE	57.8778372417
19	HAVCR1P1	57.5925676883	1055	none	none	TRUE	57.5925676883
14	NRXN3	146.089421689	1622028	none	none	TRUE	56.9926734185
3	ULK4	132.362462312	715832	none	none	TRUE	56.9550830037
16	Z84812.4	128.265965339	17681	none	none	TRUE	56.8065996106
2	NCKAP5	129.37216507	896660	none	none	TRUE	56.7805298513
6	GRIK2	181.100628623	671294	none	none	TRUE	56.552278129
1	FAF1	110.625066051	520785	none	none	TRUE	55.9263558793
3	ZNF385D	126.321816659	954897	none	none	TRUE	55.7151617174
3	ARMC10P1	55.6831394797	854	none	none	FALSE	55.6831394797
14	GPHN	156.542514627	674395	cgc	none	TRUE	55.193035137
13	LPAR6	279.564443489	55133	none	none	TRUE	55.133637628
19	LINC01002	502.705722687	4614	none	none	FALSE	55.01771602
16	AXIN1	165.777520334	65233	cgc	tsg	TRUE	54.947553406
2	NRXN1	78.7227351044	1114031	none	none	TRUE	54.5633763404
3	ERC2	183.595178009	960055	none	none	TRUE	54.280959994
1	RP11-417J8.1	123.731754699	5877	none	none	FALSE	54.1656548334
17	RP11-219A15.1	157.488096114	114192	none	none	FALSE	54.049014471
7	THSD7A	83.0899978248	461840	none	none	TRUE	53.7529392028

6	RP11-452D24.1	151.824409865	162	none	none	TRUE	53.7069815921
4	KCNIP4	113.387361925	1220183	none	none	TRUE	53.6955353305
5	CTD-2013M15.1	55.2624545862	5754	none	none	TRUE	53.62845014524
10	ATAD1	320.842579514	89831	none	none	TRUE	53.596404731
6	IRF4	179.856520698	19708	cgc	none	FALSE	53.192906625
4	FSTL5	177.387307186	780138	none	none	TRUE	53.147911598
13	RNF219-AS1	157.920210319	697639	none	none	TRUE	53.141056248

Supplemented_Table 6: Potential synthetic lethality partners of PTEN deletions

chr	gene_name	cgc	RA_p	RA_OR	sample_size	RA_PTEN_mutations	RA_gene_mutation	RA_overlaps	RA gene mutations in indications of interest	Of interest
3	VHL	TRUE	0,200612135069417	0,454867017741067	1741	105	69	2	69	no
X	MAGEC1	FALSE	0,680514818576531	0,864379986805011	1741	105	19	1	18	no
X	TM4SF2	FALSE	0,324530039415268	0	1741	105	18	0	16	no
3	SETD2	TRUE	0,367975920289565	0	1741	105	16	0	16	no
X	MAGEC3	FALSE	0,391811498099808	0	1741	105	15	0	15	no
X	MED12	TRUE	0,391811498099808	0	1741	105	15	0	12	potentially
3	BAP1	TRUE	0,444164539064048	0	1741	105	13	0	12	potentially
X	HEPH	FALSE	0,817492492156675	1,30105971367632	1741	105	13	1	12	no
5	CHD1	FALSE	0,503437531001216	0	1741	105	11	0	11	
17	ZNF18	FALSE	0,324530039415268	0	1741	105	18	0	10	no
11	OR4A5	FALSE	0,503437531001216	0	1741	105	11	0	10	no
16	CBFB	TRUE	0,503437531001216	0	1741	105	11	0	10	potentially
X	PTCHD1	FALSE	0,535947334663656	0	1741	105	10	0	9	no
X	DACH2	FALSE	0,535947334663656	0	1741	105	10	0	9	no
X	PABPC5	FALSE	0,535947334663656	0	1741	105	10	0	9	no

5	GPR98	FALSE	0,570535208136111	0	1741	105	9	0	9	potentially
9	C9orf66	FALSE	0,535947334663656	0	1741	105	10	0	8	no
X	WNK3	FALSE	0,535947334663656	0	1741	105	10	0	8	potentially
X	PHKA1	FALSE	0,535947334663656	0	1741	105	10	0	8	no
X	SLITRK2	FALSE	0,535947334663656	0	1741	105	10	0	8	no
X	AFF2	FALSE	0,535947334663656	0	1741	105	10	0	8	no
3	ROBO2	FALSE	0,570535208136111	0	1741	105	9	0	8	no
10	WDFY4	FALSE	0,570535208136111	0	1741	105	9	0	8	no
X	SMARCA1	FALSE	0,570535208136111	0	1741	105	9	0	8	yes
X	PASD1	FALSE	0,570535208136111	0	1741	105	9	0	8	no-potentially
6	PARK2	FALSE	0,417175374363976	0	1741	105	14	0	7	no
X	ZMYM3	FALSE	0,472882073917327	0	1741	105	12	0	7	potentially
17	PIK3R5	FALSE	0,503437531001216	0	1741	105	11	0	7	no
13	CYSLTR2	FALSE	0,535947334663656	0	1741	105	10	0	7	yes
17	MYH1	FALSE	0,535947334663656	0	1741	105	10	0	7	no
17	NCOR1	TRUE	0,535947334663656	0	1741	105	10	0	7	
X	HDX	FALSE	0,535947334663656	0	1741	105	10	0	7	no
11	OR4C46	FALSE	0,570535208136111	0	1741	105	9	0	7	no
13	FNDC3A	FALSE	0,570535208136111	0	1741	105	9	0	7	no
11	ATM	TRUE	0,607332626351278	0	1741	105	8	0	7	
17	DOC2B	FALSE	0,607332626351278	0	1741	105	8	0	7	no

17	MYH2	FALSE	0,607332626351278	0	1741	105	8	0	7	no
22	MYO18B	FALSE	0,607332626351278	0	1741	105	8	0	7	no
3	DNAH12	FALSE	0,646479296558083	0	1741	105	7	0	7	no
10	C10orf112	FALSE	0,646479296558083	0	1741	105	7	0	7	no
X	BTK	TRUE	0,646479296558083	0	1741	105	7	0	7	
4	CCSER1	FALSE	0,503437531001216	0	1741	105	11	0	6	no
X	STS	FALSE	0,535947334663656	0	1741	105	10	0	6	potentially
8	SGCZ	FALSE	0,570535208136111	0	1741	105	9	0	6	no
17	MYH10	FALSE	0,570535208136111	0	1741	105	9	0	6	no
11	MUC5B	FALSE	0,607332626351278	0	1741	105	8	0	6	
17	MYO15A	FALSE	0,607332626351278	0	1741	105	8	0	6	no
5	SLCO4C1	FALSE	0,646479296558083	0	1741	105	7	0	6	no
X	RPS6KA6	FALSE	0,646479296558083	0	1741	105	7	0	6	
X	PCDH11X	FALSE	0,646479296558083	0	1741	105	7	0	6	no
2	XIRP2	FALSE	0,688123668422253	0	1741	105	6	0	6	
8	RP1L1	FALSE	0,688123668422253	0	1741	105	6	0	6	no
X	AMMECR1	FALSE	0,688123668422253	0	1741	105	6	0	6	no
X	BRS3	FALSE	0,688123668422253	0	1741	105	6	0	6	no
17	DNAH2	FALSE	0,503437531001216	0	1741	105	11	0	5	
21	POTED	FALSE	0,535947334663656	0	1741	105	10	0	5	
1	HSPG2	FALSE	0,570535208136111	0	1741	105	9	0	5	
9	NFIB	TRUE	0,570535208136111	0	1741	105	9	0	5	

10	SORCS3	FALSE	0,607332626351278	0	1741	105	8	0	5	
X	GRPR	FALSE	0,607332626351278	0	1741	105	8	0	5	
9	JAK2	TRUE	0,646479296558083	0	1741	105	7	0	5	
9	BNC2	FALSE	0,646479296558083	0	1741	105	7	0	5	
10	KNDC1	FALSE	0,646479296558083	0	1741	105	7	0	5	
13	MLNR	FALSE	0,646479296558083	0	1741	105	7	0	5	
13	CDADC1	FALSE	0,646479296558083	0	1741	105	7	0	5	
17	EIF4A1	FALSE	0,646479296558083	0	1741	105	7	0	5	
17	KCNJ12	FALSE	0,646479296558083	0	1741	105	7	0	5	
21	TMPRSS2	TRUE	0,646479296558083	0	1741	105	7	0	5	
22	CELSR1	FALSE	0,646479296558083	0	1741	105	7	0	5	
X	SATL1	FALSE	0,646479296558083	0	1741	105	7	0	5	
X	CPXCR1	FALSE	0,646479296558083	0	1741	105	7	0	5	