Title

Classifying Cancer Genomic Alterations Using Machine Learning and Multi-Omic Data

Author

Haan, David

Publication Date

2019

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CLASSIFYING CANCER GENOMIC ALTERATIONS USING MACHINE LEARNING AND MULTI-OMIC DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

David Haan

September 2019

The Dissertation of David Haan
is approved:

Professor Josh Stuart, Chair

Professor Angela Brooks

Professor Christopher Benz

Dean Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Classifying Cancer Genomic Alterations Using Machine Learning and Multi-Omic Data

by

David Haan

In 2018, an estimated 1,762,450 new cases of cancer will be diagnosed in the United States and 606,880 people will die from these diseases. Cancer is a group of diseases characterized by the overgrowth of abnormal cells as the result of genomic mutations. Mutations that initiate tumorigenesis are called driver mutations whereas those which cannot are called passenger mutations. Driver mutations define a tumor's subtype and can be used as therapeutic targets thus, deciphering driver mutations from passenger mutations is of utmost importance as we strive to improve cancer treatment. As the cost of genome sequencing is decreasing, the amount of available tumor data is increasing, making it possible to conduct large scale computational analysis with machine learning to identify novel tumor characteristics. There have been numerous recent collaborations to collect, sequence, and analyze human tumors. The largest of these collaborations, the Cancer Genome Atlas (TCGA), is a comprehensive analysis of 9,000 patients and 33 sub types cataloging mutation data, DNA, mRNA, methylation, and protein expression. Whereas the TCGA is mostly whole exome sequencing, the International Cancer Genome Consortium (ICGC) has begun contributing data from the whole genome sequencing of a few thousand tumors. Using both the TCGA and ICGC data, I performed four new variant classification analyses using both unsupervised machine learning techniques and a novel supervised machine learning technique to identify tumor subtypes, driver mutations and potential therapeutic targets. I first present an analysis in which I used supervised machine learning to determine the most important genomic features responsible for accurate gene fusion detection among a set of fusion detection methods. Next, I present a method of unsupervised machine learning in which I classify non-coding variants of splicing factors as potential driver mutations in a number of tumor types. Third, I analyze telomere data from ICGC whole genome sequencing data using unsupervised machine learning to identify 4 subtypes of telomere maintenance mechanisms(TMM) among 2,500 tumor samples. Lastly, I present a new variant classification method called LURE, which uses supervised machine learning to classify variants based on existing signatures from known driver mutations.

Chapter 1

Introduction

Approximately 39.3 percent of men and women will be diagnosed with cancer of any site at some point during their lifetime[2]. In 2016, there were an estimated 15,338,988 people living with cancer of any site in the United States[2].

Cancer is a genetic disease caused by mutation and selection in somatic cells. Mutations in normal cells are usually repaired or result in apoptosis. Whereas in cancer cells mutations accumulate, leading to uncontrolled growth and tumorigenesis. There are two broadly defined types of mutations: drivers and passengers. Tumors contain around 2-5 driver mutations that cause and accelerate cancer, and about 10-200 passenger mutations which are accidental byproducts of thwarted DNA repair mechanisms[3]. Driver mutations define key characteristics of the tumor and may offer therapeutic targets, yet identifying them amid the myriad of passengers remains a challenge.

The Cancer Genome Atlas (TCGA) is a publicly accessible dataset of cancer samples from the National Cancer Institute (NCI)[**?**]. TCGA catalogues mutations, mRNA, miRNA, DNA methylation, copy number variation, and protein expression data for roughly 11,000 patients across 33 cancer types[**?**]. The identification of driver mutations played a fundamental role in many TCGA analyses. Whereas the TCGA is mostly whole exome sequencing, the International Cancer Genome Consortium (ICGC) has begun contributing data from the whole genome sequencing of a few thousand tumors. Using both the TCGA and ICGC data, I performed four new variant classification analyses using both unsupervised machine learning techniques and a novel supervised machine learning technique to identify tumor subtypes, driver mutations and potential therapeutic targets.

There are two types of machine learning methods: supervised and unsupervised. Unsupervised learning methods, such as hierarchical clustering and T-distributed Stochastic Neighbor Embedding(t-SNE), are effective at visualizing high dimensional data and identifying clusters or patterns in the data[4]. Supervised machine learning uses prior knowledge or labels to divide data into two or more classes[4]. By training a classification model on prior knowledge, supervised machine learning can make predictions about unknown or unlabeled data[4]. In this thesis, I present unsupervised and supervised machine learning methods I performed on cancer genomic data to classify mutations or variations found in the genomes of human tumors.

In chapter 3, I recapitulate my participation in a bioinformatics community challenge ranking user-submitted gene fusion detection methods. In this challenge, we spiked gene fusion transcripts into a few human cancer cell lines and asked participants to run their own gene fusion detection methods on these datasets. I was tasked with performing analysis on the data to determine if the presence of certain genomic features near a fusion prevented methods from detecting the fusion. After collecting a set of genomic features for each fusion, I used a random forest, a supervised machine learning method, to predict the false positive and false negative fusion events called by each method. I then preformed a feature importance analysis to identify the genomic features responsible for each method's incorrect fusion detection.

Next, I sought to classify mutation variants by developing a supervised machine learning tool called Learning UnRealized Events (LURE). LURE, discussed in chapter 4, associates alterations between samples by finding similar signatures in feature data, such as mRNA expression data or methylation data. LURE achieves this by training a classifier on tumor sample feature data of a known driver mutation (the bait), applying the classifier to find ”bait”-absent samples with a high classifier score, and identifying other alterations (the ”catch”) in those samples that correlate with the high classifier score. Using LURE, I identify new associations across pan-cancer data and find putative new driver alterations involved with MAPK signaling. In addition, I associate new driver alterations with the alternative lengthening of telomeres (ALT), a telomere maintenance mechanism (TMM), in sarcomas.

In chapter 5, I discuss variant classification of non-coding mutations using supervised learning. First, using hierarchical clustering of pathway expression data, I relate non-coding mutations of splicing factors with coding mutations of the same splicing factors. Next, using t-SNE on telomere related features, I identify 4 unique groups of tumor samples that employ different TMMs and associate mutations with each of these groups. In chapter 6, I outline conclusions based on my research and discuss the future directions of cancer genomics.

Chapter 2

Background

2.1 Cancer is a Genetic Disease

Cancer is a genetic disease typically resulting from an accumulation of mutations[3]. Mutations in normal cells generally result in repair or cell suicide. However, in cancer cells, the mutations accumulate leading to an uncontrolled growth otherwise known as a tumor. There are two broadly defined types of mutations, driver and passenger mutations. Tumors contain around 2-5 driver mutations which cause and accelerate cancer, and about 10-200 passenger mutations that are accidental byproducts of thwarted DNA repair mechanism. Driver mutations define a tumor's subtype and can be used as therapeutic targets.

2.2 The Cancer Genome Atlas

TCGA (cancergenome.nih.gov) is a publicly accessible collection of data from the NCI[5]. This atlas of data is a comprehensive analysis of 9,000 patients and 33 cancer subtypes cataloging mutation data, DNA, mRNA, methylation, and protein expression[5]. The majority of the TCGA data is whole exome sequencing that covers only protein coding regions of the genome[5]. As the cost of sequencing is decreasing, whole genome sequencing of numerous and diverse patient samples has become more practical. This has enabled the creation of the PanCancer Analysis of Whole Genomes (PCAWG), a data base of 2,500 patient-derived tumor sequencing profiles of various cancer subtypes[6].

2.3 Mutation Types

There are numerous types of mutations and other genetic alterations that directly initiate tumorigenesis[3]. Single nucleotide variations (SNVs) result from the insertion, deletion or subsitution of a single nucleotide and are the most common genomic mutations. SNVs in non-coding regions of the genome can result in aberrant activity of promoters, enhancers, and silencers. SNVs in coding regions are typically single nucleotide substitutions that change a three-base amino acid codon sequence resulting in either the production of the same amino acid (silent mutation), a different amino acid (missense mutation), or a stop codon (truncating mutation). Missense and truncating mutations are more deleterious because they change the amino acid composition of a protein. Insertions and deletions of nulceotides in the coding sequences can also affect protein composition by causing a frameshift mutation that typically results in an early stop codon and thus, a truncated protein product. Copy number alterations are large deletions or amplifications of a genomic region that result in either decreased or excessive protein production. Gene fusions occur when two genes at the DNA level are joined through a translocation, interstitial deletion, or chromosomal inversion and produce a chimeric protein.

2.4 Machine Learning

There are two types of machine learning: unsupervised and supervised. Unsupervised machine learning is a method that finds new patterns in a dataset without pre-existing labels and is typically used in cancer genomics to identify subtypes or clusters of tumors that share a similar set of features. Supervised learning is used when a computer algorithm cannot be used to solve a given problem without implementing example data or previous experiences[7]. A supervised machine learning method builds a model by defining a set of features and labels for each data point or observation in an example dataset. There are many types of supervised machine learning such as linear regression, logistic regression, random forest and neural networks.[7]. To illustrate the utility of supervised machine learning, consider an example in which a logistic regression model is used to distinguish between different species of flowers. The length and width of the sepals and petals may be considered flower features and the flower species (a label) is assigned to each flower based on their specific features. The model is trained on this existing data, known as training data, and can be used to further predict the species of a new flower given its sepal and petal measurements. In this thesis, I present my work based on similar principles of machine learning to determine a tumor's mutation status based on gene expression or methylation data and determine which features are most important for tumor classification.

2.5 Modeling Imbalanced datasets

Biological data is often imbalanced, leading to difficulty in building accurate classification models. In cancer genomic data, there is more data collected from tumors of particular subtypes, usually those subtypes exhibiting well-studied genetic mutations, like PTEN or BRCA. This leads to misleading analysis results when metrics are not weighed per class. For example, when using such data to classify multiple different tumor subtypes, a model may achieve better overall accuracy by classifying both the larger and smaller classes as the larger class. Thus, it is imperative that the correct accuracy metric is chosen to avoid a false depiction of good accuracy. The Receiver Operating Characteristic Area Under the Curve (ROC AUC) is a very popular metric but is highly inaccurate for imbalanced models because it relies on the true positive rate and false positive rate, the later being weighted for the negative class[8]. Overall, this results in the inaccurate measurement of false positives. To further explore the cause of this defect, imagine classifying a dataset with a 5-95 (true-false) split. If the classifier guessed all 5 true positives as positive but classified 5 false positives as positive as well, then the false positive rate for the ROC curve would be 5/95, or 0.0526, which is very low. Although only 0.0526 of the negative class was incorrectly classified, equal numbers of samples were guessed wrong as were guessed right therefore, this classification is unreliable. Usually in the case of imbalanced classes, there is more concern with how well the classifier can identify the smaller class or classes. When trying to model a rare genetic mutation, accurately determining the mutated samples is crucial for analysis while the negative class is of little interest. In this situation, it is best to use precision and recall to measure accuracy of imbalanced classification models because they are weighted for only the positive class. Precision is a measure of the false positives that is weighted for the positive class. This quantifies how often a model incorrectly guessed a negative label as a positive per the number of true positive samples. In the example above where the false positive rate was only 0.0526, the precision is 0.5 (false positives/(correct true positives+false positives)). With a precision of only 0.5, it would most likely be necessary to discredit the classification model. Recall is the true positive rate; it measures how often the model accurately identifies the positive labels as positive per the number of true positives. Together, these two metrics can be used to create either an F1 score, which is the harmonic mean of precision and recall, or a precision-recall area under the curve(PR AUC)[8]. When classifying imbalanced datasets, precision-recall AUC or F1 scores can be used to add a description of the model's accuracy.

2.6 Driver Discovery Tools

There are several existing computational tools that try to decipher driver from passenger mutations[9]. EPoC uses network modeling of the transcriptional effects of copy number aberrations to identify driver mutations in glioblastoma (GBM)[10]. DriverNet employs a probabilistic model to locate driver mutations using transcriptional networks[11]. These methods can predict novel drivers given a set of SNVs or copy number alterations and the corresponding mRNA gene expression data. In addition, there are methods that identify modules of driver genes based on mutual exclusivity in certain tumor types, such as CoMEt[12] and MEMo, the latter of which incorporates prior knowledge such as pathway data into driver gene module discovery[13].

Chapter 3

Accurate Fusion Detection

3.1 Introduction

Genomic rearrangements in cancer cells produce fusion transcripts that may give rise to protein products not present in normal cells. These can serve as robust diagnostic markers, e.g. TMPRSS2-ERG in prostate cancer[14] or drug targets, e.g. BCR-ABL1 in chronic myeloid leukemia[15]. Ongoing research efforts are beginning to unveil the potential clinical relevance of aberrant processing of RNA in cancer, such as defects in alternative-splicing. An assortment of computational methods are needed to fully document the transcriptomic differences between tumor cells and their normal counterparts. Increasing the alterome of tumors by fully characterizing their RNA landscapes will expand our understanding of cancer mechanisms, provide new biomarkers and reveal possible new RNA-based therapeutics, improving personalized patient treatment.

Gene fusions occur when two genes at the DNA level are joined through a translocation, interstitial deletion, or chromosomal inversion. A fusion may also occur at the RNA level, resulting from a ligation between two transcripts. Gene fusions often play an important role in the initial steps of tumorigenesis. Specifically, gene fusions have been found to be the driver mutations in neoplasia and have been linked to various tumour subtypes. An increasing number of gene fusions are being recognized as important diagnostic and prognostic parameters in malignant haematological disorders and childhood sarcomas. Reviews have estimated that gene fusions occur in all malignancies and that 20% of human cancer cases harbour at least one to RNA fusion events[16, 17].

The goal of this challenge is to use a crowd-based competition to identify the optimal methods for quantifying isoforms and detecting mRNA fusions from RNA-seq data. Several methods have been published that detect and quantify cancer-associated RNA abundance species. Yet, it is not clear which methods are best used and under what contexts. Recent systematic comparisons have been performed[18, 19] to evaluate RNA-Seq analysis methods. Most comparisons have been performed by an author of one of the competing methods and so may suffer self assessment bias. One of the more recent evaluations, performed by an impartial list, was the study by Kumar et al 2016 that compared 12 different methods based on their accuracy, length of execution time, and memory requirements. The work performed by the authors in Creason, et al, includes several newly developed tools, the use of spike-ins for an unbiased assessment of sensitivity, an objective evaluation framework in which the administrators ran submitted methods to generate all predictions, and a statistical procedure to infer background fusions to accurately measure precision. In addition to a new evaluation of methods, our work provides a tool for simulating RNA isoforms and fusions, a new benchmark dataset against which forthcoming methods can be compared, and all of the tested methods in standardized workflow for re-execution, which should further progress this area of study. My contribution to the community challenge was to rank the isoform challenge and run a feature importance pipeline on the fusion detection methods.

3.2 Dream Challenge Results

The fusion detection evaluation challenge received 63 entries, of which the organizers were able to run 37. The final evaluation consisted of 17 entries, which included multiple submissions by the same team (up to three per team allowed). Two different datasets were created to evaluate methods-a computationally simulated dataset and another experimentally generated set using spike-ins (Figure 2a). The simulated dataset was used to evacuate the preliminary rounds. The simulated data were generated with the program rnaseqSim (in preparation) that created synthetic reads from computationally constructed fusions. On average, the simulated tumor samples contained 39 fusions per transcriptome. The second evaluation dataset of spike-ins was used for the final evaluation of methods. The spike-in data were created in the lab using a prescribed series of fusion products spiked in to several cancer cell lines, using 18 different fusion constructs. Scoring of the fusion methods was performed using the F1 score against the spike-in set. To account for fusions found in the cell line background, fusions called by multiple methods across multiple replicates were collected for PCR based validation, and utilized as an imputed truth set to augment the spike-in set (see methods). An additional F1 score was calculated for the combined (spike-in plus imputed) truth set. All submitted entries were ranked according to their F1 score on the combined data. Two of the submitted methods emerged as the overall winners of this sub-challenge – Arriba and STAR-Fusion – based on their performance in spike-in benchmarks with the addition of the imputed truth set. Arriba achieved an F1 of 0.73, and STAR-Fusion had an F1 of 0.70. The next closest entry, another permutation of STAR-Fusion, was at 0.63. Based on a bootstrap analysis, no other methods were found to have achieved results as accurate as these top two entries. The next method, not submitted by one of the winning teams was fusioncatcher, submitted by the challenge administrators, at 0.58. Finally, the highest scoring method, not submitted by the top two teams or the administrators, was STAR-SEQR with an F1 of 0.47.

3.3 Features Influencing the Accuracy of Fusion Detection Methods

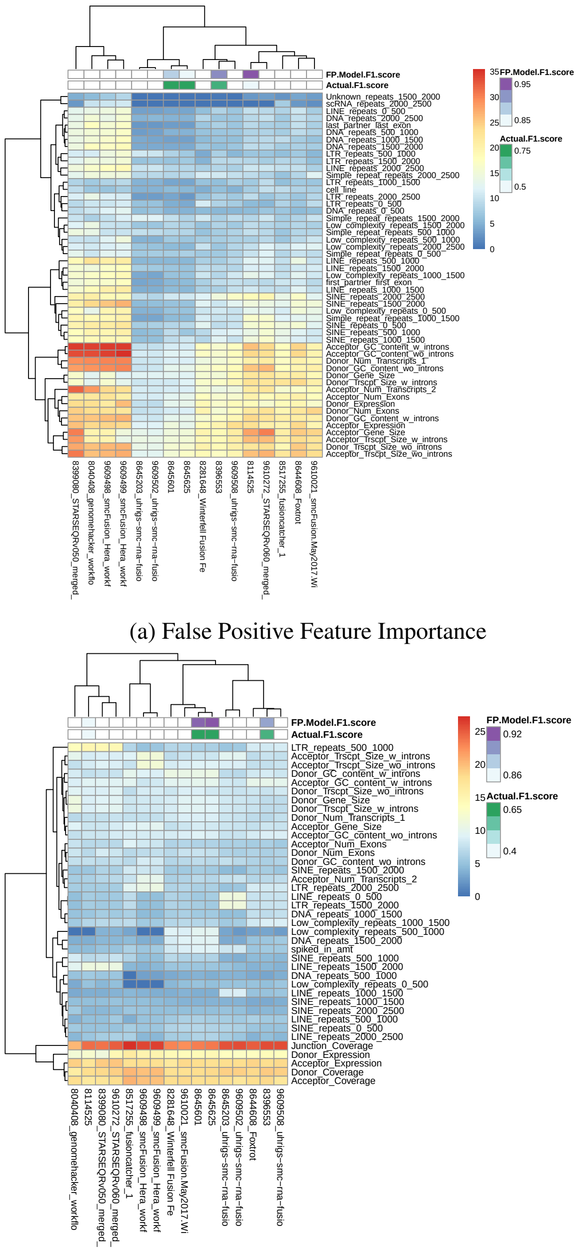
In an effort to determine what factors influence methods to incorrectly call fusion events, I created a fusion feature importance pipeline, similar to what was done for the SMC-DNA challenge[20]. To start, I collected 128 genomic features for each predicted fusion event, including gene length, transcript length, distance from the breakpoint to repeats, and the abundance for each fusion partner. Next, I built a random forest (RF) classification model to predict the false-positive fusion events from each submission. The RF was trained to select features that predict when a method erroneously calls a fusion event when no such event was present according to the i-truth. Also, I built a second RF model to select features that predict false negative events; i.e. the RF predicts when a method fails to detect a spiked-in fusion construct.

To determine feature importance among our classification models I applied our random forest models to the algorithm Boruta, an all relevant feature selection algorithm[21][22](Figure 3.1).

Boruta determines feature relevance by comparing the original importance with the importance achievable at random, estimated using permuted versions of a feature, and progressively eliminating irrelevant features to stabilise the test. In order to determine the features most relevant to all the methods, Iincluded the submission ID as a feature and created one false positive classification model achieving an out-of-bag (OOB) error rate of only 0.26%. The false negative RF model was more difficult to predict due to fewer observations, but achieved an error rate of 7.64%. The Boruta algorithm revealed that the number of transcripts and GC content were the most important features for determining the false positives among all fusion methods whereas submission id, coverage across junction and expression were the top features for the false negative model(Figure 3.2). Further analysis of the top features for the false negative model revealed a marked decrease in coverage and expression for the false negative fusions(Figure 3.3).

3.4 Conclusion

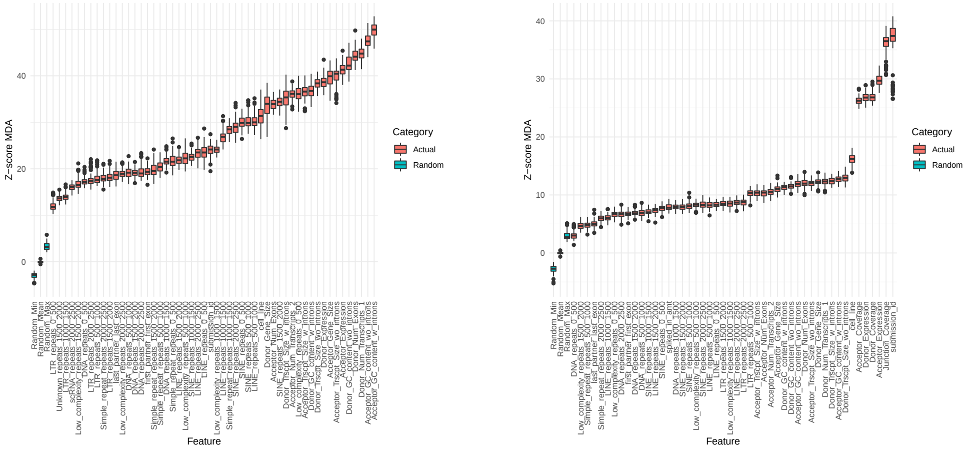
Here I presented some of the results of the SMC-RNA DREAM challenge in which we synthetically introduced fusions and asked contestants to run fusion detection methods and report the results. Through my feature importance analysis I determined that coverage, expression, GC content, and the number of transcripts of each fusion partners are the most important features. Identifying alterations in tumor sequence data is very important, but even more so is



(a) False Positive Feature Importance

(b) False Negative Feature Importance

Figure 3.1: **Boruta Feature Importance Analysis By Fusion Submission.** Heatmap showing results from performing the Boruta algorithm on each submissions false positive fusion events(A) and false negative fusion events(B). Each cell in the heatmap represents the Z-score Mean Decrease in Accuracy. Higher Z-scores are in red and represent more important features. Rows are the fusion submission names and columns are the features. Only features which had a mean value greater than Borutas shadow max value are shown.



(a) False Positive Feature Importance

(b) False Negative Feature Importance

Figure 3.2: **Boruta Feature Importance Analysis Across All Fusion Submissions for the False Positive Model(A) and False Negative Model(B).** Boxplot showing results from performing the Boruta Algorithm on all fusion submissions. The y-axis represents the Z-score MDA and features are across the x-axis. The red plots are the Z-scores of the actual features and blue are Borutas shadow features which are considered the randomized background features. Only features which performed significantly better ($p < .05$) than the shadow features are shown in this plot.

determining if these alterations would be considered drivers and contribute to tumorigenesis. Is it possible that a fusion event is capable of driving cancer? What other events could be driving cancer and how can we identify them?

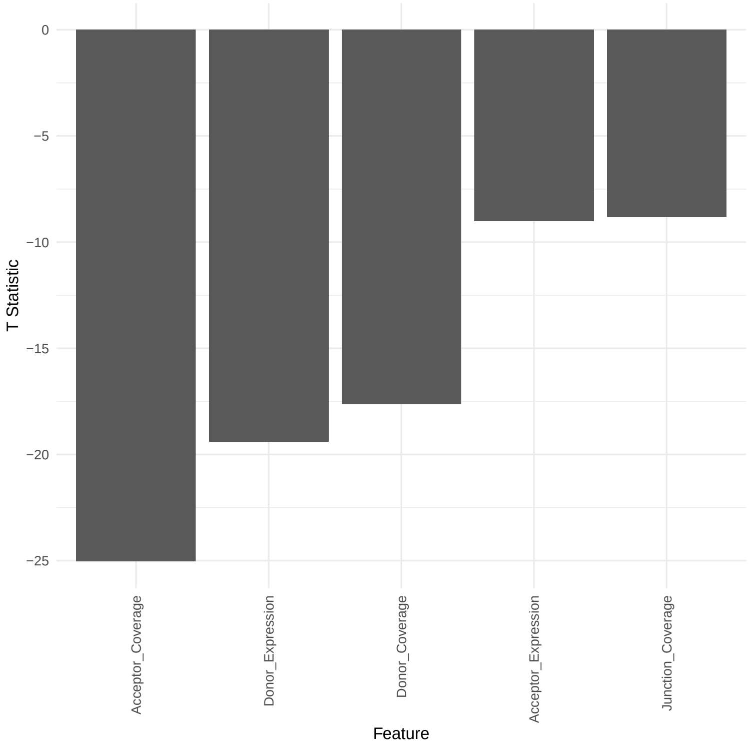


Figure 3.3: **False Negative Feature Analysis T Statistic.** Students t-test of top features identified by False Negative Random Forest model. Students t-test was performed individually on the 5 features comparing false negative fusions missed by the submission methods to the accurately identified fusions. A negative t statistic represents a decrease in the feature values for false negative fusions.

Chapter 4

LURE: Classifying Coding Variants of Unknown Significance

LURE (Learning UnRealized Events): Finding New(or Equivalent) Driver Mutation Events using Supervised Machine Learning

4.1 Introduction/Background

There are several existing computational tools that try to decipher driver from passenger mutations[9]. EPoC uses network modeling of the transcriptional effects of copy number aberrations to identify driver mutations in glioblastoma (GBM)[10]. DriverNet employs a probabilistic model to locate driver mutations using transcriptional networks[11]. These methods can predict novel drivers given a set of SNVs or copy number alterations and the corresponding mRNA gene expression data. In addition, there are methods that identify modules of driver genes based on mutual exclusivity in certain tumor types, such as CoMEt[12] and MEMo, the latter of which incorporates prior knowledge such as pathway data into driver gene module discovery[13]. In contrast, LURE uses mRNA data to identify mutations in “driver-unknown” samples with similar expression signatures to known drivers, thereby implicating a novel set of mutations as possible drivers.

Several studies have built gene expression signatures to identify samples with certain driver events. For example, studies have identified a TP53 gene expression signature as a reliable and independent predictor of disease outcome in breast cancer[23, 24]. In addition, in patients with epithelial ovarian cancer, a BRCAness gene expression signature is just as predictive of chemotherapy responsiveness and outcome as mutation status.[25] While creating signatures as prognostic markers to guide treatment is important in a clinical setting, there has been little work using such gene expression signatures to find related mutational events. LURE identifies gene expression signatures across the 723 COSMIC cancer genes[26] and then uses iterative semi-supervised learning to discover potentially related events.

A similar method, REVEALER is a computational method that identifies combinations of genomic alterations correlated with functional phenotypes, such as the activation or gene dependency of oncogenic pathways or sensitivity to a drug treatment[27]. While the concept of REVEALER is very similar to LURE, there are few differences which make LURE better. For one, at every iteration of the process LURE produces a new classification model slightly more accurate than the previous as the newly discovered events are no longer false positives and now aide in determining features relevant to the signature. Second, REVEALER utilizes a mutual exclusive relationship between new events which may limit results as mutation calls are not 100% accurate. Allowing some overlap between predicted events LURE can account for possible mutation call errors and identify modules containing co mutated events. In an effort to compare methods, my undergraduate mentee, Ruikang Tao ran REVEALER using one of our test sets described in our positive controls and REVEALER was unable to identify the leave-out test sets (Supplemental Figure 4.1).

4.2 Method

LURE attempts to associate alterations between samples by finding similar signatures in feature data such as mRNA expression data. LURE achieves this by training a classifier using the samples of a known driver mutation (the bait) applying it to find ”bait”-absent samples with a high classifier score, and looking for other alterations (the ”catch”) in those samples that correlate with the high classifier score.

The first step (Step 0) is to establish a known driver mutation as the initial bait, for example event A (Figure 4.2A,E). In the next step (Step 1) LURE trains a logistic regression classification model using gene expression as features and bait mutation status as the label to be predicted. A cross-validation is run for the classification task and baits can be filtered by model performance, e.g. area under the precision recall curve (PR AUC). LURE then uses the classification model to score each sample in the dataset (Figure 4.2A,E). Notwithstanding the inherent bias towards overfitting, there might be some negative samples not mutated in the bait gene that still receive a high classifier score. These false positives from the classification task show the same expression signal as the bait-mutated samples hinting to a different driver event having a similar effect on the cancer cells.

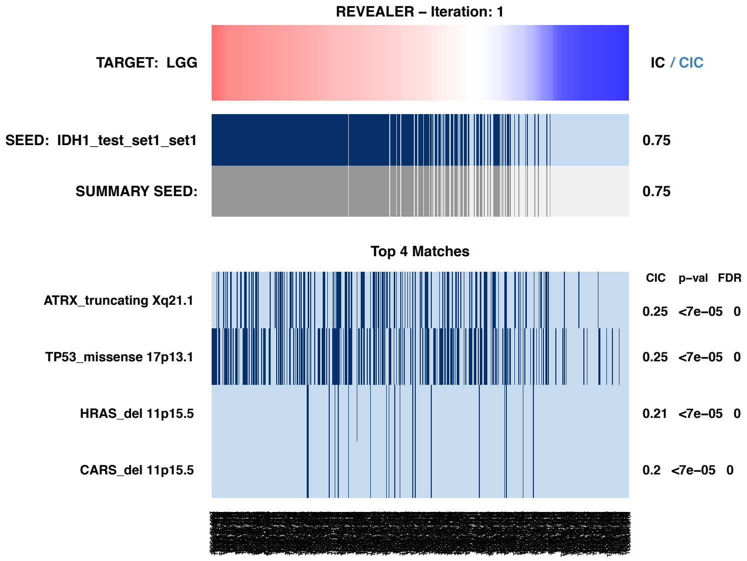


Figure 4.1: **REVEALER results of using IDH1 positive control test sets.** Each column represents the samples in the TCGA Lower Grade Glioma (LGG) dataset. The top row is the classifier score assigned to each sample using an initial classifier trained on the first IDH1 test set of 150 samples. The tick marks in the seed row represent the 150 IDH1 test set samples. REVEALER identified 4 matches, ATRX truncating mutations, TP53 missense mutations, and a 11p15.5 focal deletion. While these are not invalid results as it has correctly found genes co-mutated with IDH1, it was unable to find the 60 left out IDH1 mutants.

Therefore, in Step 2, LURE takes the false positive samples and the rest of the negative samples and runs SSEA (Sample Set Enrichment Analysis). Our SSEA runs the GSEAPreranked tool in which the mutation status forms sample sets and the classification score is the sample ranking. SSEA tests each mutation event if the samples having that mutation are associated with the false positive samples (Figure 4.2A). The events with significant SSEA association are called ”catch” events. For each catch event, LURE then combines the positive samples for both catch and bait event into a new, intermediate bait event and trains a new classifier for these samples (Figure 4.2B,E). A cross-validation is run for the new classification model and the PR AUC results are compared to the initial classifier to ensure the model improves when including the new positive samples (Student's t-test t-statistic > 0). In addition, the new classifier has to outperform a null model background distribution by adding the same number of randomly chosen catch samples to the true positives and running a cross-validation (Student's t-test p-value $< .05$). After establishing that the new additional event both improves the original classifier and significantly outperforms a random background distribution, the new classifier is run on all samples again (Figure 4.2B,E) in an effort to search for the next set of catch events. This iterative event discovery builds new classifiers by adding one catch event at a time until no further events are found by SSEA or classifier performance is not increased anymore. LURE builds an Event Discovery Tree and recursively returns to the root after exploring all events at each node (Figure 4.2C).

In Step 3 LURE builds a final classification model from the union of all events in the event discovery tree. It then runs a set coverage algorithm between all events in the event discovery tree and the samples predicted to be positive by the final classification model (Figure 4.2D,E). The set coverage algorithm identifies the minimum set of events which cover all the positive samples allowing us to remove completely overlapping mutations. The minimal set of events representing the positive samples is called the ”Catch” Cover.

4.3 Results

4.3.1 TCGA Positive Controls

Splicing Factor 3b Subunit 1 (SF3B1) is a well known splicing factor which is recurrently mutated in many tumor types, including Uveal Melanoma (UVM). Missense mutations in SF3B1 leads to aberrant splicing and a unique gene expression signature[28]. In order to test LURE's ability to discover known 'catch' events, I created a test set of bait and catch events using SF3B1 missense mutations in the TCGA UVM sample set. Of the 80 UVM samples, 18 samples have missense mutations in SF3B1. I created an initial bait out of 8 of those samples and left out 2 sets of 5 SF3B1-mutated samples for discovery. LURE re-discovered both held out sets correctly, collecting all of the SF3B1 missense events in the Catch Cover (Figure 4.3A).

Isocitrate Dehydrogenase 1 (IDH1) is one of three isocitrate dehydrogenase isozymes, which when mutated causes hypermethylation and subsequent altered gene expression in Gliomas[29]. I chose this driver gene as another positive control and created a test set using IDH1 mutations in the TCGA Lower Grade Glioma (LGG) sample set. Of the 210 LGG samples with an IDH1 missense mutation, I created an initial bait with 150 samples and three sets of 20 samples as potential catch events. LURE collected all three of the left out events in the Catch Cover, as well as including the IDH2 missense mutation event. IDH2 is another one of the three IDH

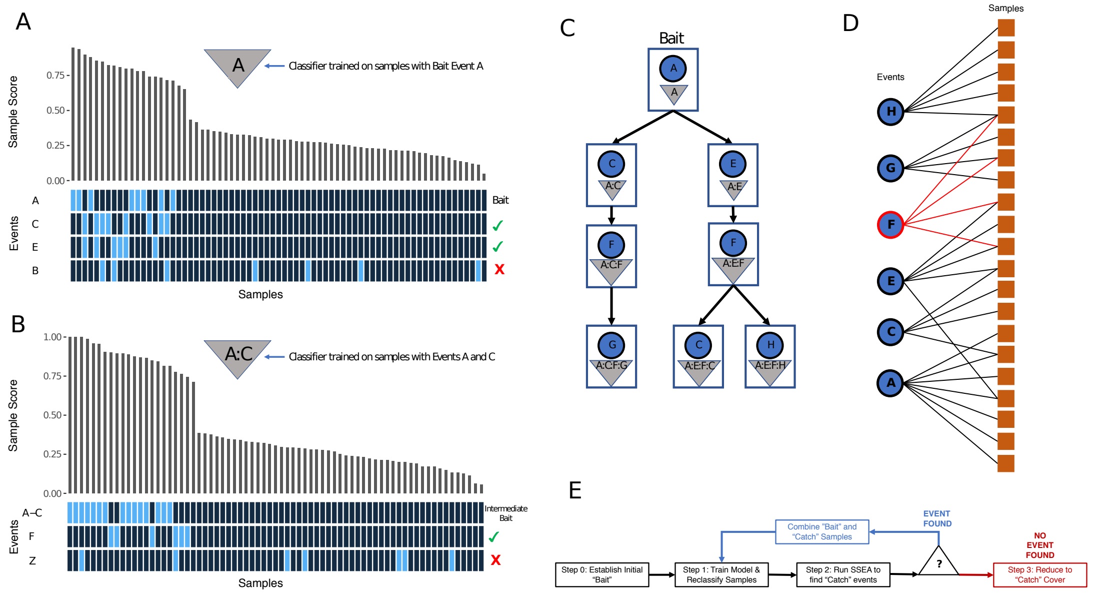

A
Sample Score
0.75
0.50
0.25
0.00
A
Classifier trained on samples with Bait Event A
Events
A
C
E
B
Bait
✓
✓
X
Samples
B
Sample Score
1.00
0.75
0.50
0.25
0.00
A:C
Classifier trained on samples with Events A and C
Events
A–C
F
Z
Intermediate Bait
✓
X
Samples
C
Bait
A
A
C
A:C
E
A:E
F
A:C:F
F
A:E:F
G
A:C:F:G
C
A:E:F:C
H
A:E:F:H
D
Samples
Events
H
G
F
E
C
A
E
Step 0: Establish Initial "Bait"
Step 1: Train Model & Reclassify Samples
Step 2: Run SSEA to find "Catch" events
?
EVENT FOUND
Combine "Bait" and "Catch" Samples
NO EVENT FOUND
Step 3: Reduce to "Catch" Cover

Figure 4.2: **LURE Method. (A) LURE Oncoprint of Bait Mutation A.** Triangle symbol represents the classification model for samples having event A. Barplot showing the score given to each sample by the model representing the probability the sample has the mutation event A. The annotation bars below indicate mutations present in each sample. The check mark annotations on the right mark if events passed Sample Set Enrichment Analysis (SSEA, see Supplemental Methods) ($p < .05$, fdr $< .25$). **(B) LURE Oncoprint of Intermediate Bait Mutation A:C.** Results from a classification model containing both events A and C. **(C) LURE Event Discovery Tree.** Directed graph shows LURE's iterative Event Discovery Tree. Each node is a classification model built on the events shown in the triangle symbol within each node. The blue circles within each node represent the newest event added to the model. **(D) LURE Set Cover Algorithm.** Bipartite graph shows the result of running a Set Cover Algorithm on the mutations collected from the Event Discovery Tree and the samples predicted to be mutated in these genes by the final classification model. The red node and edges mark Event F, which is completely overlapping with other mutations and therefore removed from the final event set ('Catch' Cover). **(E) LURE Method Flowchart.**

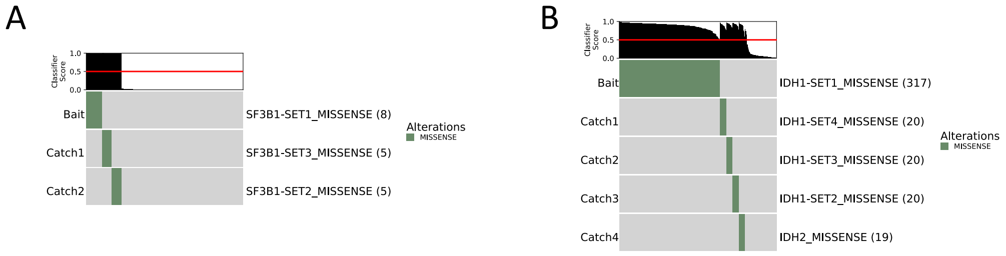


Figure 4.3: **LURE Bait and Catch Oncoprints of positive controls.** Samples are represented across the rows. Colored tick marks represent the different types of alterations present in the samples. Barplot on top shows the LURE classifier score for each sample after the final iteration. **(A) SF3B1 test set in UVM.** Initial bait is set to 8 SF3B1 missense in UVM, and LURE finds the 2 left out sets of 5 each. **(B) IDH1 test set in LGG.** Initial bait is set to 150 samples of IDH1 Missense in LGG and Lure finds the 3 left out sets of 20 each.

isozymes and a mutation has the same oncogenic effect as an IDH1 mutation[30](Figure 4.3B).

4.3.2 LURE on the PANCAN Dataset

In order to look for novel associations between genes already associated with cancer, I decided to run LURE across all tumor types in TCGA, the Pan-Cancer dataset, restricting both baits and catches to mutation events in the 723 COSMIC genes[26]. I created bait events for missense mutations, truncating mutations, homozygous focal point copy number deletions, splice site mutations, and gene fusions and required at least 10 alterations per tumor type. I restricted our classification models to within tumor types, as different tumor types typically have unique expression patterns so unless our mutation status was equally stratified across tumor types our models would simply predict tumor type and not mutation status. By creating baits for different alteration types in the same gene, as opposed to one bait for any alteration in a gene, I are able to identify associations between different alteration types of the same gene as well as identify functional alterations. For example, -thalassemia mental retardation X-linked (ATRX), a gene recurrently mutated in Lower Grade Gliomas (LGG), only has an oncogenic effect with a loss of function mutation, either truncating mutation or copy number loss, whereas a missense mutation may not have an effect[30]. I then trained logistic regression models on the resulting 3,053 bait/tumor type combinations.

I tested both random forest and neural network and found the linear model to consistently score higher in the majority of models (Suppl. Figure 1-Rick's Figure). In order to limit the number of putative false positive results, I restricted the number of bait classifiers by considering only those with PR AUC > 0.5, precision > 0.4, and recall > 0.75 (Supplemental Figure 4.4). Since the objective of LURE is to classify false positives, I were more lenient with the precision and placed more restriction on the recall. Among the bait classifiers passing these thresholds, the most common bait across all tumor types was TP53, and the tumor types with the highest number of passing bait classifiers were Lower Grade Glioma (LGG), Thyroid Carcinoma (THCA), and Prostate Adenocarcinoma (PRAD) (Supplemental Figure 4.6) After trimming the models, I ran LURE with the 81 remaining baits (Figure: 4.5) using missense mutations, truncating mutations, splice site mutations, gene fusions, and focal point copy number amplifications and homozygous deletions of COSMIC genes as possible catches.

LURE found significant catch-associations for 35 of the 81 baits tested. By adding catches to each initial bait event, the classifier PR AUC was increased in varying amounts across the different baits (Supp Figure 4.7A). The most common bait among the 35 were associations with TP53 (Supp Figure 4.7E). Tumor type was evenly distributed and SNVs dominated the

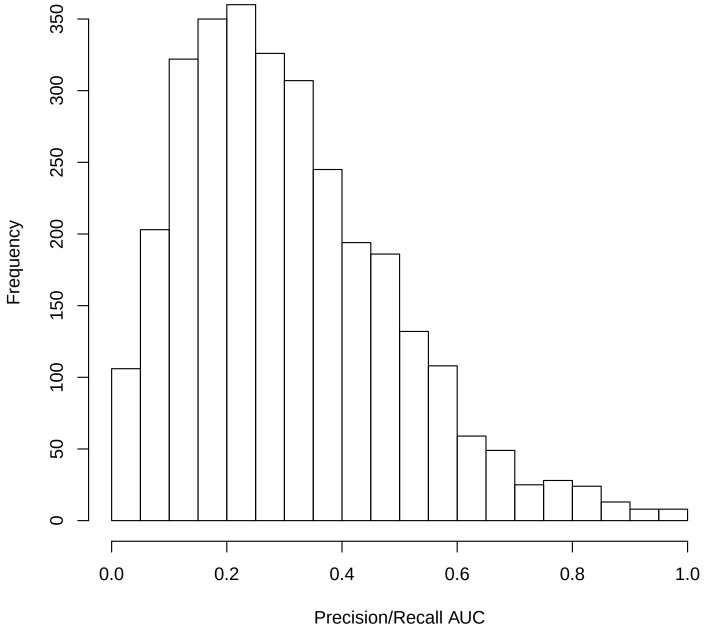


Figure 4.4: **Histogram of Bait Precision-Recall Area Under the Curve**. Plot shows a histogram of the 3,053 Precision Recall AUC cross validation scores from training a tumor type specific logistic regression classifier on the different 723 COSMIC genes. Classifiers were created only in tumor types with more than 10 alterations.

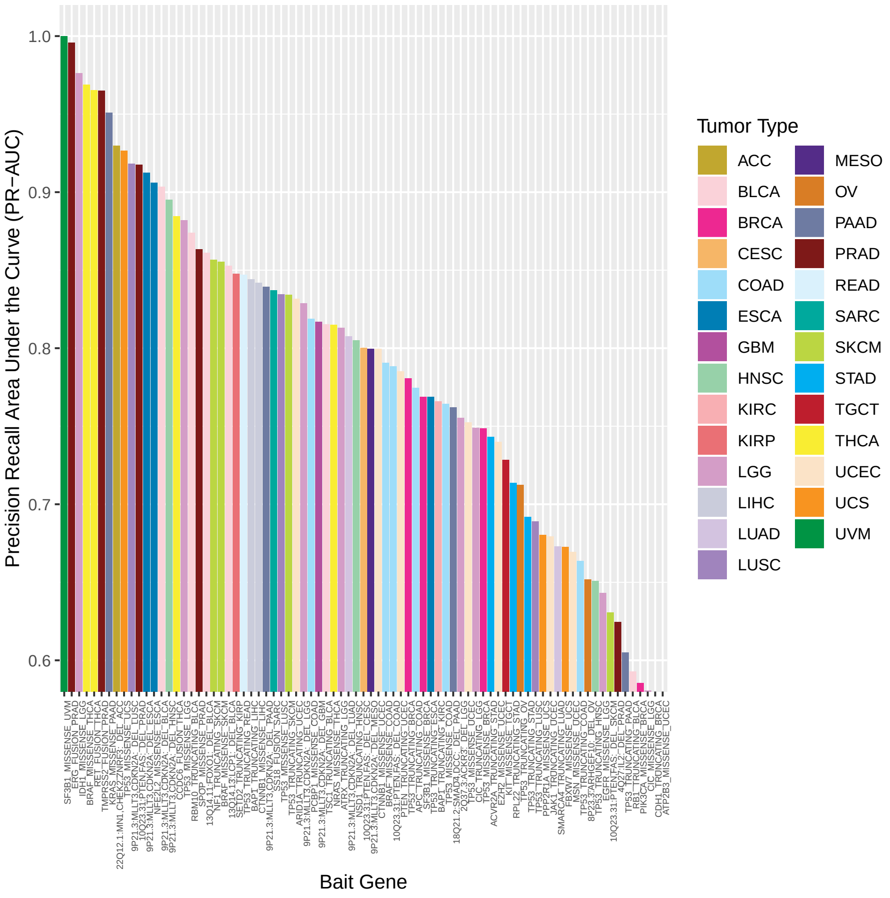


Figure 4.5: **Barplot of 81 Bait Classifiers**. Barplot shows Precision-Recall AUC test scores for 81 bait-tumor type classifiers. Bar is colored by tumor type.

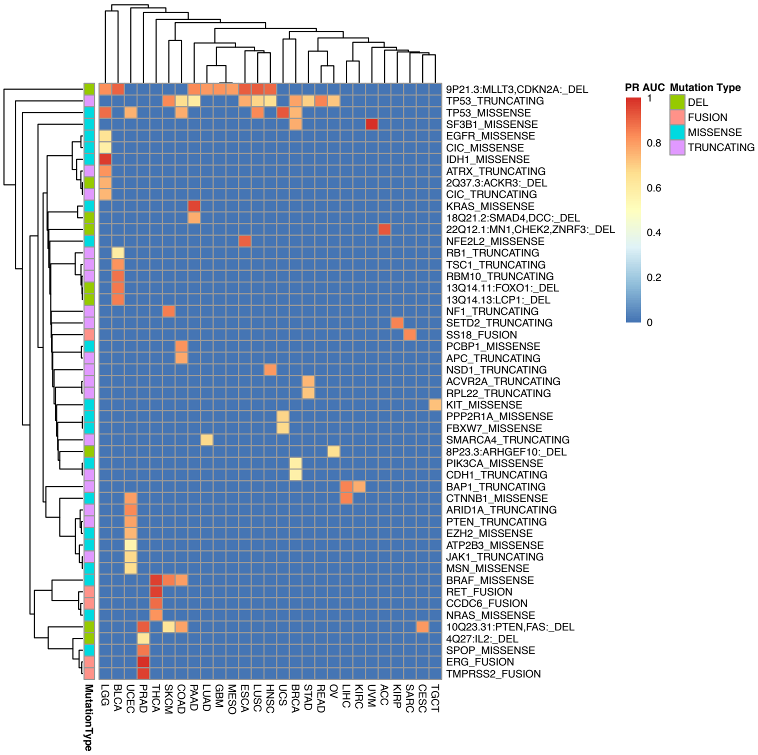


Figure 4.6: **Heatmap of baits used in PANCAN LURE analysis.** Rows are bait and alteration type, and the columns are tumor type. Each cell in the heatmap corresponds to the PR AUC score of that bait/alteration/tumor type classifier. The higher scoring baits are in red/orange and blue denotes no classifier due to lack of alterations in that bait tissue combination. TP53 was the most common bait across tissues and the tumor types with highest number of high scoring baits were LGG, THCA, PRAD.

bait mutation type (Supp Figure 4.7F).

Among the high confidence results with a final classifier PR AUC > 0.8, 14 of 59 associations were between different alterations types within the same gene, such as TP53 truncating, splice site, and missense mutations in various tumor types (Figure 4.8). There were four associations within the same gene families, e.g. IDH1/2 or the RAS protein family. In addition, I identified gene fusion events partners in BRAF and RET which associated with BRAF missense mutation in THCA. I also identified that for 20 of the 59 high confidence bait-catch associations both genes were present in the same human pathway gene set (not regarding gene sets with > 1000 genes)[**?**].

When all Pan-Cancer results are considered, the resulting association network does not group by tumor type but surprisingly by pathway (Suppl. Figure showing large cytoscape plot), and in particular four canonical pathways emerge (Figure 4.9). LURE identified very interesting associations for PTEN, in particular between PTEN and CTNNB1, a connection supported by recent research which suggests PTEN plays a role in regulating the subcellular localization of β-catenin[31]. Another striking association is between PTEN and EGFR, for which recent findings provide evidence that PTEN regulates EGFR signaling[32]. These LURE associations for PTEN reveal cross talk between pathways and provide further evidence that alterations in PTEN influence EGFR signaling and/or β-catenin signaling.

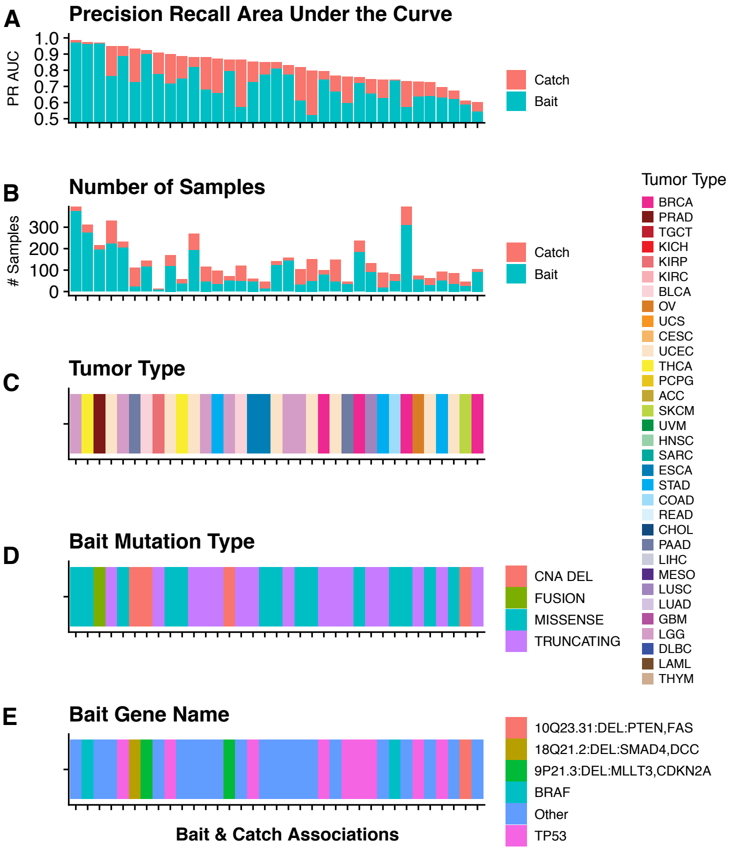


Figure 4.7: **Bait Catch Association Data Panel.** Panel shows bait and catch annotations for the 35 baits in which associations were found. Baits are across the x-axis and sorted by final Bait and Catch PR AUC. **(A) Precision Recall Area Under the Curve.** Stacked bar plot shows the original PR AUC score of the bait and the final PR AUC score including the new catch. **(B) Number of Samples** Stacked bar plot shows the number of samples in the bait and the additional samples in the catch. **(C) Tumor Type.** Annotation bar shows the tumor type in which the bait-catch association was found.**(D) Bait Mutation Type.** Annotation bar shows the type of bait mutation. **(E) Bait Gene Name.** Annotation bar shows the name of the bait gene.

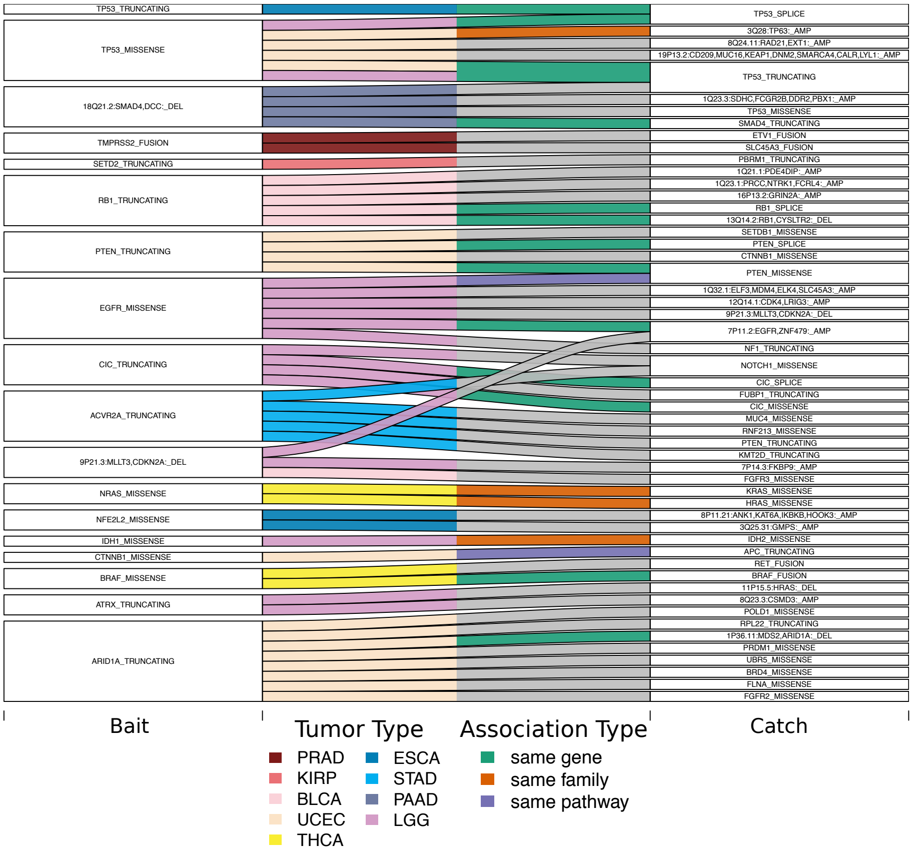


Figure 4.8: **High Confident LURE PANCAN Results.** Sankey plot shows the high confident 18 bait-catch associations with a final PR AUC greater than 0.8. Bait gene and mutation type are on the left side and the Catch bait gene and mutation type are on the right side. The flows represent an association between the bait and catch gene. The color of the left side of each flow is the tumor type in which each association was found. The color of the right side of each flow is the association type.

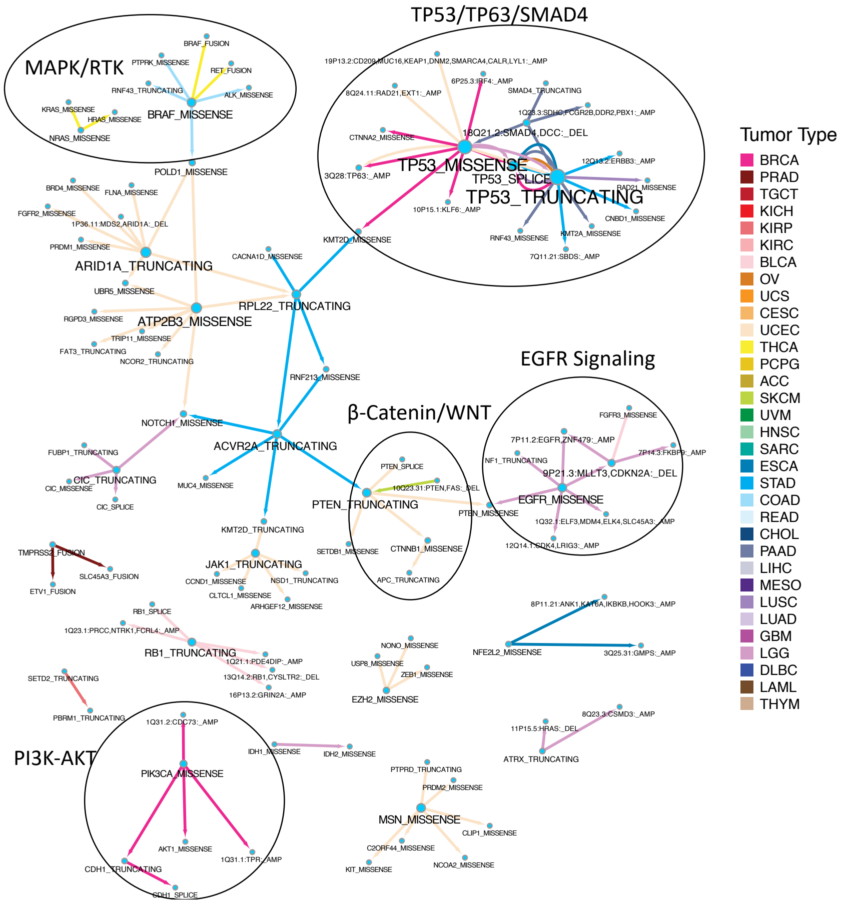


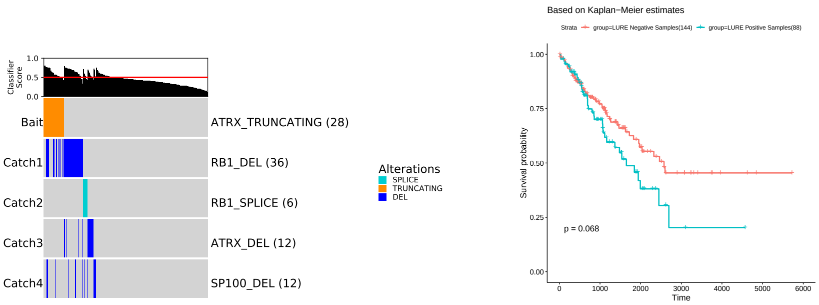
Figure 4.9: **LURE "Event Net" of LURE PANCAN Results** LURE Event Net shows all associations resulting from the 35 successful cosmic bait genes. Edges are colored by the tumor type in which the association was found. Edges are directed from Bait to Catch. Pathway associations are annotated with circles.

4.3.3 LURE finds new drivers of the Alternative Lengthening of Telomeres in Sarcomas

Tumors must employ Telomere Maintenance Mechanisms (TMM) to extend their telomeres in order to avoid senescence and multiply rapidly[33]. To date, there are two known mechanisms which tumor cells use to avoid telomere erosion: the overexpression of telomerase, an enzyme with the ability to extend telomeres, or the Alternative Lengthening of Telomeres (ALT) pathway. The vast majority of tumors overexpress telomerase in some way, whereas a small portion (10-15%) use ALT[34]. ALT+ samples lengthen telomeres through homologous recombination, mediated by loss-of-function mutations in the ATRX and DAXX genes[35]. Approximately 80% of tumors with ALT harbor mutations in ATRX or DAXX[36] leaving 20% with no known driver and in particular there are Lower Grade Glioma tumor samples which do not harbor a mutation in ATRX or the TERT promoter[37]. Using LURE, I sought to identify new driver mutations of ALT pathway using gene expression signatures of samples harboring ATRX loss of function mutations. Sarcomas and Lower Grade Gliomas have the highest prevalence of ALT+ samples, and ATRX is recurrently mutated, so I chose these tumor types to search for new drivers of the ALT pathway[35]. I restricted our gene set to a manually curated set of genes associated with telomere maintenance from the TelNet database[38], therefore I can expect results of telomere associated genes, but in particular I am looking for associations with ALT. Since TP53 is commonly mutated in ALT-positive (ALT+) samples and is not known to solely cause ALT[39], I decided to exclude any TP53 alterations in the possible catches in an effort to hone in on new ALT drivers.

Using ATRX truncating mutations as bait, LURE associated four mutations in Sarcomas (Figure 4.10A). While it is expected to associate ATRX truncating mutations with a ATRX copy number deletion[37], the deletions in RB1 and SP100 are novel results. I suggest the expression signature LURE identified in this analysis is classifying ALT+ TMM samples and the associated alterations are possibly driving the TMM. Previous work has associated RB1 alterations with long telomeres in the absence of TERT mutations and ATRX inactivation[40]. In addition, mouse models have revealed that the knock-out of Rb-family proteins causes elongated telomeres[41]. LURE also identified SP100 deletions as an ALT driver, and while SP100 deletions have not been directly reported to be involved in ALT, overexpression of SP100 in ALT+ cell lines resulted in suppression of ALT characteristics[42]. I therefore think it is possible that a SP100 deletion leads to unhindered ALT TMM activity. To further investigate the subset of LURE classified ALT+ samples, I performed a survival analysis and found the ALT+ samples show a significantly worse prognosis agreeing with recent research[43].

In Lower Grade Gliomas (LGG), also using ATRX truncating mutation as bait, LURE found similar results identifying associations with other ATRX alterations such as ATRX deletions, splice site, and missense mutations (Suppl. Figure X-oncoprint). Surprisingly, ATRX missense mutations were associated with truncating and as previously thought these mutations were not drivers, but this evidence suggests otherwise. Together these findings implicate new single and/or combinations of driver mutations required for the initiation of the ALT telomere maintenance mechanism and with further testing could prove to be therapeutic targets.



(a) LURE Oncoprint

(b) Survival Plot

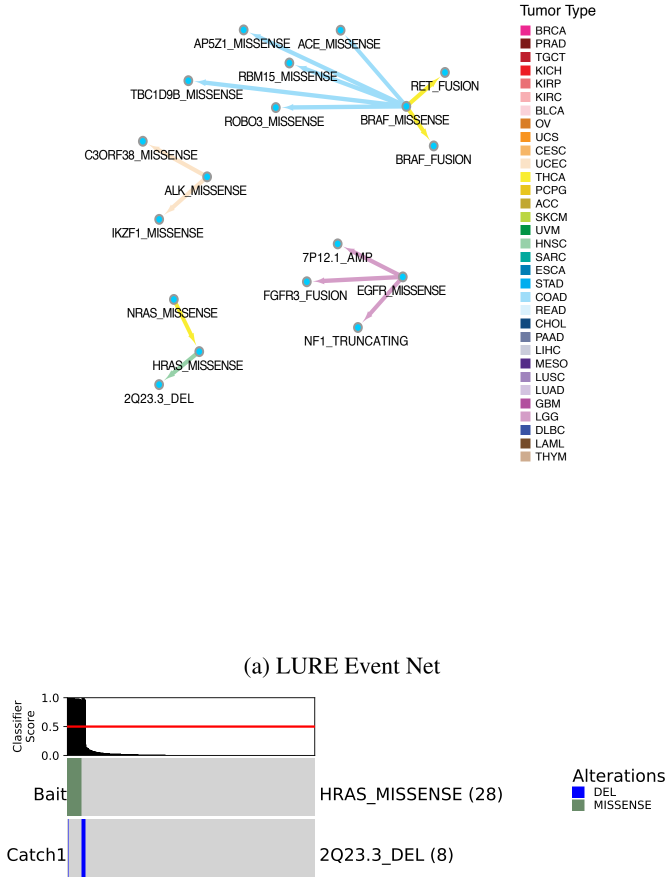
Figure 4.10: **LURE Oncoprint and Survival Plot.** **(A)** Using a known driver for ALT, ATRX truncating mutations, for bait in Sarcomas(SARC), LURE found three catch events: copy number deletions in ATRX, RB1, and SP100. **(B)** Survival plot shows the ALT classification using the final set cover classifier from the LURE method.

4.3.4 LURE identifies associations within the MAPK signaling pathway

Oncogenic mutations of the HRAS, NRAS, or KRAS genes are frequently found in human tumors and known to throw off the normal balance of signaling networks controlling cellular proliferation, differentiation, and survival. Oncogenic mutations in a number of other upstream or downstream components of MAPK/RTK signaling pathways, including membrane receptor tyrosine kinases (RTKs) or cytosolic kinases, have been detected more recently to be associated with a variety of cancer types[44]. The oncogenic RAS mutations and other mutation events within the MAPK/RTK signaling pathways are often mutually exclusive, indicating that the deregulation of Ras-dependent signaling is the essential requirement for tumorigenesis[44]. Previous studies have shown that tumor samples harboring Ras protein mutations have a unique gene expression signature and Ras-dependent samples can be more accurately defined by using the signature instead of the mutation status alone.[45] Building on this knowledge I was able to use LURE to not only train an accurate Ras-dependent classifier as was done in Way et al[46], but also identify new alterations which may be activating the MAPK/RTK signaling pathway in samples without a Ras protein mutation. To begin I restricted our baits to genes known to be involved in the MAPK/RTK signaling pathway[47]. Of those initial baits 23 event classifiers scored greater 0.5. I ran LURE for these 23 bait events and did not place any restrictions on the ”catch” set. The resulting ”Event Net” revealed known as well as new associations (Figure 4.11A). One interesting association found by LURE in Head-Neck Squamous Cell Carcinomas (HNSC) is between HRAS missense mutations and a focal deletion in the 2q23.3 locus (Figure 4.11B). The samples were by and large mutually exclusive with only one sample having both a 2q23.3 and HRAS alteration and no samples with alterations in either KRAS or NRAS. Among the 61 genes in the 2q23.3 locus, CHST11 has been shown to regulate MAPK/RTK pathway activity in hepatocellular carcinoma[48]. I suggest that in the absence of a HRAS mutation, MAPK signaling is activated by a deletion of the 2q23.3 locus in Head-Neck Squamous Cell Carcinomas (HNSC).

4.3.5 LURE ran on the PCAWG dataset (using classifiers from the PANCAN dataset)

In an effort to use LURE to find a functional impact of non-coding mutations by association with known coding driver mutations I incorporated the new PanCancer Analysis of Whole Genomes (PCAWG) dataset into our analysis. Using even a larger set of bait classifiers which included not only the bait classifiers from the previous exercise I also included copy number alterations. After preparing the bait classifiers, I quantile normalized the two different datasets to remove any platform batch effects and ran LURE which identified 4 new associations (Figure 4.12). Of particular interest, the association between MDM2 amplifications and



(a) LURE Event Net

(b) LURE Oncoprint

Figure 4.11: **MAPK LURE Pathway Analysis (A) LURE ”Event Net” showing MAPK/RTK associations.** Each node represents an event. The directed edges represent an association and the direction of the LURE discovery. The color of each edge represents the tumor type in which the association was found. **(B) LURE Oncoprint of HRAS in HNSC.** Using HRAS as a bait in HNSC, LURE found a delection catch event of the 2q23.3 region.

coding mutations of TP53 is expected and worked as a positive control. Another association of CTNNB1 coding mutations with enhancer mutations in PSIP1 enhancer mutations in Liver Hepatocellular Carcinoma is novel and interesting as there has been recent work correlating expression between PSIP1 and CTNNB1 in gliomas[49].

4.4 Discussion

Here I presented a new method to classify genomic alterations in tumors with other well known alterations using a specific feature dataset such as gene expression. This method works best with a large dataset because in order to build an accurate bait classifier there needs to be an adequate number of positive samples. In addition, identifying catch samples requires a fair amount of samples with the same alteration in order to pass the FDR correction in the sample set enrichment test (SSEA). The most common associations found with LURE in the PANCAN analysis were between different alteration types within the same gene, such as TP53. Also a considerable amount of associations were found within the same gene family as well as pathway. In addition to associating known alterations across PANCAN, I also demonstrated LURE's ability to find novel drivers of the ALT TMM and the MAPK/RTK pathway. LURE's inability to find novel non-coding associations with coding variants lies in the fact the power was lacking in the smaller number of non-coding tumor samples. I believe there needs to be an increase of whole genome sequencing of tumors in order to correlate non-coding mutations with coding.

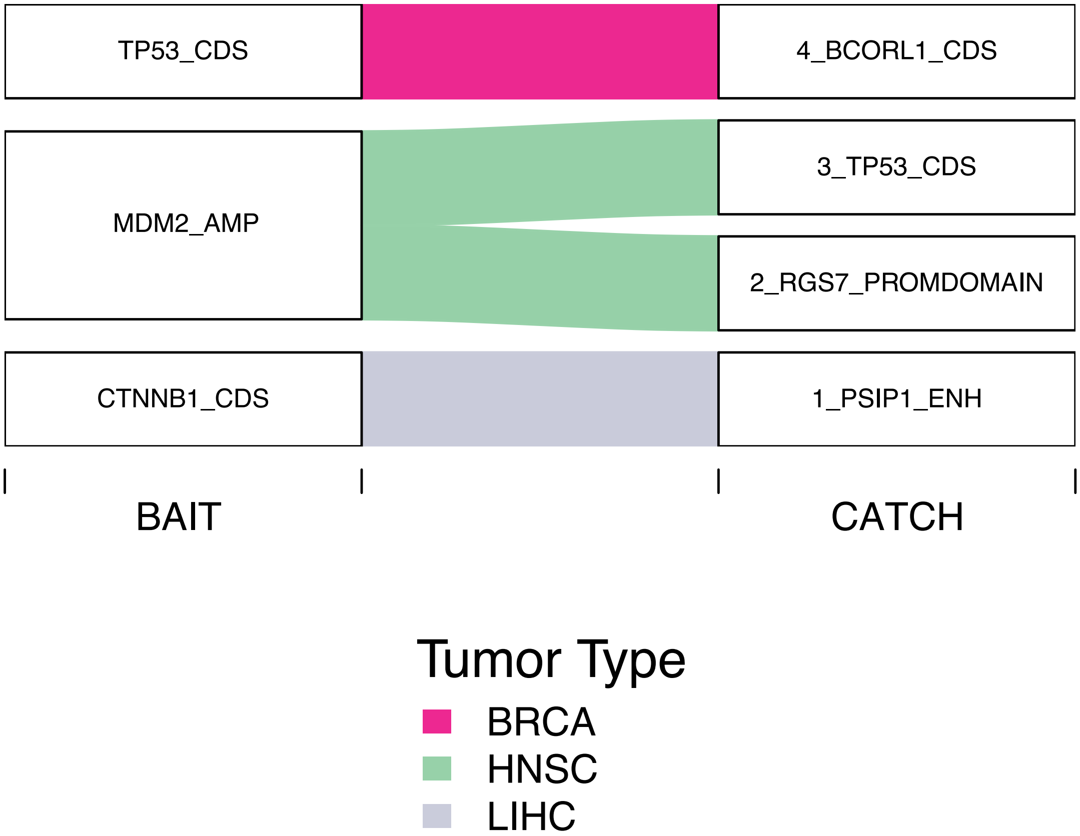


Figure 4.12: **Alluvial Plot of LURE PCAWG Results.** The left side of the plot shows the Baits and the right side shows Catches. The flow represents the tumor type in which an association was discovered by LURE.

Chapter 5

Non-coding Variant Function Discovery

5.1 Introduction

Over the past decade, cancer genome sequencing efforts such as The Cancer Genome Atlas (TCGA) have identified millions of somatic genetic aberrations; however, the annotation and interpretation of these aberrations remains a major challenge[9]. Specifically, while some aberrations occur frequently in specific cancer types, there is a long tail of rare aberrations that are difficult to distinguish from random passenger aberrations in modestly sized patient cohorts[50, 3] In many cancers, a significant proportion of patients do not have known coding driver mutations[51], suggesting that additional driver mutations remain undiscovered. To date, the vast majority of known driver mutations affect protein-coding regions. Only a few non-coding driver mutations, most notably mutations in the TERT promoter[52, 53, 54], have been identified; somatic expression Quantitative Trait Loci (eQTLs) correlate with gene expression changes in some cancer types[55]. Recent studies from the Pan-Cancer Analysis of Whole

Genomes (PCAWG) project of the International Cancer Genome Consortium (ICGC) reveal few recurrent non-coding drivers in analyses of individual genes and regulatory regions[54].

In this chapter I will discuss a pathway and network analysis of coding and non-coding somatic mutations from 2,583 tumors from 27 tumor types compiled by the Pan-Cancer Analysis of Whole Genomes (PCAWG) project of the International Cancer Genome Consortium (ICGC)[56], the largest collection of uniformly processed cancer whole genomes to date.

In addition to unlocking the non-coding areas of the genome enabling further analysis of the tumor variants, whole genome sequencing has allowed for sequencing for other regions such as telomeres to be sequenced. This chapter also includes a discussion on such sequencing and the analysis I performed confirming results of a possible new driver of the ALT pathway found by LURE in Sarcomas.

5.2 Using Gene Pathway Networks to Annotate Non-coding Variants

Cancer driver mutations unlock oncogenic properties of cells by altering the activity of hallmark pathways[57]. Accordingly, cancer genes are known to cluster in small number of cellular pathways and interacting subnetworks[3]. Previously, pathway and network analysis has proven useful for implicating infrequently mutated genes as cancer genes based on their pathway membership and physical/regulatory interactions with recurrently mutated genes[58, 57]. However, the interactions between coding and non-coding driver mutations have not been systematically explored.

In work described elsewhere, Reyna, Haan et al[59] employed seven distinct pathway and network analysis methods and derived the consensus of the methods predictions as pathway-implicated driver (PID) genes(Figure 5.1). In total, we identified a consensus set of 93 high-confidence pathway-implicated driver genes using non-coding variants (PID-N) and a consensus set of 87 pathway-implicated driver genes using coding variants (PID-C)(PID venn diagram figure). Both sets of PID genes, particularly the PID-N set, contain rarely mutated genes that were not identified by individual recurrence tests but interact with other well-known cancer genes. In total, 121 novel PID-N and PID-C genes are revealed as promising candidates, expanding the landscape of driver mutations in cancer.

Furthermore, we examined the contribution of coding and non-coding mutations in altering biological processes, finding that while chromatin remodeling and some well-known signaling and proliferation pathways are altered primarily by coding mutations, other important cancer pathways, including developmental pathways such as Wnt and Notch pathways, are altered by both coding and non-coding mutations in PID genes. Intriguingly, we find many non-coding mutations in PID-N genes with roles in RNA splicing, and samples with these non-coding mutations exhibit similar gene expression signatures as samples with well-known coding mutations in RNA splicing factors(Figure 5.3). I sought to identify an orthogonal analysis supporting the RNA splicing module of non-coding mutants. The splicing module consisted of 17 PID-N genes belonging to splicing-related pathways (pathway oncoprint plot), including several heterogeneous nuclear ribonucleoproteins (hnNRP) and serine and arginine rich splicing factors (SRSFs). None of these PID-N genes were significantly mutated according to single-element tests of the PCAWG driver discovery analysis. I did not find any significant ($FDR < 0.1$) in

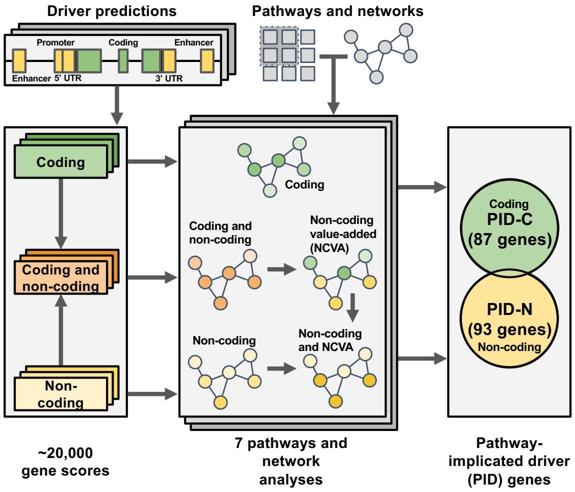


Figure 5.1: **Overview of the pathway and network analysis approach.** Coding, non-coding, and combined gene scores were derived for each gene by aggregating driver p-values from the PCAWG driver predictions in individual elements, including annotated coding and non-coding elements (promoter, 5 UTR, 3 UTR, and enhancer). These gene scores were input to five network analysis algorithms (CanIsoNet [Kahraman et al., in preparation], Hierarchical HotNet, an induced subnetwork analysis [Reyna and Raphael, in preparation], NBDI, and SSA-ME), which utilize multiple protein-protein interaction networks, and to two pathway analysis algorithms (ActivePathways [Paczkowska, Barenboim et al., in submission] and a hypergeometric analysis [Vazquez]), which utilize multiple pathway/gene-set databases. We defined a non-coding value-added (NCVA) procedure to determine genes whose non-coding scores contribute significantly to the results of the combined coding and non-coding analysis, where NCVA results for a method augment its results on non-coding data. We defined a consensus procedure to combine significant pathways and networks identified by these seven algorithms. The 87 pathway-implicated driver genes with coding variants (PID-C) are the set of genes reported by a majority (4/7) of methods on coding data. The 93 pathway-implicated driver genes with non-coding variants (PID-N) are the set of genes reported by a majority of methods on non-coding data or in their NCVA results. Only 5 genes (CTNNB1, DDX3X, SF3B1, TGFBR2, TP53) are both PID-C and PID-N genes.

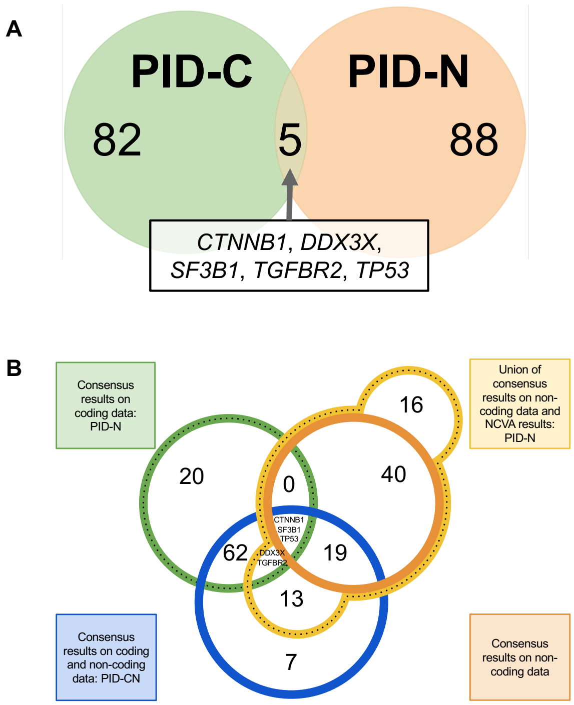
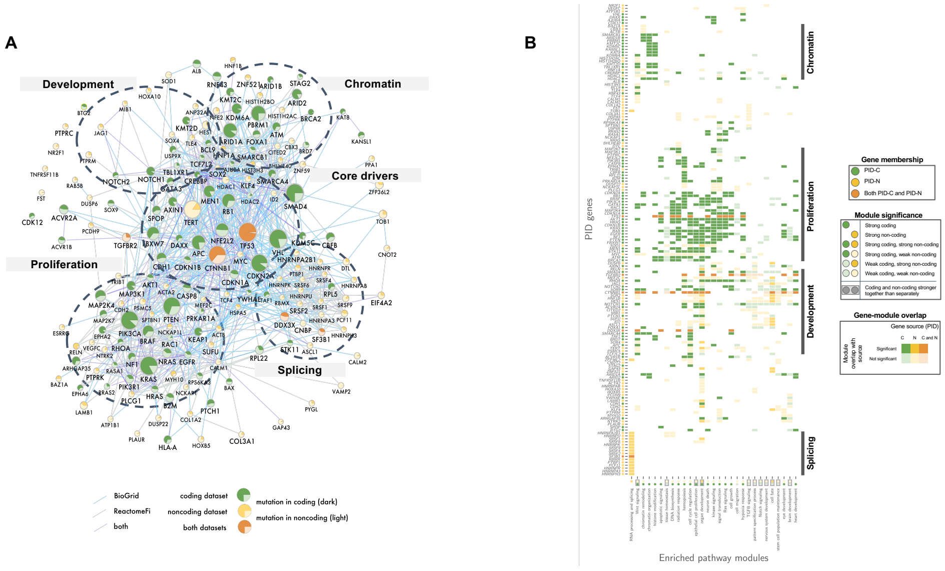

A
PID-C
PID-N
82
5
88
CTNNB1, DDX3X,
SF3B1, TGFBR2, TP53
B
Consensus results on coding data: PID-N
Union of consensus results on non-coding data and NCVA results: PID-N
16
20
0
40
CTNNB1
SF3B1
TP53
62
DDX3X
TGFBR2
19
13
7
Consensus results on coding and non-coding data: PID-CN
Consensus results on non-coding data

Figure 5.2: **Overlap of consensus results for pathway and network methods. (A) PID-C and PID-N genes have negligible overlap.** Only 5 genes (CTNNB1, DDX3X, SF3B1, TGFBR2, TP53 are both PID-C and PID-N genes. **(B) Overlap of all consensus results.** Four-circle Venn diagram for the overlap of the consensus results on coding data, i.e., PID-C genes; consensus pathway/network results on non-coding data; consensus pathway/network results on coding and non-coding data; and the union of the consensus results on non-coding data and the non-coding value-added (NCVA) results, i.e., PID-N genes.



(a)

(b)

Figure 5.3: **Pathway and network modules containing PID-C and PID-N genes. (A) Network of functional interactions between PID-C and PID-N genes.** Nodes represent PID-C and PID-N genes and edges show functional interactions from the ReactomeFI network (grey), physical protein-protein interactions from the BioGRID network (blue), or interactions recorded in both networks (purple). Node color indicates PID-C genes (green), PID-N genes (yellow), or both PID-C and PID-N genes (orange);node size is proportional to the score of the corresponding gene; and the pie chart diagram in each node represents the relative proportions of coding and non-coding cancer mutations associated with the corresponding gene. Dotted outlines indicate clusters of genes with roles in chromatin organization and cell proliferation, which predominantly contain PID-C genes; development, which includes comparable amounts of PID-C and PID-N genes; and RNA splicing, which contains PID-N genes. A core cluster of genes with many known drivers are also indicated. (B) Pathway modules containing PID-C and PID-N genes. Each row in the matrix corresponds to a PID-C or PID-N gene, and each column in the matrix corresponds to a pathway module enriched in PID-C and/or PID-N genes (see Methods). A filled entry indicates a gene (row) that belongs to one or more pathways (column) colored according to gene membership in PID-C genes (green), PID-N genes (yellow), or both PID-C and PID-N genes (orange). A darkly colored entry indicates that a PID-C or PID-N gene belongs to a pathway that is significantly enriched for PID-C or PID-N genes, respectively. A lightly colored entry indicates that a PID-C or PID-N gene belongs to a pathway that is significantly enriched for the union of PID-C and PID-N genes but not for PID-C or PID-N genes separately. Enrichments are summarized by circles adjacent each pathway module name and PID gene name. Boxed circles indicate that a pathway module contains a pathway that is significantly more enriched for the union of the PID-C and PID-N genes than the PID-C and PID-N results separately. The enriched modules and PID genes are clustered into four biological processes: chromatin, development, proliferation, and RNA splicing as indicated, with differing contributions of PID-C and PID-N genes.

cis associations between non-coding mutations and altered expression of these genes. Thus, I explored potential in trans effects on pathway expression changes. I found that non-coding mutations in splicing-related PID-N genes largely recapitulate a recently published association by TCGA[1] between coding mutations in several splicing factors and differential expression of 47 pathways (Figure 5.5). Three clusters of mutations were identified using hierarchical clustering based on patterns of differential expression (C1, C2, and C3;Figure 5.5). The clusters were found to be robust to the choice of clustering approach as a highly overlapping set of clusters was produced by t-distributed stochastic neighbor embedding (top annotation bar in Figure 5.5). Further support for robustness of clusters was found via silhouette scores and bootstrapping (silhoutte plot). Each of these clusters contained at least one coding mutation in the splicing genes SF3B1, FUBP1, and RBM10 as reported in [1], with non-coding mutations in splicing-related PID-N genes showing similar gene expression signatures. The joint analysis of coding and non-coding mutations in splicing factors also recovered the two groups of enriched pathways (P1 and P2 in Figure 5A) reported in [1]. One group (P1) is characterized by immune cell signatures and the other group (P2) reflects mostly cell-autonomous gene signatures of cell cycle, DDR, and essential cellular machineries[1]. The similarity between the gene expression signatures for non-coding mutations in several PID-N splicing factors and coding mutations in splicing factor genes[1] supports a functional role for splicing-related PID-N genes in altering similar gene expression programs.

This splicing analysis and the other analysis by [59] demonstrates that somatic non-coding mutations in untranslated and cis-regulatory regions constitute a complementary set of genetic perturbations with respect to coding mutations, affect several biological pathways and

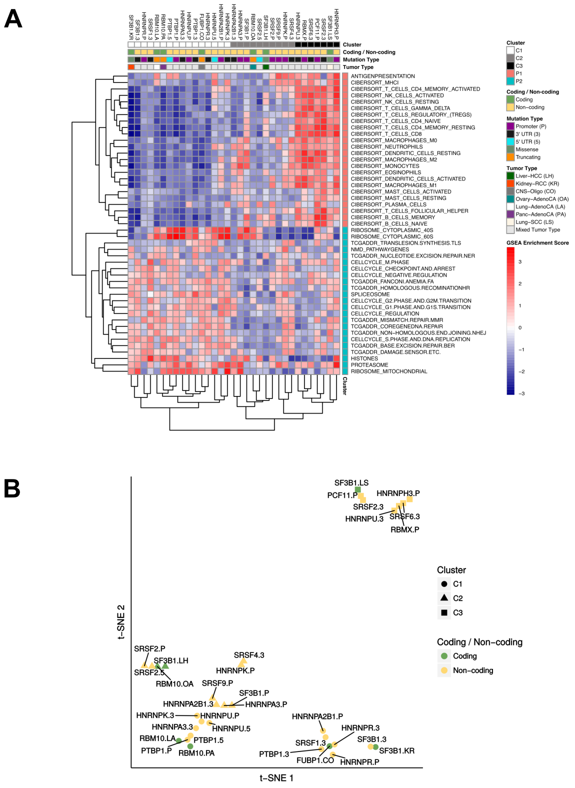


Figure 5.4: **RNA splicing factors are targeted primarily by non-coding mutations and alter expression of similar pathways as coding mutations in splicing factors. (A) Heatmap of Gene Set Enrichment Analysis (GSEA) Normalized Enrichment Scores (NES).** The columns of the matrix indicate non-coding mutations in splicing-related PID-N genes and coding mutations in splicing genes reported in [1] and the rows of the matrix indicate 47 curated gene sets[1]. Red heatmap entries represent an upregulation of the pathway in the mutant samples with respect to the non-mutant samples and blue heatmap entries represent a downregulation. The first column annotation indicates mutation cluster membership according to common pathway regulation. The second column annotation indicates whether a mutation is a non-coding mutation in a PID-N gene or a coding mutation[1], with the third column annotation specifies the aberration type (promoter, 5 UTR, 3 UTR, missense, or truncating). The fourth column annotation indicates the cancer type for coding mutations from [1]. The mutations cluster into 3 groups: C1, C2, and C3. The pathways cluster into two groups[1]: P1 and P2, where P1 contains an immune signature gene sets and P2 contains cell autonomous gene sets as reported in [1]. **(B) tSNE plot of mutated elements illustrates clustering of gene expression signatures for samples with non-coding mutations in splicing-related PID-N genes with gene expression signatures for coding mutations in previously published splicing factors.** The shape of each point denotes the mutation cluster assignment (C1, C2, or C3), and the color represents whether the corresponding gene is a PID-N gene with non-coding mutations or a splicing factor gene with coding mutations[1].

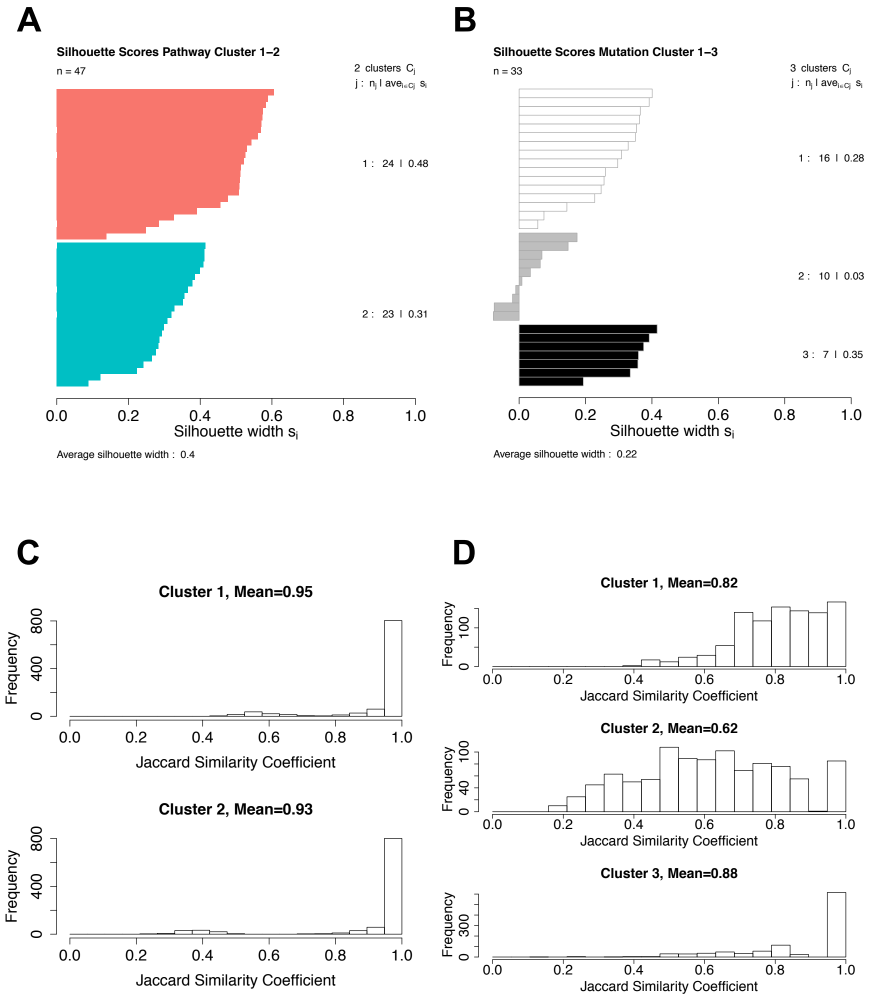


Figure 5.5: **Cluster stability analysis.** **(A)** Silhouette scores of pathway clusters 1 and 2. Silhouette width is on the x-axis, and the 47 pathways across y-axis. Average silhouette score per cluster is shown to the right of the bar plots. **(B)** Silhouette scoring of the 3 mutation element clusters. **(C, D)** Histograms representing the results of a cluster bootstrapping analysis using the Jaccard similarity coefficient to identify how often pathways **(C)** or mutation elements **(D)** clustered together in each bootstrap.

molecular interaction networks, and should be further investigated for their role in the onset and progression of cancer.

5.3 Unsupervised Classification of Non-coding Telomere Regions

One of the hallmarks of cancer is its ability to evade the normal cellular mechanisms of senescence[60]. Normal somatic cells typically have finite cell division potential, with telomere attrition one mechanism to limit numbers of mitoses[33]. Cancers enlist multifarious strategies to achieve replicative immortality. Over-expression of the telomerase gene, TERT, which maintains telomere lengths, is especially prevalent. This can be achieved via point mutations in the promoter that lead to de novo transcription factor binding[52, 53]; hitching TERT to highly active regulatory elements elsewhere in the genome[61]; insertions of viral enhancers upstream of the gene[62]; and increased dosage through chromosomal amplification. In addition, there is an alternative lengthening of telomeres (ALT) pathway, in which telomeres are lengthened through homologous recombination, mediated by loss-of-function mutations in the ATRX and DAXX genes[35].

Overall, 16% of tumours in the PCAWG dataset exhibited somatic mutations in ATRX, DAXX and/or TERT with a number of altered genes correlating with individual telomeric features such as TP53, ATRX, PLCB2, MEN1, TSSC4, RB1, and DAXX. TERT alterations were detected in 270 samples, whereas 128 tumours had somatic mutations in ATRX or DAXX, of which 71 were protein-truncating.

Whole genome sequencing includes reads of DNA from the telomeres, which can be recognised by the preponderance of the characteristic hexameric repeat, TTAGGG. In work described elsewhere in Sieverling, et al, 12 features of telomeric sequence were measured on the PCAWG cohort, including counts of nine variants of the core hexameric sequence, the number of ectopic telomere-like insertions within the genome, the number of genomic breakpoints, and the telomere length as a ratio between tumour and normal. I clustered the PCAWG samples based on their telomere sequence-associated features, normal and tumour samples formed distinct, non-overlapping clusters (Figure 5.6), suggesting that the biology of telomere maintenance is nearly universally altered in cancer.

Unexpectedly, I noticed the tumour samples formed four distinct sub-clusters (Figure 5.7A), suggesting a more diverse array of telomere maintenance mechanisms than the usual TERT/ALT dichotomy. Clusters C1 (47 tumours) and C2 (42 tumours) were enriched for traits of the ALT pathway, having longer telomeres, more genomic breakpoints, more ectopic telomere insertions, and variant telomere sequence motifs (Figure 5.8). C1 and C2 were distinguished from one another by the latter having striking elevation in the number of TTCGGG and TGAGGG variant motifs among the telomeric hexamers. Samples in clusters C3 (33 tumours) and C4 (2396 tumours) had relatively short telomere lengths, fewer genomic breakpoints and more normal frequencies of variant hexamers. Interestingly, samples in C2 and C3 only infrequently underwent whole genome duplication, as opposed to those in C1 and C4 ($p < 6.985x10-6$). Thyroid adenocarcinomas were strikingly enriched among C3 samples (26/33 C3 samples; $p < 2.2x10-16$); the C1 cluster (ALT subtype 1) was common among sarcomas; and both pancreatic endocrine neoplasms and low-grade gliomas had a high proportion of samples in the C2 cluster (ALT subtype 2) (Figure 5.7B). Interestingly, some of the

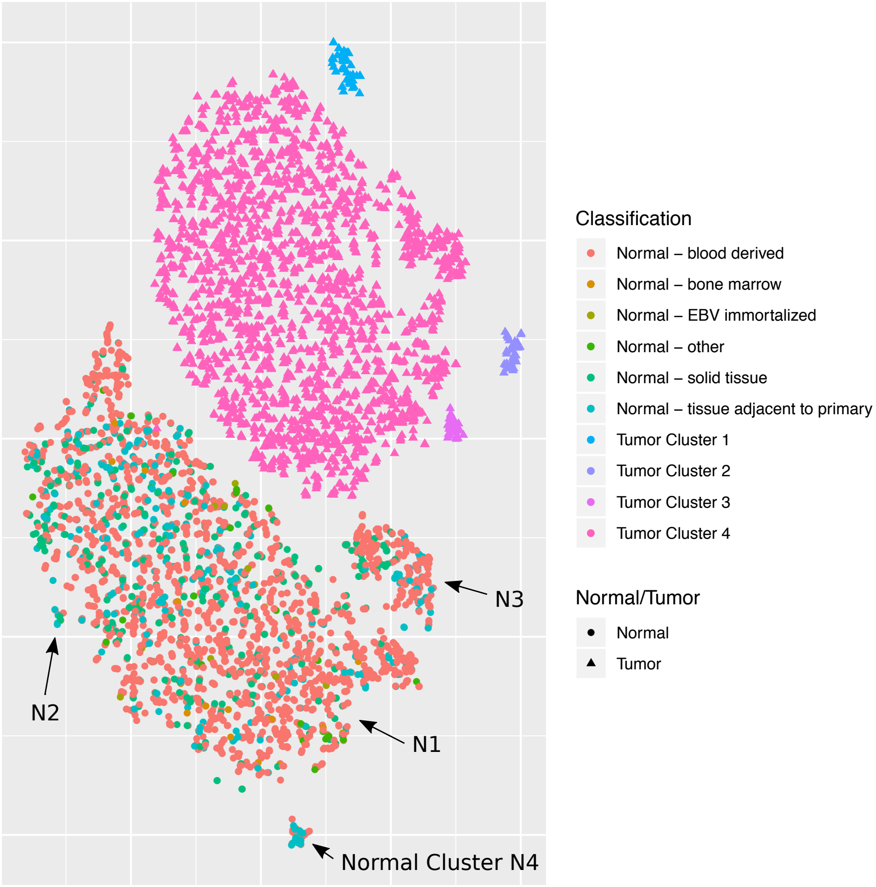


Figure 5.6: **Differences between normal and cancer-associated telomere properties.** Scatter plot showing the four clusters of telomere patterns identified across PCAWG cancers, together with the more homogeneous cluster of matched normal samples, generated by t-Distributed Stochastic Neighbour Embedding.

thyroid adenocarcinomas and pancreatic neuroendocrine tumours that cluster together (Cluster C3; (Figure 5.7) have matched normals that also cluster together (Normal cluster N3, Figure 5.6). This suggests that a subtype of these tumor types share telomeric maintenance properties: for example, the GTAGGG repeat is overrepresented among samples in this group (Extended Data Figure 19). Somatic driver mutations were also unevenly distributed across the four clusters (Figure 5.7C). C1 tumours were enriched for RB1 mutations or structural variants (p=3.1x10-5), as well as frequent structural variants affecting ATRX (p=6x10-14), but not DAXX. The RB1 and ATRX mutations were, by and large, mutually exclusive (Extended Data Figure 20A). In contrast, C2 tumours were enriched for somatic point mutations in ATRX and DAXX ($p < 6.4x10-5$), but not RB1. The enrichment of RB1 mutations in C1 remained significant when only leiomyosarcomas and osteosarcomas were considered, confirming that this enrichment is not merely a consequence of the different distribution of tumour types across clusters. C3 samples had frequent TERT promoter mutations (30%; p=2.3x10-6). Patterns of RB1, ATRX, DAXX and TERT expression confirmed the associations revealed in the mutation analysis (Figure 5.10B).

The predominance of RB1 mutations in C1 was striking. Nearly a third of the samples in C1 contained an RB1 alteration, evenly distributed across truncating SNVs, SVs and shallow deletions, with a mix of clonal and subclonal events (Figure 5.11). Interestingly, previous work has shown that RB1 mutations are associated with long telomeres in the absence of TERT mutations and ATRX inactivation[40], and mouse models have revealed that knockout of Rb-family proteins causes elongated telomeres[41]. The association with the C1 cluster here suggests that RB1 mutations can represent another route to activating the ALT pathway.

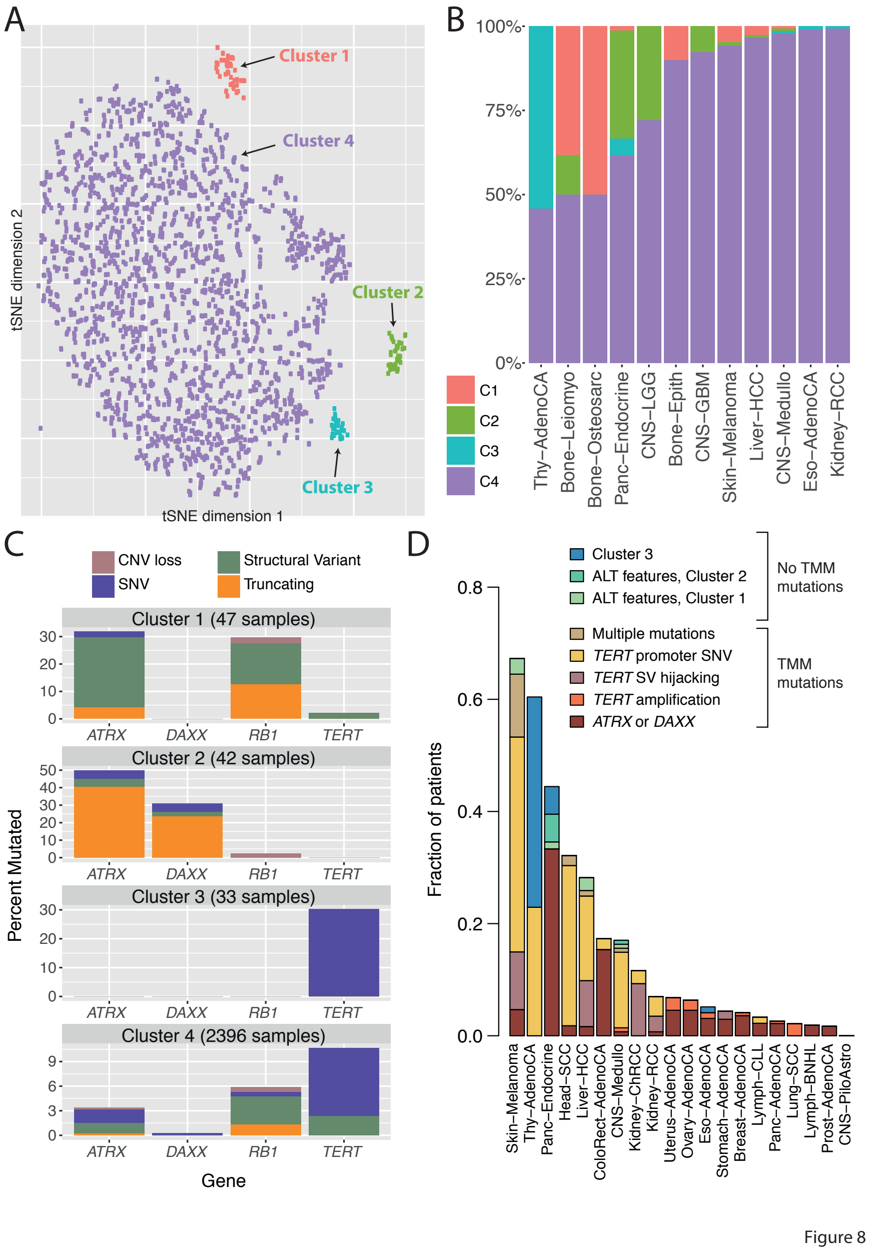


Figure 5.7: **Telomere sequence patterns across PCAWG.** (A) Scatter-plot showing the four clusters of telomere patterns identified across PCAWG by t-Distributed Stochastic Neighbour Embedding (tSNE). (B) Distribution of the four clusters of telomere patterns in selected tumour types from PCAWG. (C) Distribution of relevant driver mutations associated with alternative lengthening of telomere and normal telomere maintenance across the four clusters. (D) Distribution of telomere maintenance abnormalities across tumour types with more than 40 patients in PCAWG. Samples classified as tumour cluster 1-3 if they fall into a relevant cluster without mutations in TERT, ATRX or DAXX and have no ALT phenotype.

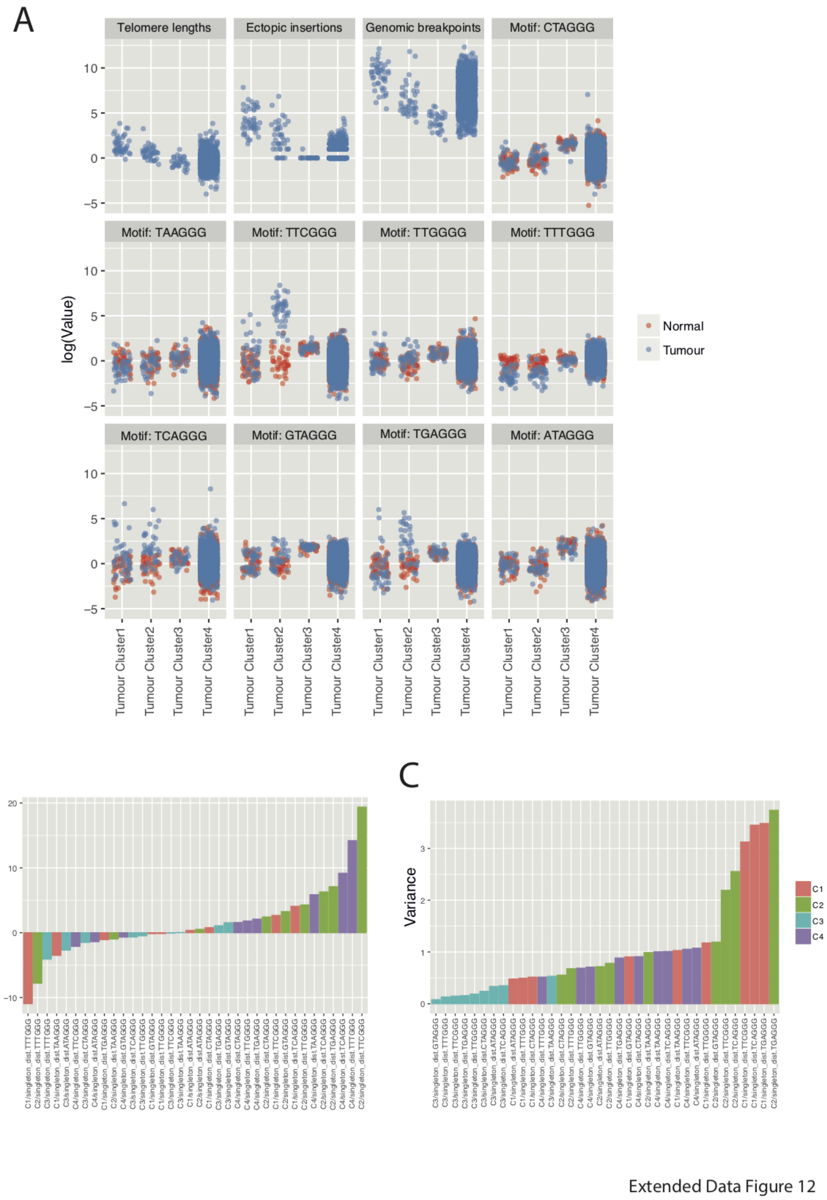


Figure 5.8: **Properties of telomeres across different tumor clusters.** (A) Distribution of telomere sequence and properties across samples in the four clusters, with both tumour (blue points) and matched normal (red points) shown. (B) Enrichment (positive T statistics) or depletion (negative T statistics) of different variant sequence motifs in the four clusters of telomere properties. (C) Variance of frequency of different sequence motifs across the four clusters.

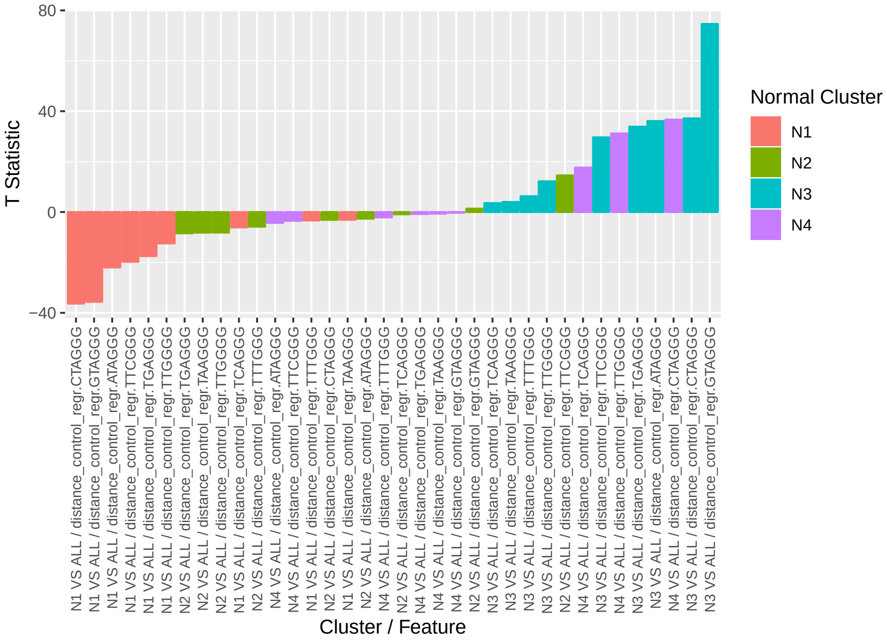


Figure 5.9: **Properties of telomeres across different normal clusters.** Properties of telomeres across different normal clusters. Enrichment (positive T statistics) or depletion (negative T statistics) of different variant sequence motifs in the four normal clusters of telomere properties.

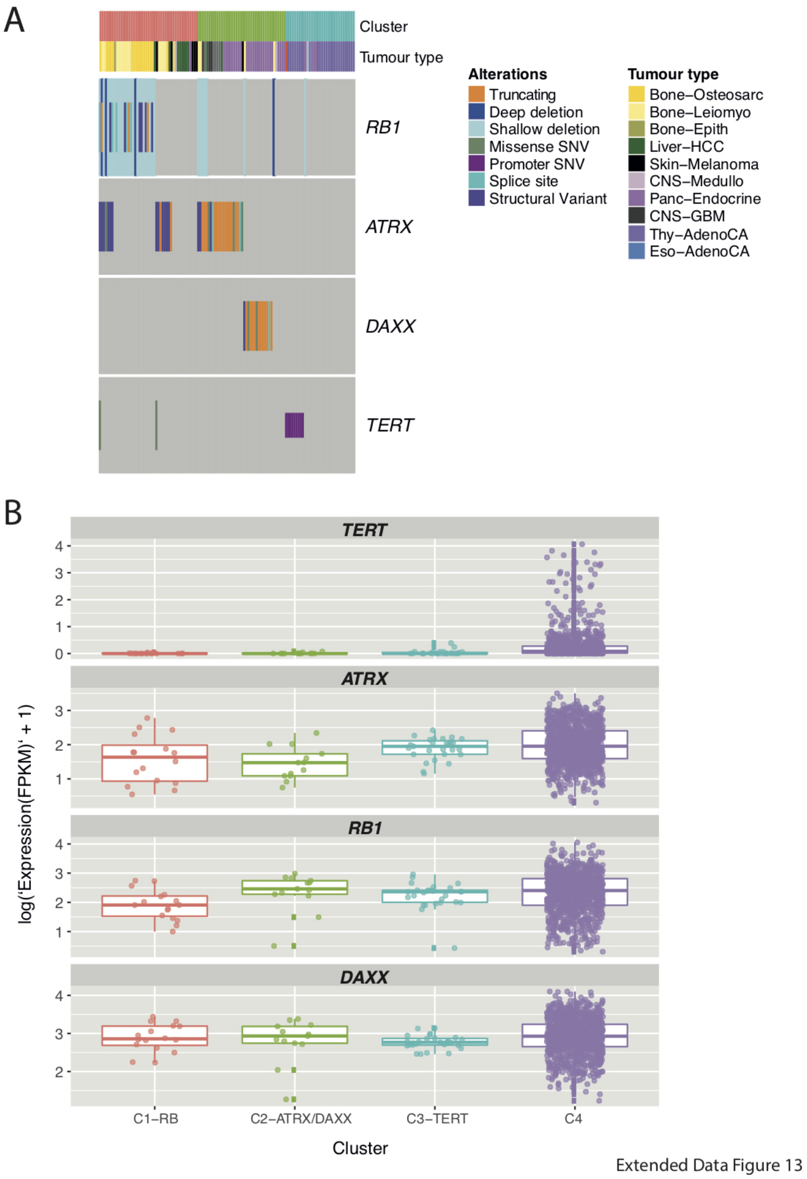


Figure 5.10: **Co-mutation and expression levels of genes related to telomere maintenance across the four clusters of telomere properties.** (A) Patterns of co-mutation of the relevant driver mutations across individual patients. Columns in plot represent individual patients, coloured by type of abnormality observed. (B) Box and whisker plots of expression levels of key telomere maintenance genes across the four clusters of telomere properties. The boxes demarcate the interquartile range, with a horizontal line to mark the median.

This route gives subtly different properties of telomeric sequence compared to tumours with ALT through inactivated DAXX, which fall almost exclusively in cluster C2. The finding of a distinct cluster of telomeric features, C3, so strongly associated with thyroid adenocarcinomas remains unexplained.

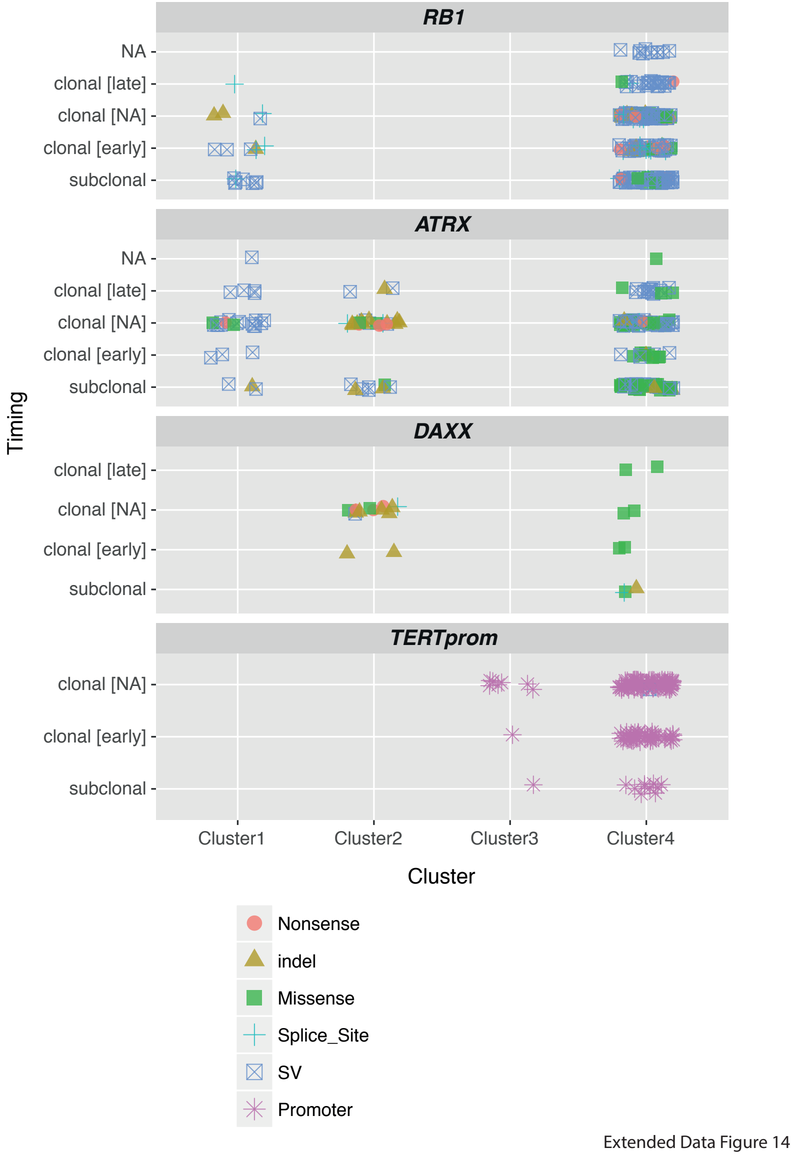


Figure 5.11: **Timing of mutation in genes related to telomere maintenance.** Clonal [early] denotes clonal mutations occurring before duplications involving the relevant chromosome (including whole genome duplications); clonal [late] to clonal mutations occurring after such duplications; and clonal [NA] to mutations occurring when no duplication was observed.

Chapter 6

Future Directions and Conclusion

6.1 Future Directions

Despite the efforts of TCGA and ICGC, there is still a lack of data for the majority of tumor types, including rare adult and pediatric tumors. For example, in the TCGA data, BRCA-type cancers have over 1,000 samples while UVM-type cancers have only about 80 samples. This is just one example of the many underrepresented tumor types among the TCGA and ICGC data.[63]. Of course, BRCA is a relatively common cancer mutation so there is an abundance of individuals to pull samples from, but there remains an immediate need for a greater variety of sample data to better understand and improve cancer treatments. LURE requires a minimum of 3-4 samples with a specific unclassified mutation in order to associate an event with a known driver, so tumor types with a lower number of samples are incompatible with LURE. In addition, it is difficult to draw statistical power with such a small number of samples. The PCAWG analysis put forth great effort to find driver mutations in the non-coding regions of the genome, but they only sequenced approximately 2,500 tumor samples and about half of these samples were not analyzed for gene expression data. Some specific tumor types had fewer than 10 samples. As I learned from the fusion detection method challenge, sequencing coverage was vital in accurate fusion detection. Thus, in addition to more sample types, we need samples with higher coverage to accurately make confident mutation calls.

Improving cancer detection technologies is an important goal of cancer genomics research, as early detection of cancer improves patient response to treatment and overall prognosis[64]. One particular method of early detection involves analyzing cell free DNA (cfDNA) in blood plasma for specific biomarkers to reveal the existence of a tumor in the body. cfDNA contains both tumor-derived DNA, or ”circulating tumor DNA” (ctDNA), and DNA derived from non-tumor cells such as hematopoietic and stromal cells[65]. ctDNA shares unique characteristics of tumor DNA, such as cancer-associated mutations, translocations, and/or large CNVs not typically found in the cfDNA of healthy patients[66]. In patients with cancer, ctDNA generally represents a small fraction of the cfDNA, ranging from 5-10% in late-stage disease to 0.01-1.0% in early-stage disease. This value is even lower in premalignant conditions, so methods that consider both ctDNA and total cfDNA could be very beneficial[67]. The goal of an early cancer detection method would be to identify the existence of a tumor as well as the tumor's subtype. Recently, a method was published that uses machine learning to predict the existence of colon cancer among 546 patients and 271 non-cancer control subjects using a supervised machine learning[65]. The cancer patients were divided equally by gender and consisted of 80& early-stage (stages I and II) patients[65]. The cfDNA of the 817 total subjects was sequenced using whole genome sequencing and their machine learning method was able to correctly diagnosis the cancer patients with a mean ROC AUC of 0.92[65]. To date, this was the largest study using only cfDNA whole genome sequencing in patients for the early detection of colon cancer. While the group was successfully able to build a classification model to determine the existence of a tumor, an analysis of feature importance was absent. Such feature importance analysis could possibly reveal novel biomarkers, which would limit the amount of sequencing necessary to make a prediction and allow hospitals to perform early tumor detection without invasive procedures.

6.2 Conclusion

It is widely accepted that tumorigenesis is a multistep process that depends on the sequential accumulation of mutations within cells[68]. Although tumor cells often exhibit a large number of mutations[69][70], only a relatively small subset is crucial for neoplastic development[70][71][72]. In this thesis, I presented multiple methods and the necessary data to categorize and prioritize mutations as potential drivers for tumorigenesis using both supervised and unsupervised machine learning. In chapter one, I showed that low sequencing coverage is the most important feature that causes fusion detection methods to fail in identifying fusion breakpoints. By increasing sequencing coverage, we can identify more fusions and potentially identify more cancer drivers in samples that have no known drivers. In chapter two, I presented a new method called LURE, which associates different alterations based on mRNA expression signatures using supervised machine learning. I identified associations between alterations in the same gene, the same gene family, and within the same pathway. Using LURE I also found new associations in Sarcomas, linking ATRX mutations with two new possible drivers of the ALT TMM, RB1 and SP100. In order to provide an orthogonal analysis supporting these results, I present unsupervised clustering of TMM features in Chapter three. This revealed a subgroup of samples harboring mutations in RB1 and showing characteristics of ALT. In addition, I presented results from a network and pathway analysis showing how non-coding mutations in splicing factors can affect tumors in the same manner as a coding mutation in the splicing factor. In this thesis, I presented just a glimpse of the on-going research in the field of cancer genomics. I hope to inspire new ideas and spur greater interest in the field of genomics with the intention of overcoming the extremely convoluted problems associated with cancer.

Bibliography

[1] M. Seiler et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell reports*, 23(1):282–296.e4, April 2018.

[2] W. Street. Cancer Facts & Figures 2019. pp. 76, 2019.

[3] B. Vogelstein et al. Cancer Genome Landscapes. *Science (New York, N.Y.)*, 339(6127):1546–1558, March 2013.

[4] M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, June 2015.

[5] K. Tomczak et al. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68–A77, 2015.

[6] International Cancer Genome Consortium et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[7] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, December 2009. Google-Books-ID: TtrxCwAAQBAJ.

[8] T. Saito and M. Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3), March 2015.

[9] A. Gonzalez-Perez et al. Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods*, 10(8):723–729, August 2013.

[10] T. Abenius et al. System-Scale Network Modeling of Cancer Using EPoC. In I. I. Goryanin and A. B. Goryachev, editors, *Advances in Systems Biology*, Advances in Experimental Medicine and Biology, pp. 617–643. Springer New York, 2012.

[11] A. Bashashati et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology*, 13(12):R124, 2012.

[12] M. D. Leiserson et al. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology*, 16(1), 2015.

[13] G. Ciriello et al. Using MEMo to Discover Mutual Exclusivity Modules in Cancer. *Current protocols in bioinformatics*, CHAPTER 8:Unit–8.17, March 2013.

[14] S. A. Tomlins et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia (New York, N.Y.)*, 10(2):177–188, February 2008.

[15] S. Faderl et al. The biology of chronic myeloid leukemia. *The New England Journal of Medicine*, 341(3):164–172, July 1999.

[16] A. M. Chinnaiyan and N. Palanisamy. Chromosomal aberrations in solid tumors. *Progress in Molecular Biology and Translational Science*, 95:55–94, 2010.

[17] F. Mitelman et al. The impact of translocations and gene fusions on cancer causation. *Nature Reviews. Cancer*, 7(4):233–245, April 2007.

[18] S. Kumar et al. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Scientific Reports*, 6:21597, February 2016.

[19] M. Carrara et al. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC bioinformatics*, 14 Suppl 7:S2, 2013.

[20] A. Y. Lee et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biology*, 19(1):188, 2018.

[21] F. Degenhardt et al. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 20(2):492–503, March 2019.

[22] M. B. Kursa and W. R. Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(1):1–13, September 2010.

[23] S. Takahashi et al. Prediction of breast cancer prognosis by gene expression profile of TP53 status. *Cancer Science*, 99(2):324–332, February 2008.

[24] S. Yamaguchi et al. Molecular and clinical features of the TP53 signature gene expression profile in early-stage breast cancer. *Oncotarget*, 9(18):14193–14206, March 2018.

[25] P. A. Konstantinopoulos et al. Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 28(22):3555–3561, August 2010.

[26] J. G. Tate et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*.

[27] J. W. Kim et al. Characterizing genomic alterations in cancer by complementary functional associations. *Nature Biotechnology*, 34(5):539–546, May 2016.

[28] S. Alsafadi et al. Cancer-associated *SF3B1* mutations affect alternative splicing by promoting alternative branchpoint usage. *Nature Communications*, 7:10615, February 2016.

[29] S. Turcan et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483(7390):479–483, March 2012.

[30] Cancer Genome Atlas Research Network et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England Journal of Medicine*, 372(26):2481–2498, June 2015.

[31] A. Persad et al. Active -catenin is regulated by the PTEN/PI3 kinase pathway: a role for protein phosphatase PP2a. *Genes & Cancer*, 7(11-12):368–382, November 2016.

[32] S. R. Shinde and S. Maddika. PTEN modulates EGFR late endocytic trafficking and degradation by dephosphorylating Rab7. *Nature Communications*, 7:10689, February 2016.

[33] J. W. Shay and W. E. Wright. Hayflick, his limit, and cellular ageing. *Nature Reviews. Molecular Cell Biology*, 1(1):72–76, 2000.

[34] J. D. Henson and R. R. Reddel. Assaying and investigating Alternative Lengthening of Telomeres activity in human cells and cancers. *FEBS Letters*, 584(17):3800–3811, September 2010.

[35] C. M. Heaphy et al. Prevalence of the Alternative Lengthening of Telomeres Telomere Maintenance Mechanism in Human Cancer Subtypes. *The American Journal of Pathology*, 179(4):1608–1615, October 2011.

[36] C. A. Lovejoy et al. Loss of ATRX, Genome Instability, and an Altered DNA Damage Response Are Hallmarks of the Alternative Lengthening of Telomeres Pathway. *PLOS Genet*, 8(7):e1002772, July 2012.

[37] Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine*, 372(26):2481–2498, June 2015.

[38] D. M. Braun et al. TelNet - a database for human and yeast genes involved in telomere maintenance. *BMC Genetics*, 19, May 2018.

[39] Y.-J. Chen et al. Association of Mutant TP53 with Alternative Lengthening of Telomeres and Favorable Prognosis in Glioma. *Cancer Research*, 66(13):6473–6476, July 2006.

[40] F. P. Barthel et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nature Genetics*, 49(3):349–357, March 2017.

[41] M. Garca-Cao et al. A role for the Rb family of proteins in controlling telomere length. *Nature Genetics*, 32(3):415, November 2002.

[42] W.-Q. Jiang et al. Suppression of Alternative Lengthening of Telomeres by Sp100-Mediated Sequestration of the MRE11/RAD50/NBS1 Complex. *Molecular and Cellular Biology*, 25(7):2708–2721, April 2005.

[43] R. T. Lawlor et al. Alternative lengthening of telomeres (ALT) influences survival in soft tissue sarcomas: a systematic review with meta-analysis. *BMC Cancer*, 19(1):232, March 2019.

[44] A. Fernndez-Medarde and E. Santos. Ras in Cancer and Developmental Diseases. *Genes & Cancer*, 2(3):344–358, March 2011.

[45] A. Loboda et al. A gene expression signature of RAS pathway dependence predicts response to PI3k and RAS pathway inhibitors and expands the population of RAS pathway activated tumors. *BMC Medical Genomics*, 3:26, June 2010.

[46] G. P. Way et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Reports*, 23(1):172–180.e3, April 2018.

[47] F. Sanchez-Vega et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337.e10, April 2018.

[48] H. Zhou et al. CHST11/13 Regulate the Metastasis and Chemosensitivity of Human Hepatocellular Carcinoma Cells Via Mitogen-Activated Protein Kinase Pathway. *Digestive Diseases and Sciences*, 61(7):1972–1985, July 2016.

[49] W. Xiang et al. [RNA interference of PC4 and SFRS1 interacting protein 1 inhibits invasion and migration of U87 glioma cells]. *Nan Fang Yi Ke Da Xue Xue Bao = Journal of Southern Medical University*, 36(6):802–806, June 2016.

[50] L. A. Garraway and E. S. Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, March 2013.

[51] M. S. Lawrence et al. Mutational heterogeneity in cancer and the search for new cancer genes. *Nature*, 499(7457):214–218, July 2013.

[52] S. Horn et al. TERT promoter mutations in familial and sporadic melanoma. *Science (New York, N.Y.)*, 339(6122):959–961, February 2013.

[53] F. W. Huang et al. Highly recurrent TERT promoter mutations in human melanoma. *Science (New York, N.Y.)*, 339(6122):957–959, February 2013.

[54] E. Rheinbay et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv*, pp. 237313, December 2017.

[55] W. Zhang et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nature Genetics*, 50(4):613–620, 2018.

[56] International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[57] P. Creixell et al. Pathway and network analysis of cancer genomes. *Nature Methods*, 12(7):615–621, July 2015.

[58] R. Ashkenazi et al. Pathways to TumorigenesisModeling Mutation Acquisition in Stem Cells and Their Progeny. *Neoplasia (New York, N.Y.)*, 10(11):1170–1182, November 2008.

[59] M. A. Reyna et al. Pathway and network analysis of more than 2,500 whole cancer genomes. *bioRxiv*, pp. 385294, August 2018.

[60] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[61] M. Peifer et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature*, 526(7575):700–704, October 2015.

[62] Y. Totoki et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nature Genetics*, 46(12):1267–1273, December 2014.

[63] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, July 2014.

[64] J. D. Schiffman et al. Early detection of cancer: past, present, and future. *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting*, pp. 57–65, 2015.

[65] N. Wan et al. *Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA:*. November 2018.

[66] E. Heitzer et al. Circulating tumor DNA as a liquid biopsy for cancer. *Clinical Chemistry*, 61(1):112–123, January 2015.

[67] J. C. M. Wan et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews. Cancer*, 17(4):223–238, 2017.

[68] K. R. Loeb and L. A. Loeb. Significance of multiple mutations in cancer. *Carcinogenesis*, 21(3):379–385, March 2000.

[69] A. G. Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews. Cancer*, 1(2):157–162, November 2001.

[70] E. G. Luebeck and S. H. Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23):15095–15100, November 2002.

[71] K. W. Kinzler and B. Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159–170, October 1996.

[72] T. Sjblom et al. The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)*, 314(5797):268–274, October 2006.