UNIVERSITY OF CALIFORNIA
SANTA CRUZ

USING RNA SEQUENCING DATA TO DETECT VARIANTS OF INTEREST

A dissertation submitted in partial satisfaction of the
requirements for

Doctor of Philosophy

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Jon Akutagawa

June 2022

The Dissertation of Jon Akutagawa
is approved:

Associate Professor Angela N. Brooks, Chair

Professor Joshua M. Stuart

Assistant Professor Olena M. Vaske

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Using RNA Sequencing Data to Detect Variants of Interest

by

Jon Akutagawa

The primary function for RNA sequencing (RNA-seq) is to investigate the transcriptome through differential gene expression. For cancer and other genetic diseases, detecting variants in the genome is critical for our understanding of how these diseases begin and progress. Here, I will present computational methods focused on using RNA-seq to detect disease-associated variants. We developed RNA-VACAY, a containerized high-throughput pipeline that automates somatic variant calling in RNA-seq data. We analyzed 1,349 RNA-seq samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Project and found that RNA-VACAY can accurately identify somatic variants of interest using tumor RNA-seq, alone. Our pipeline also does not require a matched normal sample to detect somatic variants, which is commonly unavailable in research or clinical settings. RNA-VACAY can also successfully identify 5' and 3' UTR variants, which are overlooked when using WES data. Additionally, we analyzed RNA-seq data to characterize splicing variants. We found a splice site variant associated with a previously detected variant of uncertain significance in a patient with an undiagnosed genetic disorder. We also developed a computational method for efficiently designing guide RNAs for a CRISPR/Cas9 screen to detect exon skipping events associated

with tumor formation. Our work demonstrates the impact of RNA-seq for detecting functional variants in genetic diseases.

To Sarah,

my everything,

for making this possible.

Acknowledgments

Graduate school has been an incredible time in my life and there are numerous people who have profoundly impacted my time here. I first want to acknowledge my advisor, Angela Brooks. Joining her lab has been one of the best decisions of my life. She has been extremely influential to me scientifically and personally and I will forever be grateful for her wonderful mentorship. Angela continues to inspire me to do good science.

I would also like to thank the other members of my committee, Josh Stuart and Olena Vaske. I decided to enroll at UC Santa Cruz because I wanted an opportunity to work with brilliant cancer researchers like them. I am thankful for their contributions to my work and for their constant support through my scientific training.

I was fortunate to have talented collaborators, which made science even more rewarding and enjoyable. I am incredibly grateful to Julie Aspden. Her keen insight and wisdom has been foundational in my research and have helped guide these projects significantly. Holly Beale and Allison Cheney were gifted collaborators, who both made the splicing analysis collaboration a success. I also wanted to thank the entire BME department for their commitment to scientific collaboration and for always being available when I needed assistance. Thank you to all of the staff, including Tracie Tucker and Theo-Alyce Gordon, for all of the behind-the-scenes administrative work you do.

I will always cherish the relationships I made within the Brooks Lab. It has been

such an phenomenal environment to do science. In particular, I want to acknowledge Alison Tang and Brandon Saint-John. I am grateful that we entered the lab together and that we were able to support each other through all the ups and downs of graduate school. Alexis Thornton warmly welcomed me to the lab and I am blessed to call her a friend. Thank you to Cameron Soulette and Max Marin for showing me the ropes when I joined the lab and for fielding my constant questions. I was honored to work closely with Allysia Mak and Beth DeVogelaere, who both helped shape the scope of these projects. I have truly enjoyed the company of everyone in our lab, including Megan Durham, Cindy Liang, Colette Felton, Dennis Mulligan, Eva Robinson, Pratibha Jagannatha, Kevyn Hart, Stefanie Brizuela, and Matthew Cattle.

My graduate student cohort has been an amazing source of joy and fun. I will miss the long nights of trivia and board games with all of you, including Bryan Thornlow, Roger Volden, Kishwar Shafin, Colleen Bosworth, and Akshar Lohith. Thank you to Andrew Bailey, who will always have a space on our boat.

I also wanted to acknowledge all of my friends who supported this crazy dream. There are so many communities that I am grateful for - college friends, high school friends, PG crew, Menlo friends. The list is long and I love you all.

I will always be grateful to our greater community here in Santa Cruz during the long nights and hazy days of parenthood. In particular, I am grateful for the support of Darrin Schultz and Micael Nuñez, who constantly looked out for us and were always

incredibly generous with their time and resources. Thank you to Jordan Eizenga, Ed Rice, and Lauren Sanders for all of the delicious meals. Thank you to Stephanie Cheung for your constant friendship and for bringing us crucial groceries.

I also wanted to thank everyone who was a part of my early scientific career, particularly Ben Braun, Kevin Shannon, and Cassandra Shu. You have all helped inspire me to pursue scientific excellence.

My family have literally made this entire experience possible. My in-laws, Wilson Lam and Mary Lam, visited weekly after our son was born to care for him and have been tireless cheerleaders. I am indebted to my sister-in-law, Allyson Lam, who practically lived with us the early months of pandemic to also help with Kota. My sister Eri has been a source of unending encouragement and support from the beginning. Thank you to my parents, Emi and Mits, for always believing in me and supporting my dreams by filling our house with all of the books and scientific gadgets that it could hold.

Thank you to my two kids, Kota and Mie. When you learn how to read, I hope you can finally understand a bit of what Dad has been working on. I love you both so much.

And finally, thank you Sarah, for supporting me each and every day of this journey. This could have never happened without your encouragement and love.

Chapter 1

Introduction

RNA sequencing (RNA-seq) has transformed our understanding of the relationships between the transcriptome and the genome. This method is most often used to quantify changes in gene expression in different conditions and has been instrumental in illuminating how alterations in the genome alter phenotype. Harnessing the power of RNA-seq data, its applications have been extended to various aspects of the transcriptome, including splicing, RNA structure, single-cell gene expression, translation, and spatial transcriptomics. Its ubiquity in molecular biology has made it an essential tool for any researcher interested in discerning how RNA biology is linked to development and disease. The motivation behind this thesis is to build methods that harness the power of RNA-seq to further reveal how dysregulation within the genome leads to cancer and other genetic diseases.

The first chapter of this thesis highlights the development of a somatic variant calling pipeline that can analyze tumor RNA-seq. I illustrate how the core components of this pipeline, RNA-VACAY, were chosen and designed. I also highlight its performance with synthetic data and a large dataset of over 1,300 samples. This pipeline can detect somatic variants well in cancer-related genes and can be used as a lightweight and lower cost approach to either confirm existing somatic variant calls or detect novel somatic mutations.

The second chapter details how splicing variant detection can be used to demonstrate how aberrant splicing could potentially contribute to diseases. I establish how MESA, our splicing analysis tool, can efficiently look for splicing changes between samples. Using RNA-seq from pediatric patients with an undiagnosed genetic disease, we were able to find multiple splicing variants of interest, including a variant in *DEGS1* that may be linked to the disease.

The third chapter focuses on my contributions to the development of a splice site CRISPR assay that seeks to model exon skipping events in cancers. We created a workflow to computationally identify potential exon skipping events in lung adenocarcinoma and create guide RNAs that serve as the backbone of the molecular assay.

These three methods together demonstrate novel applications of RNA-seq and continue to build on our current understanding of both the transcriptome and genome. The downstream applications of these methods will serve to further reveal the connections

between aberrant molecular mechanisms and genetic disease.

Chapter 2

Using RNA sequencing to detect somatic mutations

2.1 Background

The detection of somatic variants through next generation sequencing (NGS) has enabled researchers and clinicians to associate genetic mutations and disease phenotypes. The rapid improvement and falling costs of these technologies have led to the discovery of a whole host of crucial cancer-driving mutations and have opened new doors for targeted therapies in many cancers. Discovering *EGFR* mutations in lung cancer (Rusch et al., 1993) and *BRAF* mutations in melanoma (Long et al., 2011) have led directly to novel treatments (Chapman et al., 2011; Fukuoka et al., 2003) that have rede-

fined standard of care options for eligible cancer patients. Continued advances in NGS technology, particularly whole exome sequencing (WES), whole genome sequencing (WGS), and RNA sequencing (RNA-seq), have allowed researchers to generate massive amounts of NGS data and subsequently detect novel somatic variants (McKenna et al., 2010). Variant calling tools are built to differentiate somatic mutations from inherited or de novo germline mutations, neutral polymorphisms, and artifacts derived from misalignments, sequencing errors, or PCR errors (Depristo et al., 2011; Koboldt et al., 2012; H. Li, 2011; Xu et al., 2012). These existing variant callers are designed primarily to handle WES or WGS data. RNA-seq is commonly employed for gene expression and alternative splicing analyses, which has given researchers an opportunity to uncover the transcriptional and post-transcriptional phenotypes of cancer cells. The transcriptome's inherent complexity can prove to be technically challenging when detecting variants. RNA-seq data often contain reads that span intronic regions or harbor variants in genes with low expression, which pose problems for many of these current variant calling tools (Quinn et al., 2013). WES also utilizes probes designed for exonic regions, which introduces annotation biases and would miss the UTR data captured by RNA-seq. Researchers have demonstrated that identification of somatic variants in this data is possible (García-Nieto, Morrison, & Fraser, 2019; Piskol, Ramaswami, & Li, 2013; Yizhak et al., 2019), but there has yet to be an integrated pipeline with its results validated by a matched whole-genome variant list.

2.2 Somatic mutations and their importance in understanding cancer biology

Both normal cells and cancer cells are direct descendants of the fertilized egg. However, the DNA sequence within these cells has specific differences when compared to the fertilized egg genome; these differences are known as somatic mutations. Somatic mutations typically arise from replication errors or unrepaired or incorrectly repaired DNA damage. Both exogenous (UV light, chemical and radiation exposure, viruses (Talbot & Crawford, 2004) and endogenous (mitotic errors, reactive oxygen species factors) factors can cause DNA damage. Most somatic mutations found in a cell have no phenotypic effect. However, particular mutations that alter gene function or a regulatory element may allow for increased growth or survival. These driver mutations refer to positively selected mutations within a cell population (Fig. 2.1). The counterpart passenger mutations are mutations that do not give the cell a selective advantage. Cells can acquire hundreds or thousands of passenger mutations with no contribution to cancer development, but the appearance of one or more driver mutations can confer the growth advantage necessary to transform a normal cell into a cancer cell. The identification of these driver mutations and their effect on cancer cells is essential to uncovering the biological underpinnings of how these tumors arise and develop.

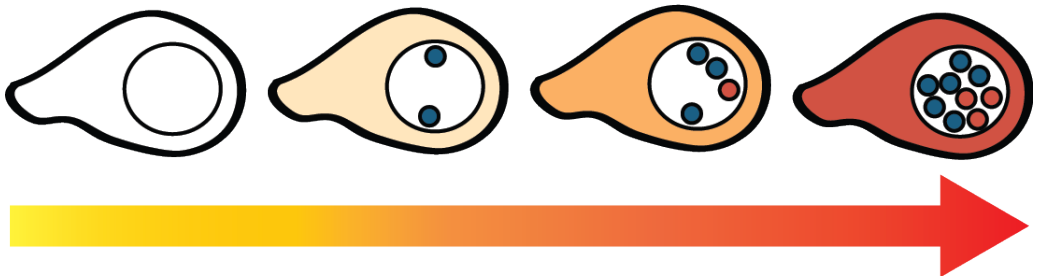


Figure 2.1: **Somatic mutations transform normal cells into cancer cells.** Normal cells accumulate mutations over time. Passenger mutations (shown in blue) are not known to be involved with oncogenesis, but driver mutations (shown in blue) are responsible for the transformation of normal cells into cancer cells.

2.3 Historic use of RNA-seq

The introduction of NGS has revolutionized our understanding of the genome. RNA-seq is the dominant technique for transcriptome profiling and the measurement of gene expression levels (Marioni, Mason, Mane, Stephens, & Gilad, 2008). Prior to the introduction of RNA-seq, cDNA microarrays were primarily used to quantify the transcriptome. Microarrays had several limitations, including requiring a priori knowledge of the genomic sequence, cross-hybridization leading to high background, and complicated normalization methods between experiments. This made it difficult to design arrays for both full sequence and transcriptome comprehensiveness and to detect splice events. RNA-seq does away with these limitations and is now routinely used to investigate differential gene expression in samples undergoing different conditions. RNA-seq also allows for the identification and quantification of isoforms and novel mRNA tran-

scripts, drastically increasing our understanding of splicing mechanisms (E. T. Wang et al., 2008) and regulation by non-coding RNAs (Djebali et al., 2012) and enhancer RNAs (Kim et al., 2010).

2.4 Previous attempts to develop/evaluate variant calling tools/workflows

The massive amounts of reads produced by NGS have given researchers an incredible source of sequencing data to better understand the genetics of cancer. Bioinformatic tools known as variant callers were initially built to distinguish these variants from noise. Typical variant calling pipelines consist of read processing, mapping and alignment of reads to a reference genome, and finally, variant calling. Sequence adapters, primers, unique molecular identifiers, and other exogenous sequences are first removed. These cleaned reads are then mapped to a reference genome and aligned. Once all reads are properly aligned, variant callers identify real variants from sequencing errors, mismapped bases, and other sources of noise. As germline variants have allele frequencies of 0.5 or 1, somatic variant callers are tasked with determining whether variants with low allele frequencies are artifacts or a rare true variant. Multiple algorithms exist to detect mutations and most prefer the inclusion of a matched normal sample during analysis. Heuristic thresholding is the most straightforward approach, applying

thresholds and statistical tests to detect potential variants. Joint genotype analysis calculates the posterior probability of the joint genotypes by Bayes' rule. Other tools extend this approach to calculate joint allele frequencies to better capture rare subclones. Haplotype-based algorithms locally assemble reads in a region using a de Bruijn graph to generate candidate haplotypes. Reads are aligned to these haplotypes and read support of each haplotype is used to calculate the likelihood of the candidate haplotypes. Recent variant callers have also incorporated machine learning techniques to classify variants as somatic mutations or artifacts. Methods such as random forest, support vector machines, logistic regression, and regression trees are trained on the features of a ground truth set of somatic variants.

Using RNA-seq as a data source for variant calling has been attempted previously. Variants detected in RNA-seq are expressed and more likely to affect the phenotype of a cell. Similarly, rare variants in highly expressed genes are less likely to be labeled as artifacts. However, there are several limitations with RNA-seq variant calling including alignment errors near splice junctions, missing variants in lowly expressed genes, variants in genes with allele specific expression, and often missing matched normal samples. Several variant callers accept RNA-seq data, including VarDict, VarScan2, RADIA (Radenbaugh et al., 2014), Seurat (Christoforides et al., 2013), Platypus, and GATK. RADIA and Seurat both use RNA-seq with matched tumor and normal DNA data to improve the variant detection performance. The remaining tools accept either

	Requires DNA	Matched Normal	Preprocessing
GATK	No	No	Yes
Platypus	No	No	Yes
VarDict	No	No	No
FreeBayes	No	No	No
Samtools	No	No	No
Strelka2	No	No	No
SNPiR	No	No	No
RADIA	Yes	Yes	No
VarScan2	No	Yes	No

Table 1: **Existing variant callers that can process RNA-seq data.** Most variant callers are built for WES or WGS data, but these 9 tools have been reported to work with RNA-seq. Some of these tools require additional data or processing, including a matched DNA or normal sample, to properly function.

DNA or RNA-seq data, but their ability to accurately detect variants in RNA-seq has not been fully explored.

2.5 Evaluation of existing variant calling tool performance with RNA-seq

We first surveyed multiple open-source variant callers (Table 1) and cataloged their relevant features, including their ability to call variants without a matched normal sample. Platypus (Rimmer et al., 2014), GATK (McKenna et al., 2010), VarDict (Lai et al., 2016), and FreeBayes (Garrison & Marth, 2012) were all evaluated for their ability to detect somatic variants in RNA-seq data. FreeBayes was quickly eliminated as an option due to its massive requirements for both time and computational resources (Fig. 2.2a). Platypus, GATK, and VarDict all have multithreading options, allowing the user to decrease the total time necessary to run each tool when using a multicore system. We first created a synthetic dataset, using normal RNA-seq aligned reads from Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG) of the International Cancer Genome Consortium (IGCG). We spiked 300 somatic SNVs into 20 of these samples to test the tools. Platypus had the best combination of recall and positive predictive value (PPV) of these three tools when analyzing this dataset (Fig. 2.2b). VarDict had the highest recall, but also detected a large number of false positives. We further curated a small subset of matched tumor and normal RNA-seq data from 8 tumor types within PCAWG to measure the performance of the tools with real world data. The variants detected in the normal RNA-seq were used to identify germline calls and potential

false positives. We compared RNA-seq-based variants with the consensus WGS variant calls from the same samples to measure recall. As expected, we saw higher recall at higher levels of expression and coverage across all tools (Fig. 2.2c). GATK had the highest recall in this analysis, but is likely due to GATK being a major component of the PCAWG WGS variant calling pipeline. Platypus had the next best performance after GATK and was again significantly faster and less resource intensive. As a result, Platypus was chosen to be incorporated into a new pipeline to detect somatic mutations in RNA-seq data.

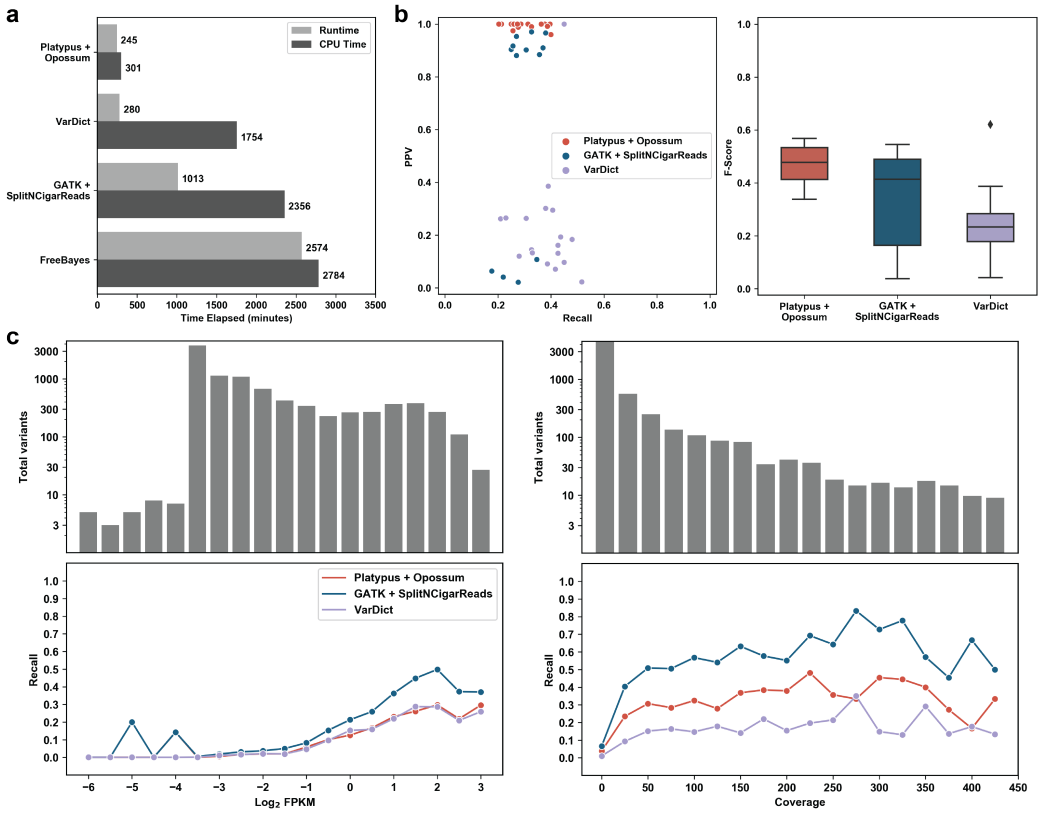


Figure 2.2: **Platypus is the tool best fit for a fast and sensitive variant calling pipeline. a**, Four variant callers were evaluated for runtime using 5 synthetic RNA-seq samples. The mean time was reported. Together Platypus and Opossum - a preprocessing step - were significantly faster and consumed less resources than all other variant callers. **b**, Three variant callers were evaluated for recall and positive predictive value (PPV) with a synthetic RNA-seq dataset of aligned reads with spiked-in somatic mutations. Platypus had the highest median F-score. **c**, The three tools were then evaluated with a small test dataset (12 samples from diverse cancer types) curated from PCAWG. All variant callers generally were more sensitive in genes with higher expression (log2 FPKM) and variants with higher coverage.

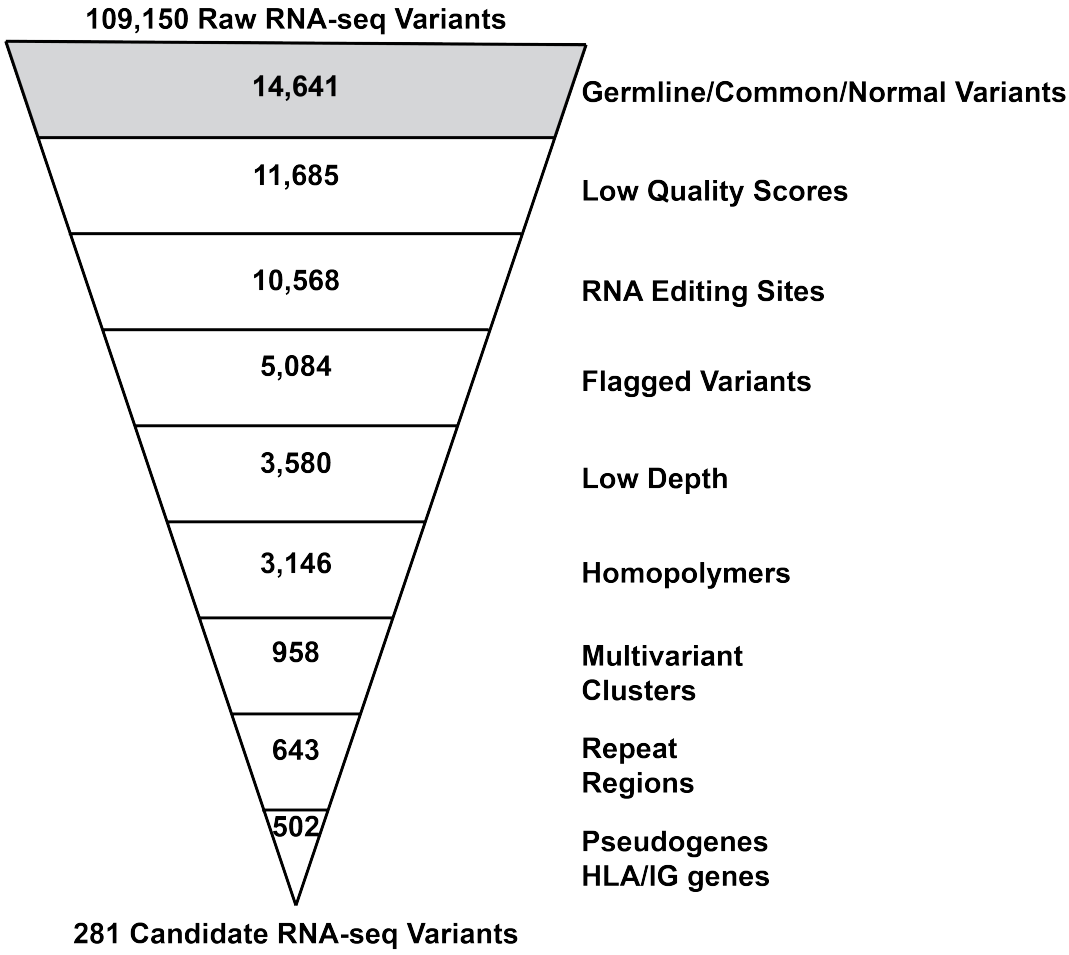


Figure 2.3: **Filtering strategy for RNA-VACAY.** An example filtering strategy for a single lung squamous cell carcinoma sample. RNA-VACAY uses multiple filters to remove false positives from the raw variant candidates. The majority of these variants are removed using a combination of either germline, common, or normal variant databases.

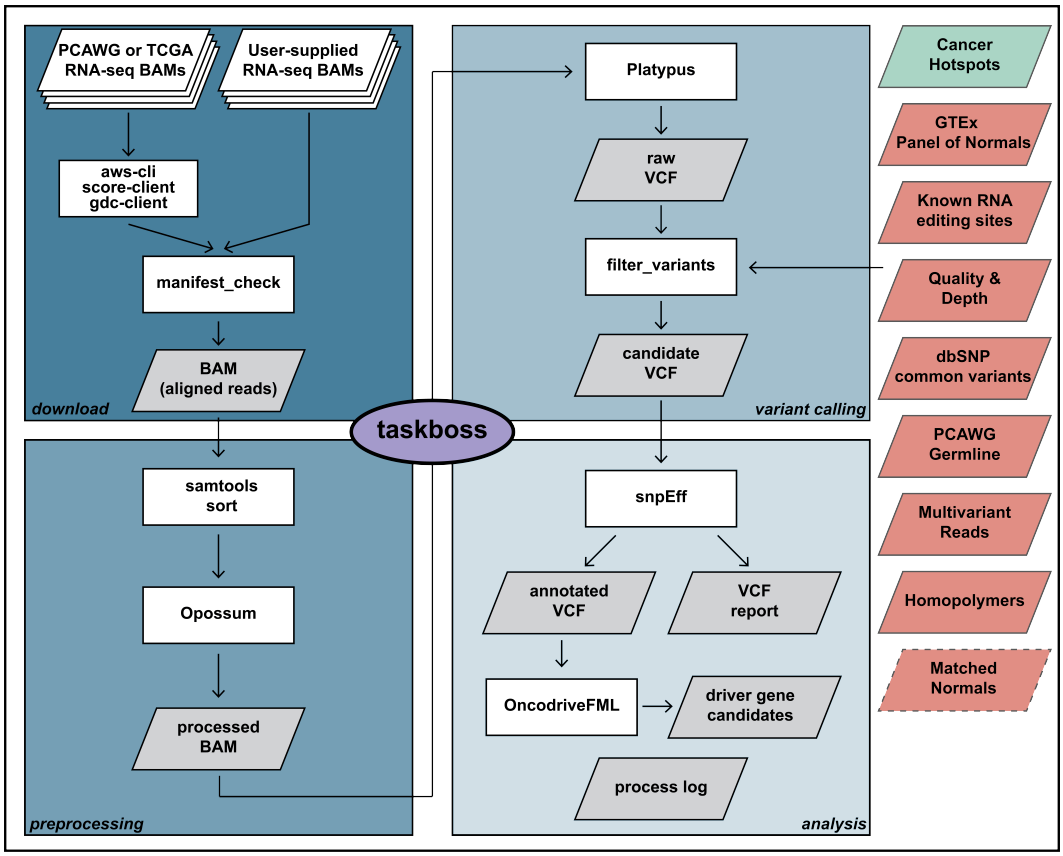


Figure 2.4: **A scalable variant calling pipeline to detect somatic variants in RNA-seq data.** Schematic of RNA-VACAY, a somatic variant calling pipeline designed for RNA-seq data. The pipeline consists of 4 main modules - data download, preprocessing, variant calling, and analysis. Taskboss, a controller module, assigns tasks according to available resources and can assist in resuming pipeline operations after interruptions. Multiple filtering options (green and red) can be toggled to keep known cancer mutations and remove likely false positives. Variants from matched normal samples, if available, can also be used to filter common variants. White boxes refer to components of the pipeline and gray parallelograms refer to outputs generated by the pipeline.

2.6 Variant calling pipeline design

Since PCAWG samples have both whole genome and RNA-seq data, we used this data as an opportunity to benchmark the RNA variant callers by comparing variant calls from RNA-seq data with known somatic variant calls from the whole genome sequencing data. Unlike previous whole exome variant calling studies, this allows us to eliminate annotation biases associated with WES (Barbitoff et al., 2020). We first broadly examined the performance of our initial pipeline detecting variants in known cancer-associated genes within a specific cancer type. We previously found that a large number of variants reported by Platypus alone did not replicate the consensus variants found in the matched WGS data. Preliminary analysis of the candidate variants revealed significant amounts of noise, particularly around insertions and deletions and in specific regions of the genome. Gene ontology analysis was performed on potential false positive variants and a striking number were found in immunoglobulin (Ig) and human leukocyte antigen (HLA) genes. Transcripts from these genes often feature extreme levels of diversity, making accurate mapping to these regions difficult for most tools (Degner et al., 2009; Watson & Breden, 2012). Previous studies also applied similar filters, such as removing variants found at known RNA editing sites and near splice junctions (García-Nieto et al., 2019). Other potential false positives were also found nearby homopolymer tracts and on reads with multiple variants in close proximity (within 50bp). These events were deemed to be likely sequencing or alignment

artifacts and were removed from the candidate variant list. These filtering steps were incorporated into the final variant calling pipeline, which we have titled RNA-VACAY (Fig. 2.3). Our pipeline can download aligned reads from popular cancer data repositories or accepts a manifest of aligned read data already in a specified location on the user's computer. Once the reads are available, the pipeline automates the entire variant calling process and delivers a list of candidate variants. RNA-VACAY is easy to use, scalable, and resource efficient (Fig. 2.4).

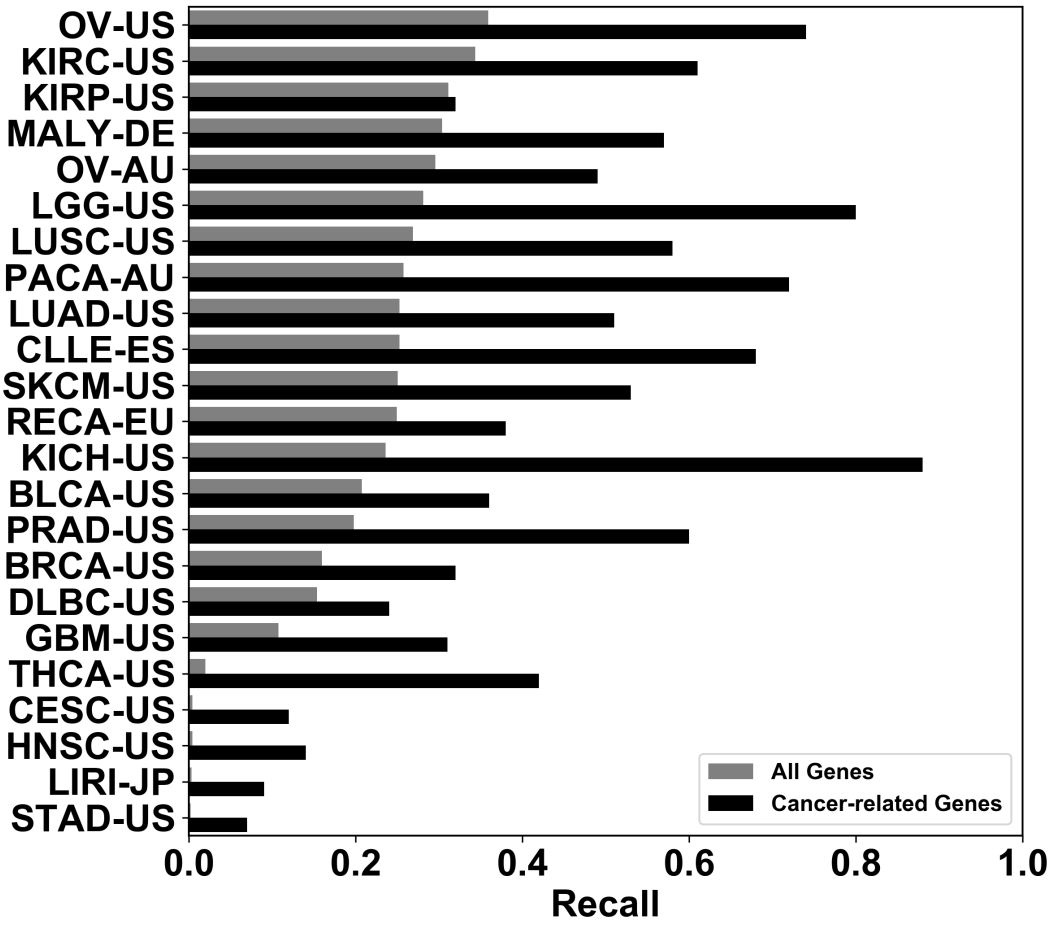


Figure 2.5: **Increased recall when focused on cancer-related genes.** Median recall across all tumor types was 0.25. Focusing on a subset of genes provided by the Cancer Gene Census, median recall increases to 0.48.

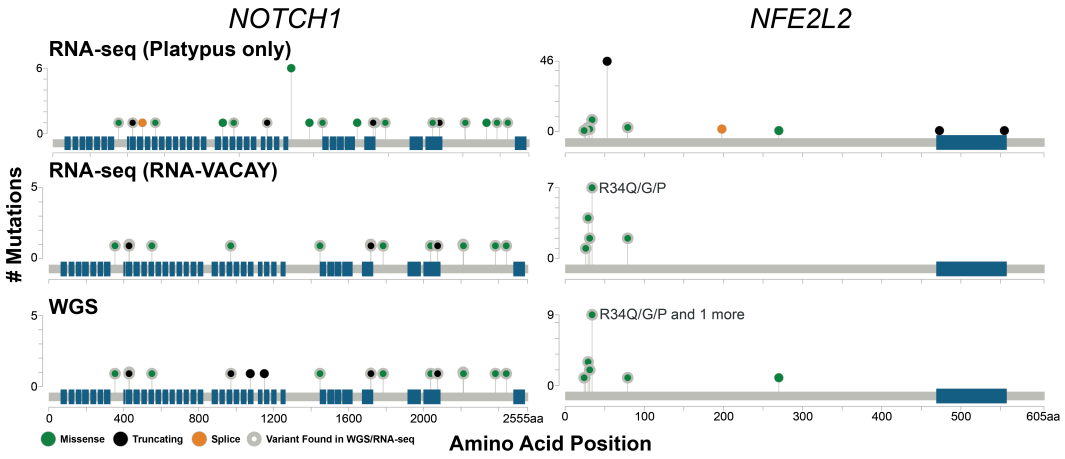


Figure 2.6: **RNA-VACAY identifies cancer-related variants found in WGS data at the gene level.** Stickplot of RNA-seq and WGS mutations found in the genes *NOTCH1* and *NFE2L2* of lung squamous cell carcinoma samples. *NFE2L2* R34* is a known hotspot mutation.

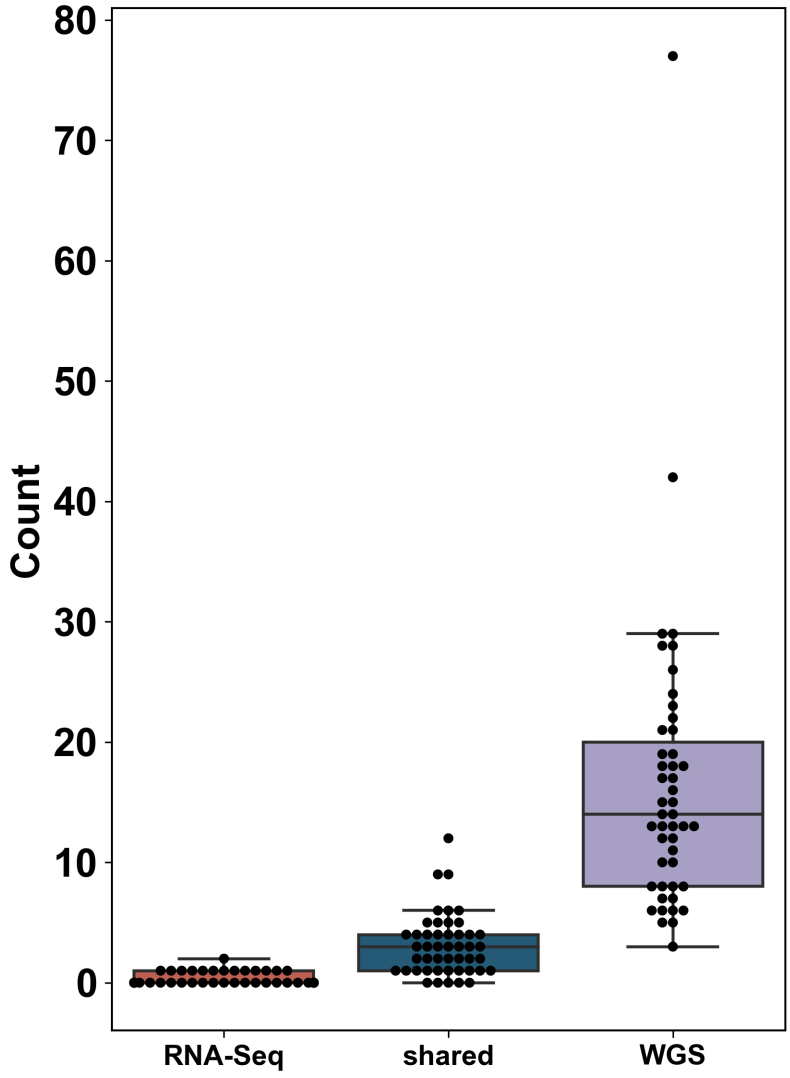


Figure 2.7: **Potential stop codon creating variants not detected in RNA-seq data.** In lung squamous cell carcinoma samples, the median count of stop codon creating variants detected in WGS data alone was 14, while the median count from RNA-seq alone was 0. The median count of stop codon creation variants found in both datasets was 3.

2.7 Gene-level performance

We found that the majority of variants reported by RNA-VACAY matched their whole genome counterparts and significantly lowered the number of false positives. For example, in lung squamous cell carcinoma (LUSC), we detected 3,577,489 variants across the entire cohort using Platypus alone. Our pipeline delivered 7,326 candidate variants; 4,319 candidate variants were found in the WGS data. When we subsetted for only cancer-related genes in this cohort, we found 241 candidate variants were found in both RNA-seq and WGS data. 177 variants found in the WGS data were not detected by RNA-VACAY, showing a marked increase in recall. This finding was mirrored across all tissue types in the study (Fig. 2.5). We specifically looked at the performance of the pipeline in two genes, *NOTCH1* and *NFE2L2*, that have been linked to cancer formation in LUSC (Fig. 2.6). While the variants reported by Platypus alone point to a false hotspot mutation, RNA-VACAY largely replicated the WGS mutations found in *NOTCH1*. The *NFE2L2* R34 hotspot mutation was detected with RNA-VACAY with no false positives across the rest of the gene. Of the missed variants, many resulted in truncations or stop codon creation (Fig. 2.7), which in turn commonly lead to degradation of the transcript by nonsense-mediated decay (Amrani et al., 2004); therefore, expression of these variants is low and subsequently were not detected by our pipeline. For example, truncating mutations in *TP53* reported in lung adenocarcinoma (LUAD) WGS data were not identified by RNA-VACAY. Many driver genes are often highly

expressed (Ohshima et al., 2017) and therefore our pipeline detects these high impact variants with confidence.

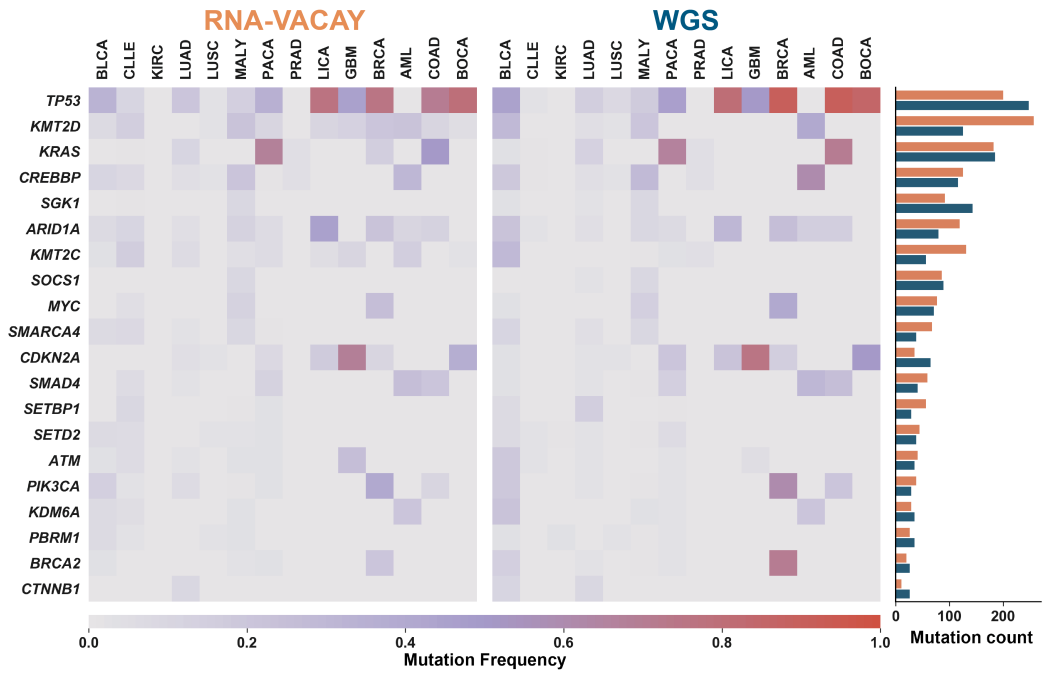


Figure 2.8: **Mutation frequencies of cancer-related genes are similar in RNA-seq and WGS** Heatmap compares frequency of samples in a tissue type containing mutations in known cancer-associated genes. Barplot displays the total number of mutations found in each gene - RNA-VACAY variants are orange and WGS variants are blue.

2.8 Cancer-type performance

Using this finalized pipeline, we detected 161,809 single nucleotide somatic variants in all 1,403 RNA-seq samples from the PCAWG dataset (PCAWG Transcriptome Core Group et al., 2020). We surveyed several known cancer-associated genes with published hotspot mutations (Chang et al., 2016) (*KRAS* G12 (Riely et al., 2008), *BRAF*

V600 (Long et al., 2011), *PIK3CA* H1047 (Mandelker et al., 2009), etc.) in multiple cancer types to assess the pipeline performance in all of the cancer types analyzed by PCAWG. Over 78% of the WGS calls were detected by RNA-VACAY, demonstrating the pipeline's ability to detect these variants in cancer-related genes while analyzing only RNA-seq data across different tumors (Fig. 2.8). *TP53*, the most frequently mutated gene in this study, recapitulated the WGS mutational frequency profile and demonstrated similar high mutational frequencies in liver cancer, colon adenocarcinoma, bone cancer, and breast adenocarcinoma using RNA-seq variants. Similarly, mutation frequencies and counts in *KRAS*, *MYC*, *CREBBP*, and *SOCS1* were very similar in both RNA-seq and WGS data. Both *KMT2D* and *ARID1A* surprisingly had a larger share of RNA-seq only variants. After individual confirmation with the WGS aligned reads, the variants were present in both datasets, suggesting that these particular WGS variants were removed during the consensus variant calling process in WGS analysis.

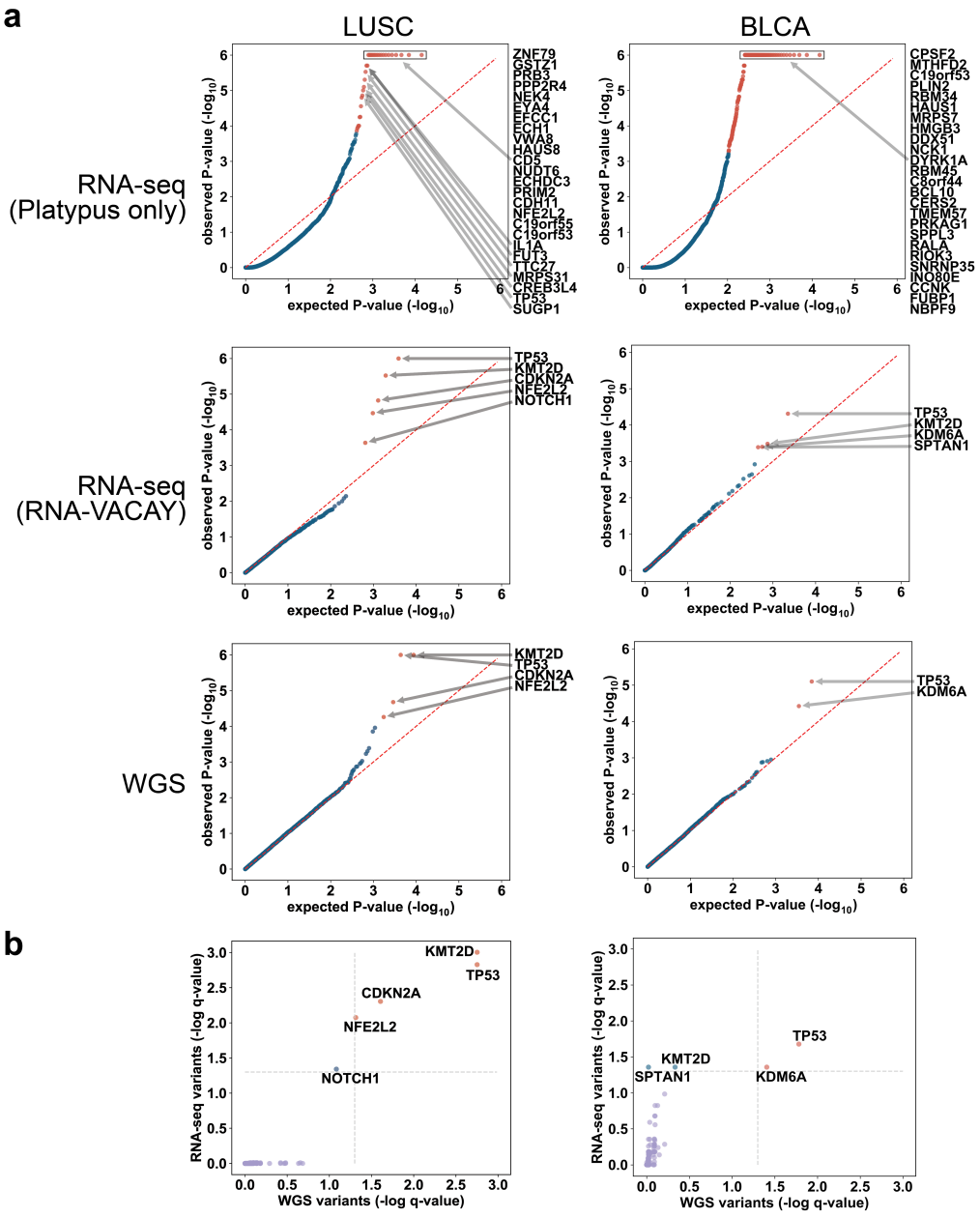
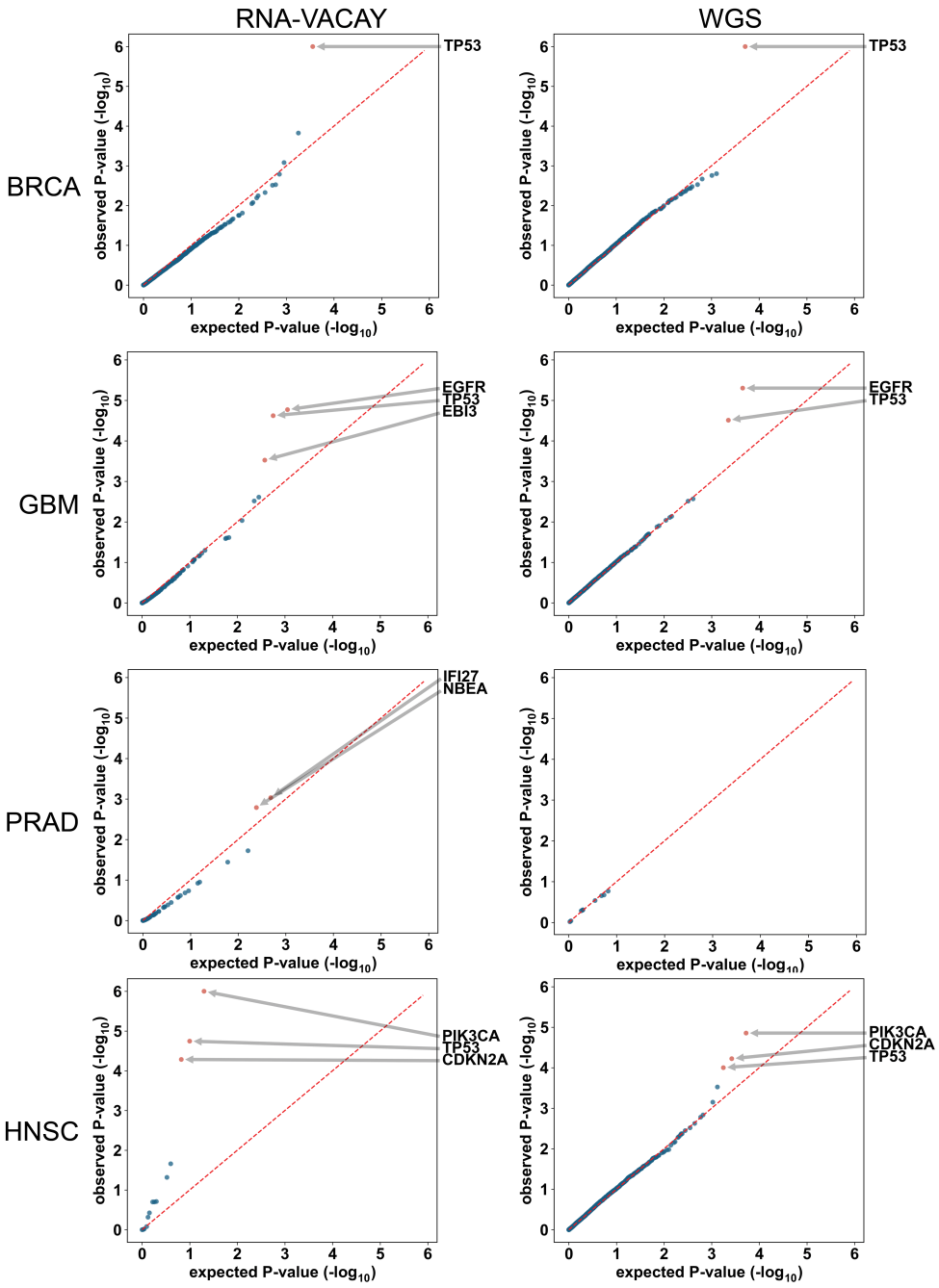
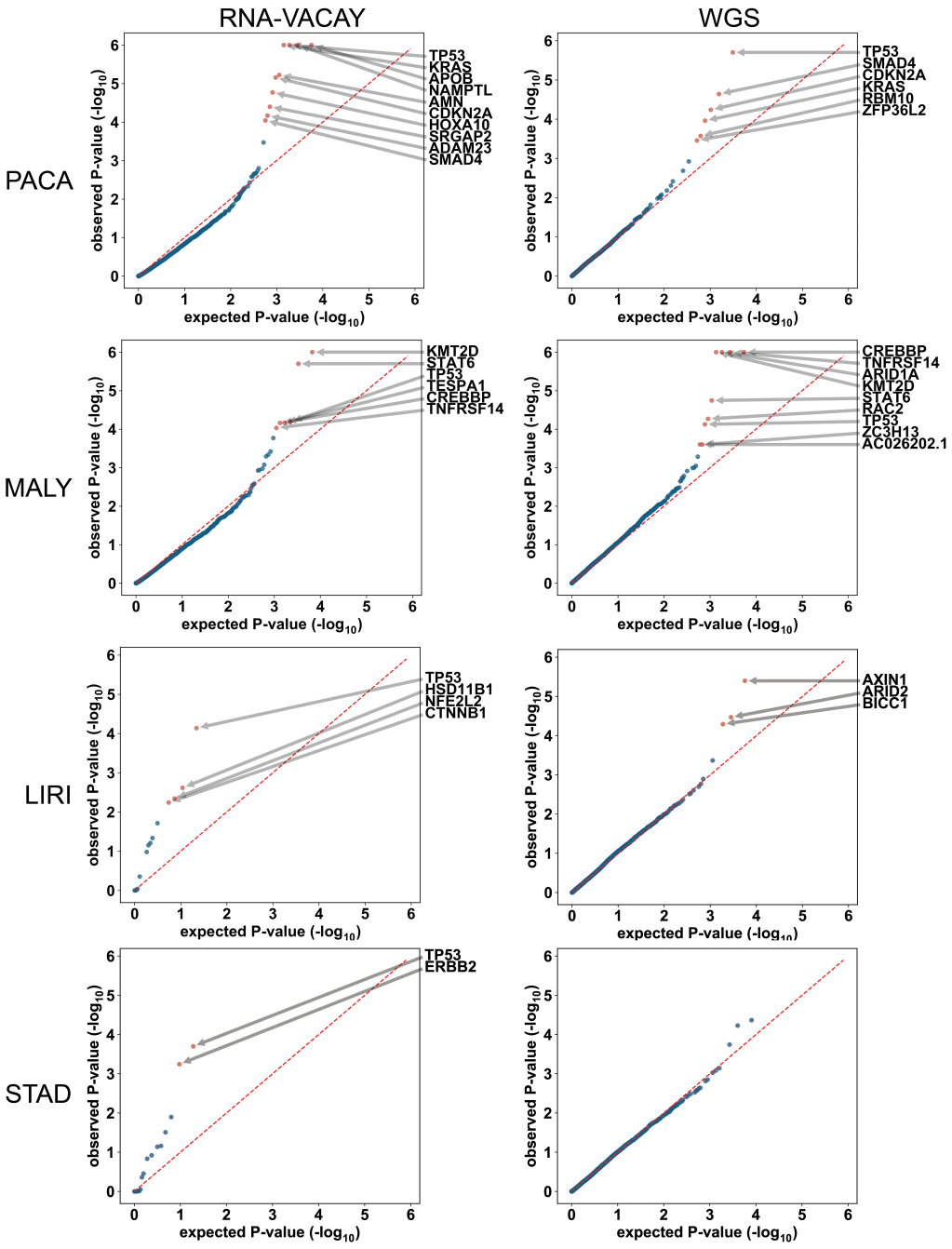


Figure 2.9: **Driver mutation profiles from RNA-VACAY variants match WGS. a**, Quantile-quantile (QQ) plots of p-values generated from oncodriveFML. Red line displays where observed and expected p-values match. Blue dots represent genes with Q-values ≥ 0.1, red < 0.1. Genes with Q-values < 0.1 are indicated. **b**, Plots of q-values from genes with detected RNA-seq and WGS variants. Quadrants show significance in RNA-seq, WGS, or both.


RNA-VACAY
WGS
BRCA
GBM
PRAD
HNSC
observed P-value (-log10)
expected P-value (-log10)
TP53
EGFR
EBI3
IFI27
NBEA
PIK3CA
CDKN2A


RNA-VACAY
WGS
PACA
MALY
LIRI
STAD
observed P-value (-log10)
expected P-value (-log10)
TP53
KRAS
APOB
NAMPTL
AMN
CDKN2A
HOXA10
SRGAP2
ADAM23
SMAD4
TP53
SMAD4
CDKN2A
KRAS
RBM10
ZFP36L2
KMT2D
STAT6
TP53
TESPA1
CREBBP
TNFRSF14
CREBBP
TNFRSF14
ARID1A
KMT2D
STAT6
RAC2
TP53
ZC3H13
AC026202.1
TP53
HSD11B1
NFE2L2
CTNNB1
AXIN1
ARID2
BICC1
TP53
ERBB2

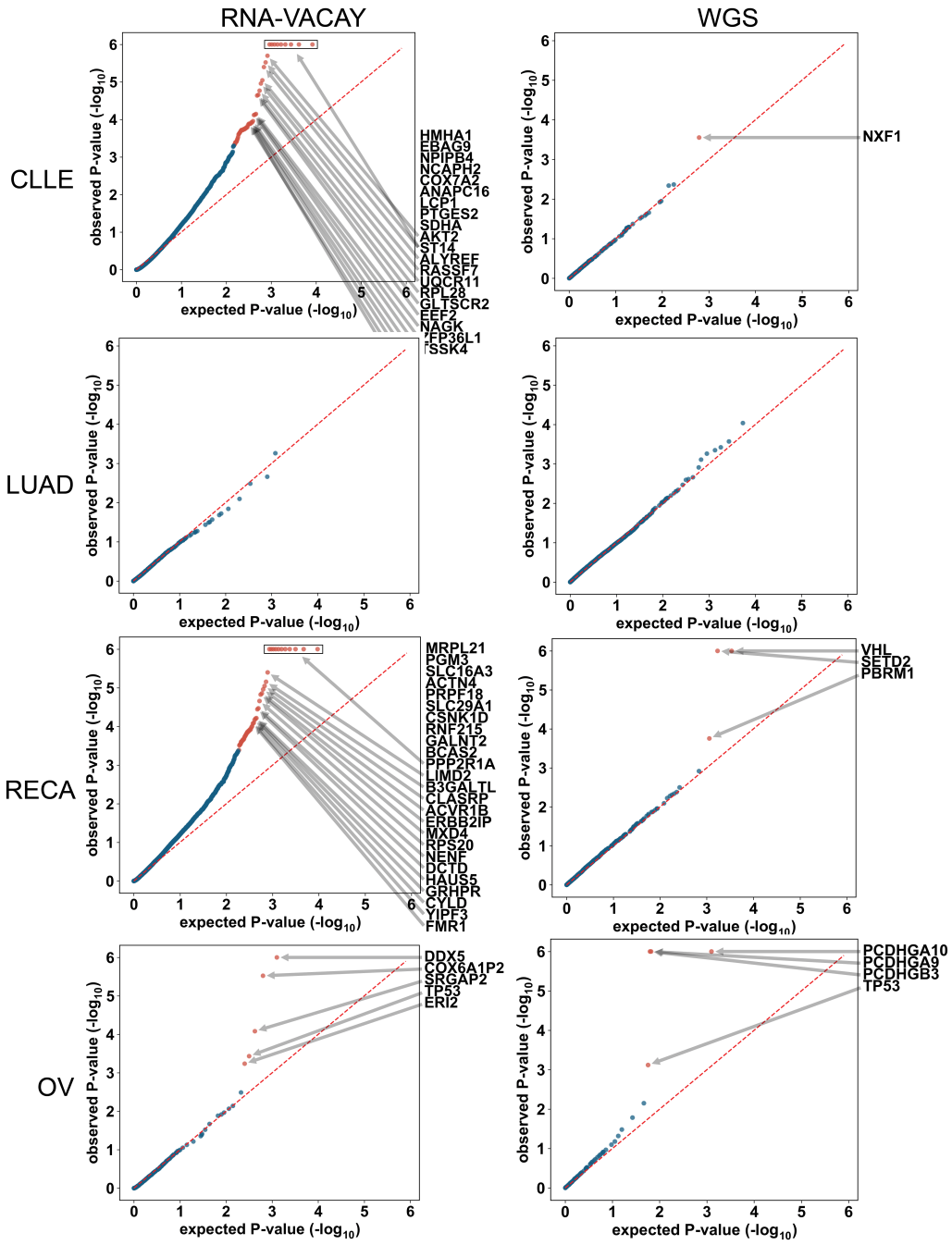


Figure 2.10: **Driver analysis for other cancer types.** Quantile-quantile (QQ) plots of p-values generated from oncodriveFML across 12 other cancer types. Red line displays where observed and expected p-values match. Blue dots represent genes with Q-values ≥ 0.1, red < 0.1. Genes with Q-values < 0.1 are indicated.

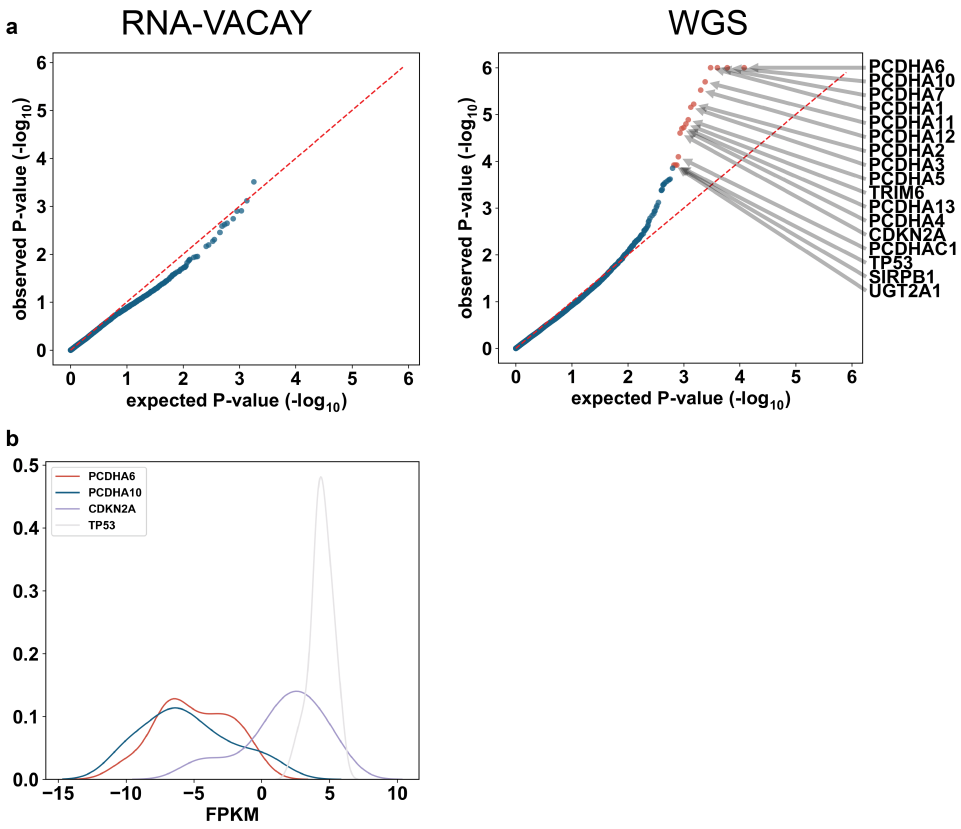


Figure 2.11: **Potential driver genes contain variants found in lowly expressed genes.** **a**, Quantile-quantile (QQ) plots of p-values generated from oncodriveFML in skin cutaneous melanoma (SKCM). **b**, Density plot of gene expression of top WGS driver gene candidates in SKCM samples.

2.9 Driver analysis

The advent of next generation sequencing technologies have revealed an entirely new landscape of somatic mutations linked to tumors. The majority of these mutations are passenger mutations, which mostly have no functional consequence. Identifying which of these mutations are driver mutations continues to be a huge hurdle for cancer

researchers. Driver mutations are mutations that undergo positive selection within a cell population and give these cells a selective advantage that often leads to abnormal proliferation and growth. Currently, many methods that detect genes with abnormal mutational patterns often focus on calculating the mutational frequency of a tumor cohort. If the mutational frequency is significantly higher than the background mutation rate, this is an indicator of positive selection and potentially, the existence of a driver gene.

In order to assess our pipeline's ability to detect cancer driver genes, we used oncodriveFML (Mularoni, Sabarinathan, Deu-Pons, Gonzalez-Perez, & López-Bigas, 2016) to compare the driver mutation profiles of matched RNA-seq and WGS samples in multiple cancer types. OncodriveFML predicts which genes harbor driver mutations using functional impact scores derived from the Combined Annotation Dependent Depletion (CADD) tool. The mean functional impact (FI) score of the mutations within a gene are compared with the distribution of mean functional impact scores of randomly generated mutations. Genes with significant differences in FI scores are likely to be driver genes. We used single nucleotide variants in coding regions from the RNA-seq data across all tumor types to generate driver gene profiles and compared these profiles to their matched WGS samples. The driver mutation profile of RNA-seq variants called by Platypus alone initially resulted in a multitude of potential driver genes, which can be attributed to the inclusion of germline or false positive variants (Fig. 2.9a). However,

RNA-seq variants called from our RNA-VACAY pipeline are often consistent with their WGS equivalents (Fig. 2.9a,b, 2.10). Known driver genes were identified such as *TP53*, *KMT2D*, *CDKN2A*, and *NFE2L2* in both LUSC RNA-seq and WGS data. *NOTCH1*, another cancer-related gene, was also predicted to be a driver gene using RNA-VACAY variants, but not WGS. Similarly, *TP53* and *KDM6A* were reported to have driver mutations in bladder adenocarcinoma RNA-seq and WGS data. *SPTAN1* and *KMT2D* were also predicted to be driver genes from RNA-VACAY variants, but not WGS. The mutations in *NOTCH1*, *SPTAN1* and *KMT2D* detected by RNA-VACAY were also found in the WGS consensus calls. However, additional synonymous mutations found in the WGS lower the mean FI score of those genes. Interestingly, the mutations in *NOTCH1* and *TP53* not found in the RNA-seq data are either synonymous or missense mutations, suggesting that the variants are not expressed and may not be functional. *TP53*, the gene with the most recurrent mutations, was most commonly reported as being a driver gene in 13 cancer types using WGS variants. RNA-VACAY delivered the same finding in 11 cancer types. In chronic lymphocytic leukemia (CLLE) and renal cell carcinoma (RECA), there were significantly more driver gene candidates found in the RNA-VACAY variants than the WGS variants (Fig. 2.10). Upon inspection, there was a significantly higher number of variants found on the same reads and in close proximity to one another, which may point to either a technical artifact introduced during sample preparation or alignment. However, in cancer types with a high mutation frequency

such as skin cutaneous melanoma (SKCM), we saw less overlap between the RNA-seq and WGS data; multiple genes from the protocadherin alpha gene cluster (*PCDHA6*, *PCDHA10*, *PCDHA7*, *PCDHA1*, etc.) were reported as potential driver using the WGS data (Fig. 2.11a). These genes are lowly expressed in this cancer type, which could explain why RNA-VACAY was unable to detect these variants (Fig. 2.11b). Overall, 16 cancer types reported one or more driver genes using WGS variants and RNA-VACAY was able to detect at least 1 matching gene in 13 of them. RNA-VACAY described 1 or more potential driver genes in 15 of 16 cancer types that were also listed in the Cancer Gene Census, a curated database of mutations implicated in cancer (Tate et al., 2019). The driver gene profiles generated from somatic variants detected by RNA-VACAY largely match the driver gene profiles generated from variants found in the corresponding WGS data, demonstrating the ability to use RNA-seq alone to find driver genes.

2.10 5' and 3' UTR analysis

Previous PCAWG studies identified recurrent noncoding point mutations in multiple genes as being strong candidate drivers (Rheinbay et al., 2020). As RNA-seq captures both 5' and 3' untranslated regions (UTRs), we decided to test RNA-VACAY's ability to detect these same UTR mutations. Somatic variants in the 5' UTR of *MTG2* and 3' UTR of *TOB1* and *NFKBIZ* were detected by RNA-VACAY (Fig. 2.12). RNA-VACAY was unable to detect 5' UTR mutations in *PTDSS1* and *DTL*, as the vast major-

ity of RNA-seq samples had virtually no aligned reads in the specified region of those genes. This may be because somatic variants in this region can often downregulate or upregulate the expression of these genes, particularly in a cancer context (Lim et al., 2021). An alternative explanation is that RNA-seq data can exhibit a 3' end coverage bias due to the cDNA amplification process, resulting in reduced 5' UTR coverage. Provided there is satisfactory coverage, RNA-VACAY is successfully able to detect recurrent UTR variants.

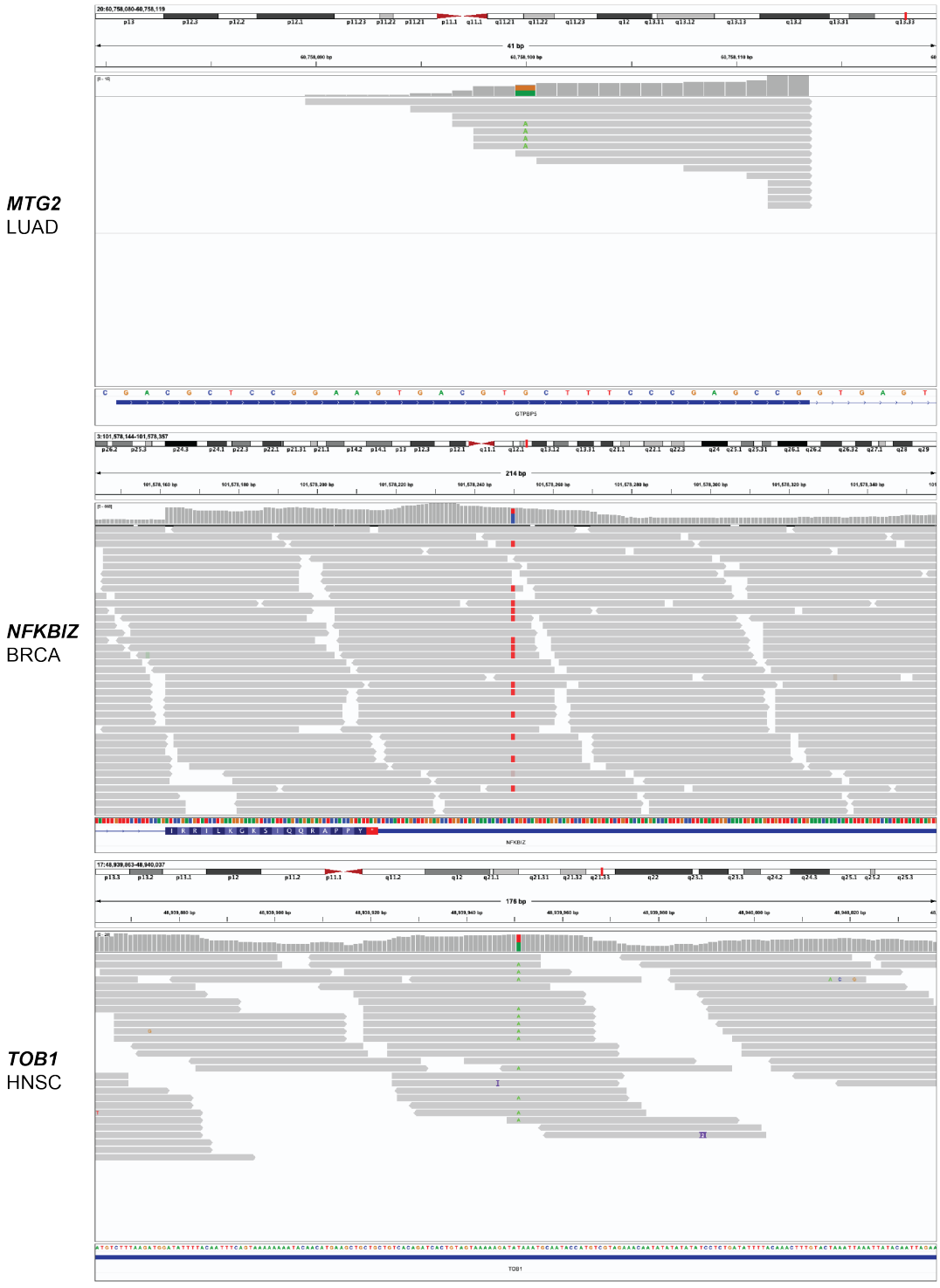


Figure 2.12: **RNA-VACAY detects 5' and 3' UTR driver mutations.** RNA-VACAY detected previously discovered candidate driver mutations in the 5' UTR of *MTG2* and and 3' UTR of *TOB1* and *NFKBIZ*.

2.11 Cost savings associated with using RNA-seq for somatic variant calling

RNA-VACAY is capable of harnessing existing RNA-seq data and provides a cost-effective and reliable option for the validation of variants found through other methods (Fig. 2.13). RNA-VACAY only requires the use of a tumor RNA-seq sample, unlike many WES and WGS methods that require both a tumor and a matched normal sample. RNA-seq data are also significantly smaller in size, so storage requirements for this data are much less demanding. For the size of the PCAWG study of 1,349 samples, we estimate that the cost of detecting somatic mutations from RNA-seq of only tumor samples to be $592,487, while the cost of WGS is estimated at $1,472,926 (DNA Technologies Core, n.d.; Yung et al., 2017). Using RNA-seq would also cut the runtime from 68,326 hours to 16,275 hours.

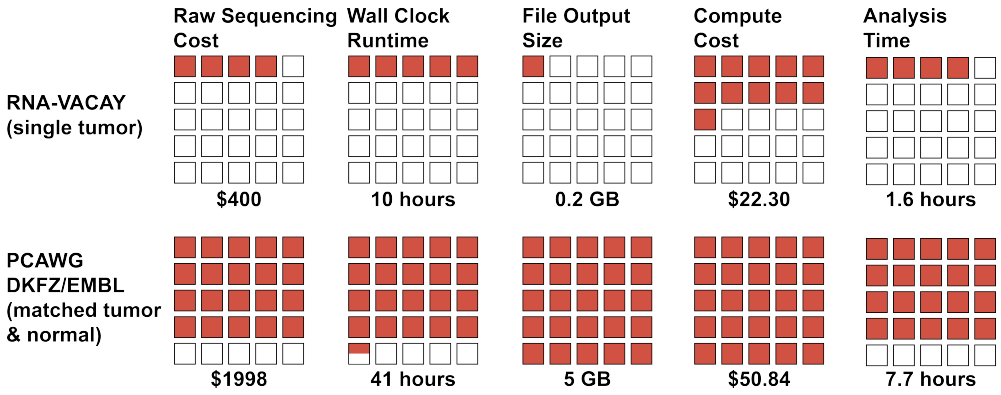


Figure 2.13: **RNA-VACAY lowers the cost of variant calling.** Variant calling with RNA-seq data significantly lowers time and cost constraints of a randomly selected sample compared to WGS sequencing.

2.12 Conclusion

WES and WGS continue to be the main sources of genomic data for identifying cancer-associated somatic variants, but the function and cost of RNA-seq make it an increasingly attractive option for characterizing tumors. In situations where no WES or WGS data are available, existing RNA-seq data collected for differential gene expression or gene fusion analysis can also be used for somatic variant detection. We applied RNA-VACAY to over 1,300 RNA-seq samples and were able to detect somatic variants with high recall. Across all samples, the median recall was 0.25, but increases to 0.48 when looking specifically at cancer-related genes (Fig. 2.5). Our study demonstrates that RNA-seq data can function both as a supplement and as a substitute for WES and WGS data when detecting somatic variants. These variants were detected in actively expressed regions, so they are more likely functionally relevant and significant. RNA-VACAY has demonstrated its ability to detect somatic variants in RNA-seq that match the driver gene profiles of variants detected in WGS.

Using RNA-seq also allows for the discovery of somatic variants in the 5' and 3' UTR, allowing for further discovery of the functional impact of these noncoding variants. While we currently filter out previously identified RNA editing sites, future applications of our pipeline could also be to measure the RNA editing profile of a transcriptome or detect novel RNA editing sites. However, our pipeline is also limited by the biological underpinnings of RNA-seq. Variants in lowly expressed genes or that

decrease expression are difficult to detect. Genes with tissue-specific expression can also make variant discovery challenging.

Our pipeline does not currently detect insertions and deletions. The preprocessing step with Opossum has only been evaluated in the context of single nucleotide variant detection. Platypus has been reported to detect indels in WES and WGS data, so extending the scope to detect indels would be a natural goal for updated versions of this pipeline. A consensus strategy, incorporating multiple variant callers into the pipeline, could also be used to increase both recall and PPV.

Tumor-only sequencing can also misidentify germline variants as being somatic variants. Our filtering approach utilizing multiple public variant and mutation databases is designed to minimize this scenario. Sequencing adjacent normal tissue can decrease the number of inaccurately defined somatic variants. Our driver analyses of cancer cohorts also decrease the chances of a rare germline mutation being identified as a significant somatic mutation. Using RNA-VACAY also lowers the cost and time often necessary for somatic mutation detection.

Next generation sequencing technologies continue to enter into the clinic and have become the gold standard in the genetic diagnosis of cancer and other genetic diseases. The importance of RNA-seq as a clinical diagnostic tool requires robust and straightforward pipelines such as RNA-VACAY to automate analysis of this data.

2.13 Methods

2.13.1 Aligned reads processing

We used data from The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) of the International Cancer Genome Consortium (ICGC). We downloaded 1,349 RNA-seq samples from the PCAWG data portal. This dataset features 30 cancer types. 161 of these samples originate from normal solid tissue or tissue adjacent to the tumor. These reads were aligned with STAR (v2.4.0i) and hs37d5 or Gencode (release 19) as the reference gene annotation. Matched WGS data for these samples were used to evaluate pipeline performance. To generate a synthetic dataset, 300 randomly selected somatic single-nucleotide variants (SNVs) were manually added to aligned reads from 20 PCAWG normal tissue RNA-seq samples to simulate tumor RNA-seq data. All variants were located in the coding regions of the genome and had random allele frequencies. A test dataset of 8 donors with matched tumor and normal RNA-seq from 8 tumor types were curated and used to evaluate the performance of the variant calling tools. Normal tissue samples were downloaded from the Genotype-Tissue Expression (GTEx) project (Ardlie et al., 2015) portal and re-aligned with identical parameters. Somatic variants detected in these samples were then used to generate a panel of normal variants. 2 samples from 11 different tissue types were incorporated into this panel.

2.13.2 Pipeline description

The RNA-VACAY pipeline is a modular workflow built on Python 2.7 that automates task assignment, downloading and preprocessing data, tool execution, and variant analysis. Each task is completed within a Docker container.

1. **Data retrieval**

   This pipeline is built to specifically handle RNA-seq reads aligned with STAR (add link to document with commands) currently stored in either PCAWG or TCGA repositories. The download module includes the recommended tools by each consortium. RNA-VACAY accepts file manifests generated by these data repository portals and automates downloading. It can also accept user-generated file manifests to call variants in previously downloaded data.

2. **Preprocessing data**

   Aligned reads are sorted and indexed by samtools (H. Li et al., 2009) if necessary and then preprocessed with Opossum (v0.2) (Oikkonen & Lise, 2017). Opossum prepares RNA-seq data for variant calling by Platypus, GATK, and other callers. It splits reads mapped across splice junctions and ensures that minimal information is lost at read ends by merging overlapping reads and modifying base qualities at the edges of these reads. Opossum also eliminates duplicate reads.

3. **Variant calling**

Platypus is a Bayesian haplotype-based variant caller that uses local de novo assembly and realigns sequences to detect variants. Platypus also shares variant information between multiple samples, increasing the confidence of calls that are weakly supported in one sample, but strongly supported in related samples.

4. **Filtering**

   Raw variant calls from Platypus are first filtered with a custom panel of normal variants generated from RNA-seq samples from the GTEx repositories. Subsequent filters incorporate a combination of preexisting common and normal variant databases - dbSNP (Sherry et al., 2001), gnomAD (Karczewski et al., 2020), and REDIportal (RNA editing sites) (Picardi, D'Erchia, Lo Giudice, & Pesole, 2017). Variants with low quality scores or sequencing depth (< 7) were filtered. Variants found in certain locations, such as known decoy regions, and repeat regions were excluded. Variants found in human leukocyte antigen genes, immunoglobulin genes, and pseudogenes were also excluded. Variants found within 50 bases of other variants with similar allele frequencies or within 10 bases adjacent to homopolymer tracts of 5+ bases were also excluded. An optional filter will prevent removal of variants found in known cancer hotspots, regardless of call quality. Normal variants from matched normal RNA-seq samples, if available, can also be incorporated as an optional filter.

5. **Annotation and analysis**

Filtered variants were annotated with SnpEff (v4.3t) (Cingolani et al., 2012). SnpEff categorizes the variants based on their genomic locations and predicts the coding effects of these variants. These candidate variants were then analyzed using custom Python scripts. Driver analysis was performed by oncodriveFML. OncodriveFML calculates a profile of somatic mutations in specific genomic regions and identifies genes that have a higher mutational frequency compared to their background mutation rate. All calls outside of the coding region and any non-single nucleotide variants were filtered before running oncodriveFML.

2.13.3 Initial variant caller evaluation

We first evaluated four open-source variant callers previously reported to be compatible with RNA-seq data – Platypus (v0.8.1.1) , GATK (v4.1.9), VarDict (v1.5.5), and FreeBayes (v1.1). We ran the tools using default recommended options and recommended preprocessing steps for Platypus (Opossum (Oikkonen & Lise, 2017)) and GATK (SplitNCigarReads). We measured the speed, recall, PPV, and resource requirements of the four variant callers processing 10 RNA-seq samples, comparing the results between the 5 pairs of normal and tumor samples.

2.13.4 PCAWG RNA-seq data analysis

RNA-seq samples were downloaded as cohorts based on cancer type. RNA-VACAY was run on multiple OpenStack instances in parallel. Custom python scripts were developed to handle and aggregate results.

2.13.5 Single gene variant comparison

The variants in particular genes were visualized as stickplots with cBioPortal (Cerami et al., 2012; Gao et al., 2013). Known cancer-related genes in specific cancer types were chosen and plotted using the MutationMapper tool. Custom python scripts were written to analyze the overlap between the RNA-VACAY and WGS variant sets.

2.13.6 Cancer type variant comparison

The RNA-VACAY and WGS mutational frequencies of the 25 most mutated cancer-related genes were compared across each PCAWG tumor type using a custom python script.

2.13.7 Driver mutation profiling

OncodriveFML (v2.2.0) was used to identify genes with potential driver mutations. We ran oncodriveFML with default settings on filtered variants, using the whole-exome sequencing option.

2.13.8 5' and 3' UTR mutation confirmation and visualization

Custom python scripts were written to uncover variants detected by RNA-VACAY that matched previously published genes with recurrent 5' and 3' UTR mutations. We used Integrated Genomics Viewer (v2.8.3) (Robinson et al., 2011) to visually confirm the variants.

Chapter 3

Using RNA sequencing to detect splicing variants of interest

3.1 Background

During RNA splicing, introns are removed from the precursor messenger RNA (pre-mRNA) and exons are joined together to form a mature transcript. Exons and introns can be differentially included (Berget, Moore, & Sharp, 1977) and excluded to create multiple transcripts from a singular template DNA. Alternative splicing (AS) is one of the main drivers of transcriptome complexity. It is a highly regulated cellular mechanism where pre-mRNA is processed into different mature mRNA molecules. 95% of genes in humans undergo some level of alternative splicing (Pan, Shai, Lee, Frey, &

Blencowe, 2008). When alternative splicing occurs, the splicing mechanisms can create multiple protein products from a single gene by modifying the splicing of the exons. These transcripts or isoforms might encode functionally different proteins, which can depend on cell type or environment. These proteins may be missing a transactivation domain, lack a DNA-binding domain, have altered affinities in their binding sites, or localize to a different region of cell. Genetic diseases associated with alternative splicing events mainly arise from changes in core splicing consensus sequences. Mutations in or near splice sites that alter the 5' GT or 3' AG can cause usage of cryptic splice sites, intron retention, or exon skipping. Mutations can also create new splice sites or reduce the strength of existing splice sites by altering the surrounding context. Altered splicing can also create frameshifts that lead to degradation of mRNA by nonsense-mediated decay (Sun, Zhang, Sinha, Karni, & Krainer, 2010). Changes in the core spliceosome machinery, enhancer and repressor sequences, RNA polymerase II, and histone modifications also alter splicing regulation and lead to disease.

3.2 Calculating percent spliced with existing tools

Historically, alternative splicing was measured by the percent spliced in (PSI) value. This score is a ratio of inclusion reads (i.e. reads overlapping with exons) and exclusion reads (i.e. reads spanning splice junctions) that summarizes the AS events across individual exons. Percent spliced (PS) is similar to PSI, but instead of focusing on exon

counts, PS uses junction counts exclusively. The two values can be similar in situations featuring alternative splice sites or mutually exclusive exons. However, these values will differ significantly when quantifying complex splicing event combinations or exon skipping events. Junction-based quantification has a distinct advantage over exon-based quantification as junctions better reflect RNA processing. Information regarding splicing changes to either side of the exon is captured. Short-read sequencing data is ideal for this data, as each junction is found on a single aligned read.

Multiple tools for analyzing splicing events exist - SUPPA2 (Trincado et al., 2018), LeafCutter (Y. I. Li et al., 2018), JuncBASE (Brooks et al., 2011), MAJIQ (Vaquero-Garcia et al., 2016), SplAdder (Kahles, Ong, Zhong, & Rätsch, 2016), and Whippet (Sterne-Weiler, Weatheritt, Best, Ha, & Blencowe, 2018). JuncBASE, LeafCutter, MAJIQ, and SplAdder are able to detect novel splice events, unlike SUPPA2 and Whippet.

3.3 MESA tool development

We developed a new splicing analysis tool named MESA (Mutually Exclusive Splicing Analysis) to detect alternative splicing events (Fig. 3.1). The tool performs junction-based quantification and counts splicing events based on donor and acceptor sites, where transcripts are connected when mapped to a reference genome. MESA specifically identifies mutually exclusive junctions - these are where intronic intervals cannot coexist on the same transcript due to overlap. These junctions must be found on

different transcripts or isoforms of a gene. For each junction, MESA creates a mutually exclusive cluster of the entire set of all junctions that overlap it. The inclusion count is calculated as being the number of reads for this junction. The exclusion count is then calculated by taking the number of other junctions from this mutually exclusive cluster appear in reads for this sample. The PS value is finally calculated by dividing the inclusion count by the sum of the inclusion and exclusion count. Each junction will have a PS value between 0 and 1, except for junctions where exclusion count is 0 and the PS value is reported as NaN. This straightforward calculation allows for a brisk analysis that analyzes all splicing events within a sample that can easily scale and is easy to interpret. MESA also does not require alignment files and can be run on junction count files alone. This significantly reduces the requirements necessary to obtain splicing information, particularly in situations where alignment or sequence data is inaccessible.

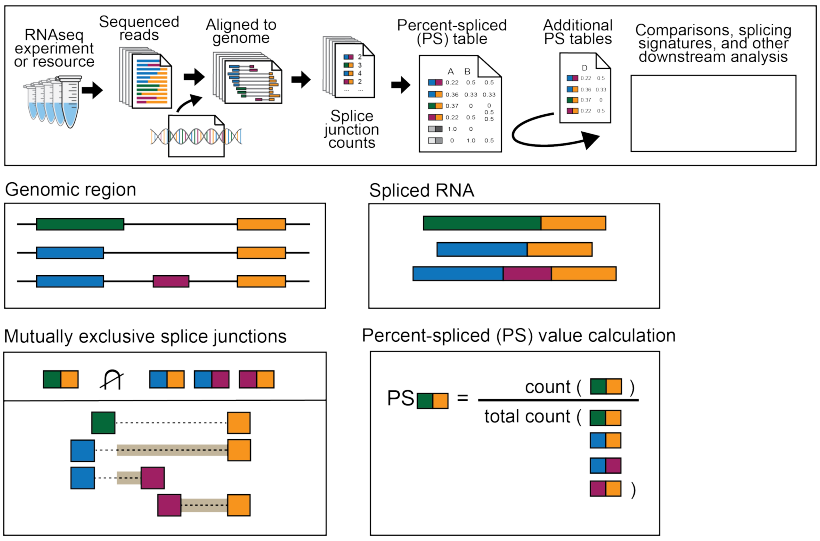


Figure 3.1: **MESA performs junction-based quantification of splicing events.** Read counts supporting each event in a mutually exclusive cluster of splice junctions are used to calculate the percent spliced value (Mulligan, 2022).

3.4 Clinical genetic splicing analysis

Germline RNA-seq data from 3 probands and their families were initially provided by UCSF. A second set of 2 probands and their families was subsequently provided. Each proband was previously diagnosed with a genetic disease of unknown origin. Variant and outlier expression analysis was completed. Some variants of unknown significance (VUS) were reported, but outlier expression analysis revealed nothing remarkable. After analysis with MESA, the splicing patterns of all three proband groups featured notable changes in junction usage. We first created a reference splicing event list, using all TCGA samples to generate distributions for each splicing event in those

samples (Fig. 3.2). These served as a reference point to compare the PS values from the clinical samples we received. In order to determine the PS values for relevant splice junctions were generated from 670 GTEx whole blood samples using MESA. For each clinical sample, the PS of each candidate splicing event was first compared to the distribution of PS values generated from the GTEx whole blood samples. Events that were initially deemed as outlier events would undergo pairwise comparisons using Fisher's exact test between the PS values of that event for each clinical sample. An event with a PS higher or lower than all other samples as well as a Fisher's P-value less than 0.05 between the proband and its parents, other probands, and all other parents were selected. Finally, gene lists previously provided by UCSF were used to curate events. Genes with variants being potentially associated with patient symptoms were generated by Phenomizer (Köhler et al., 2009, 2014) and further used to curate events.

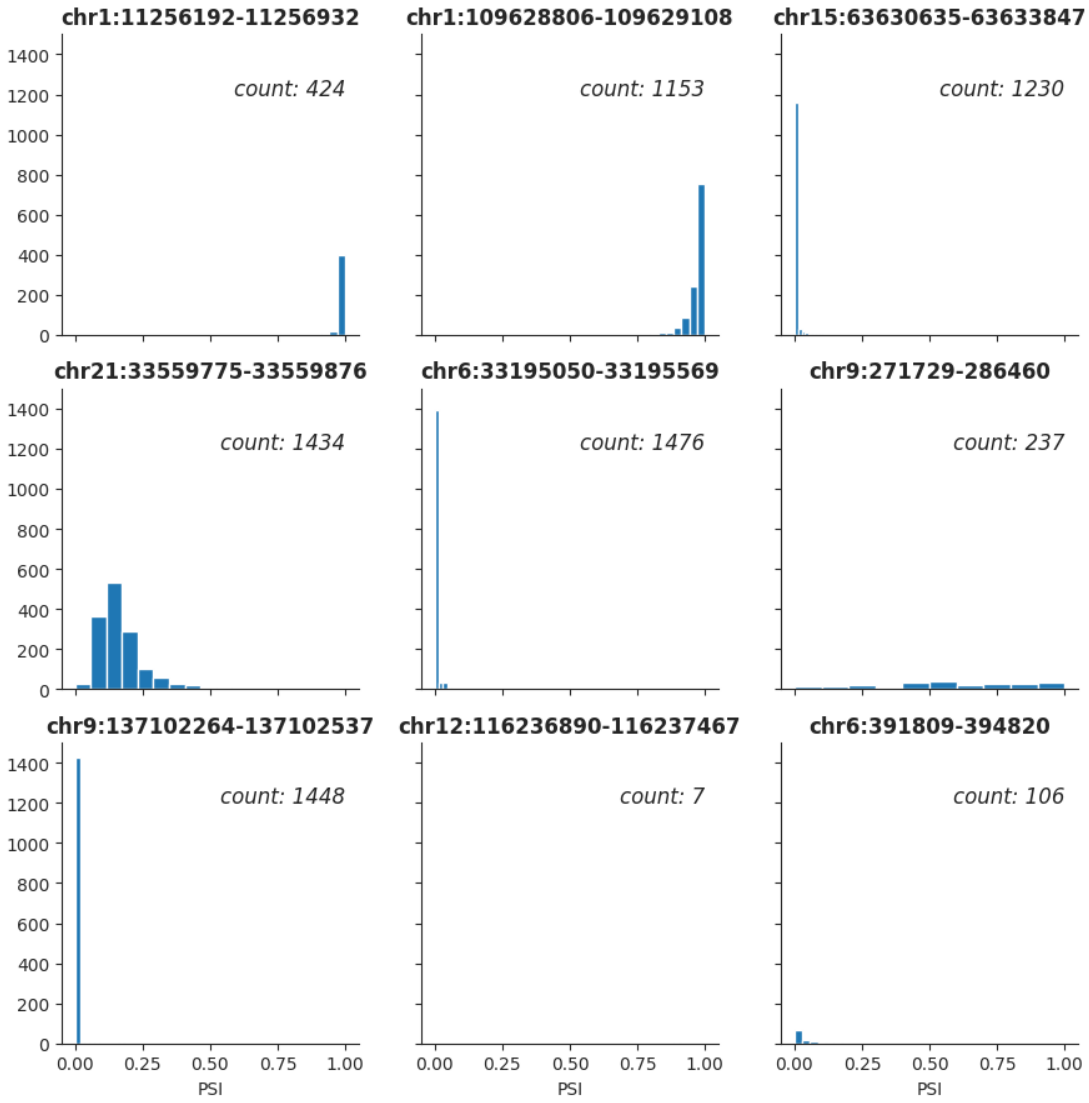


Figure 3.2: **PS distributions from TCGA.** Using MESA, we generated PSI values for all splice junctions in whole blood samples. We used 11,123 splice junction files and created distributions of the PSI values as a baseline comparison for potential splicing events. These 9 splicing events are a snapshot of distribution patterns.

3.5 *STT3B* and *SMAD4* alternative splicing events

One outlier splicing event was found in *STT3B* of proband GML1000 - the sample had an alternative 5' splice site usage at the end of exon 15 (Fig 3.3). Inclusion of this intron features a premature stop codon, which may cause the protein product to be truncated and have altered functionality. STT3B mediates post-translational glycosylation and mutations in this gene are connected with microcephaly and developmental delays (Shrimal, Ng, Losfeld, Gilmore, & Freeze, 2013). Another outlier splicing event in this proband was found in *SMAD4*, where intron inclusion between the 4th and 5th exon was found (Fig 3.4). A potential SMAD4 homolog (AC091551.1) with these exons has been previously identified, which could explain the changes in junction usage for these 3 samples. SMAD4 is a transcription factor and part of the transforming growth factor beta (TGF-β) pathway. It has been identified as a tumor suppressor gene (Liu, Pouponnot, & Massagué, 1997). However, it has not been strongly linked to developmental disorders, so these alternative splicing events in *SMAD4* are not likely to be associated with the diseases.

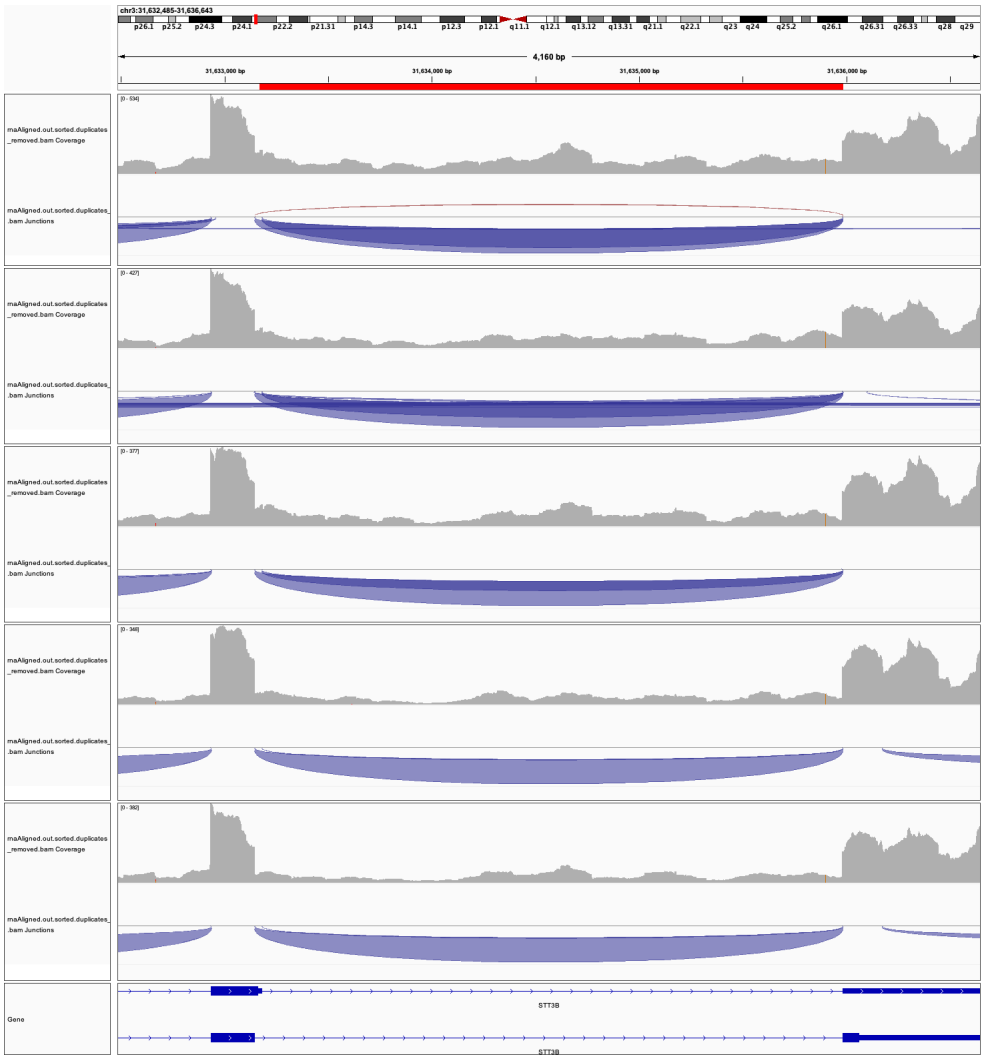


Figure 3.3: **MESA detected an alternative 5' splice site in STT3B.** *STT3B*, which encodes for a oligosaccharyltransferase subunit, was found to have alternative 5' splice site usage in the proband. mutations are the deletion of a stop codon to extend an existing uORF and the creation of a new stop codon and subsequently uORF.

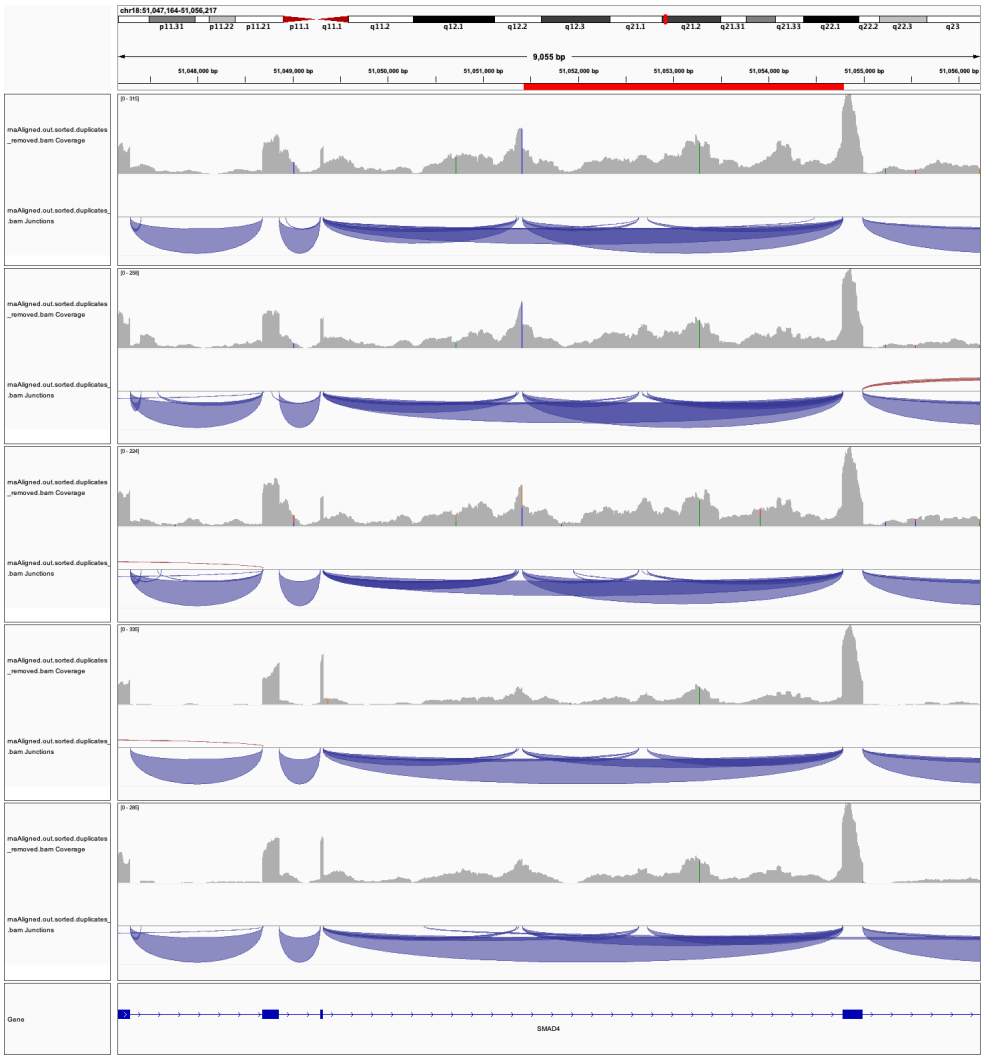


Figure 3.4: **MESA detected an intron inclusion event in *SMAD4*.** SMAD4 is a part of the TGF-β signaling pathway and was found to have an intron inclusion event between the 4th and 5th exon.

3.6 *DEGS1* alternative splicing events

Another proband, TH43_2695_S01, was also analyzed for splicing variants. Earlier variant analysis revealed a homozygous variant of uncertain significance (VUS)

in *DEGS1* - c.825+5delAGinsTT. DEGS1 is a sphingolipid desaturase that has been previously linked to hypomyelination and degeneration of the central and peripheral nervous systems (Pant et al., 2019). Using MESA, we independently detected several splicing variants in this gene and junctions from this gene were ranked in the top 100 of most differentially spliced junctions. Outlier analysis confirmed that the PS of these events were significantly lower than the rest of the samples and the GTEx whole blood samples (Fig. 3.5). We found a potential alternative 5' splice site at the end of exon 2 as well as complete skipping of this exon (Fig. 3.6). The VUS detected in *DEGS1* is located 4 base pairs away from the canonical splice site at the end of exon 2 and would alter the splice site context, potentially lowering the strength of the canonical splice site. With this splice site weakened, the use of a cryptic splice site downstream could explain the changes in splicing in this patient (Fig. 3.7).

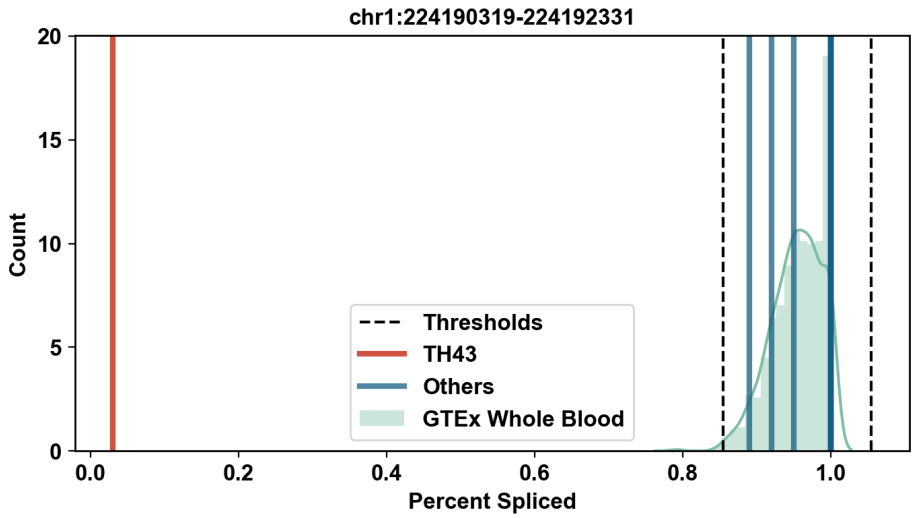


Figure 3.5: **Outlier analysis reveals *DEGS1* is differentially spliced**. Splicing outlier analysis found multiple splice variants in the proband. PS values for this *DEGS1* splice event were calculated from GTEx whole blood samples. The detected outlier splicing event is an alternative 5' splice site. The other samples had similar PS values for this event, while the proband had a significantly lower PS value.

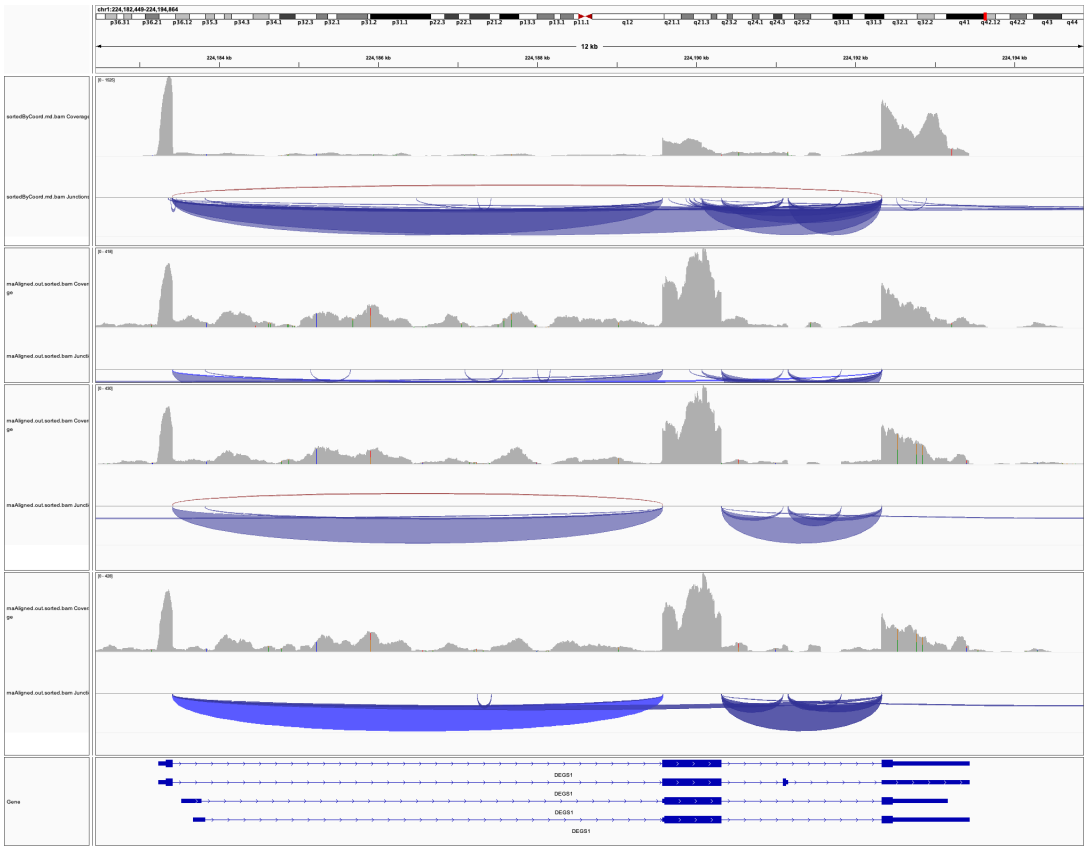


Figure 3.6: **MESA detected multiple alternative splicing events in *DEGS1*.** Both alternative 5' splice site usage and exon skipping are shown in this IGV snapshot.

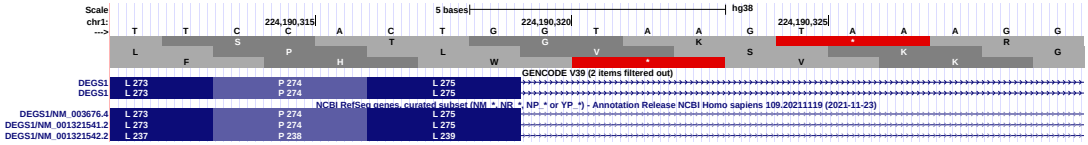


Figure 3.7: ***DEGS1* VUS affects splice site context.** This variant of uncertain significance alters context downstream of the 5' splice site and alters the strength and usage of this site, which could explain the alternative splicing events found in this gene.

3.7 Conclusion

MESA efficiently detects splicing variants and successfully performs comprehensive alternative splicing quantification. Here we demonstrated a potential clinical application of MESA and its ability to independently identify splice variants with potential links to an undiagnosed genetic disease. In particular, a variant of unknown significance was reported in *DEGS1*, which alters the splice site context of the 5' splice site of the splice junction event detected by MESA. We also outlined how a splicing event outlier analysis can be used to quickly identify splicing variants of interest with PS values from just one sample or a group of samples. As MESA quantifies splicing genome-wide, it can be used to generate splicing profiles that can further reveal splicing dysregulation can lead to disease.

3.8 Methods

Aligned reads from 670 whole blood samples were downloaded from the GTEx portal. Reads were realigned using STAR 2.4.2a to the hg38 reference genome. Realigned reads and associated junction files were analyzed with MESA. Realigned reads were also analyzed with Leafcutter for comparison between PSI and PS values. Junction locations, PS values from MESA, and PSI values from Leafcutter were compared and overlapping junctions were kept for outlier analysis. Custom python scripts to

perform pairwise comparisons with a Fisher's exact test were written to compare PS values between samples and identify the most differentially spliced events. Using the PS values generated by MESA, a distribution of PS values from the GTEx whole blood samples was used to calculate the quartile ranges. Outlier splicing events were defined as events that were larger than the third quartile (Q3) + 1.5 * the interquartile range (IQR) or smaller than the first quartile (Q1) - 1.5 * IQR. PS values from probands and family members were then compared to these cutoffs. Once outlier splicing events were identified, splicing variants were visually inspected with IGV.

Chapter 4

Modeling exon skipping events in lung cancer cell lines

4.1 Background

Alternative splicing is an efficient way to expand both transcriptome and proteome diversity by creating multiple mRNAs and proteins from a single gene through the inclusion and exclusion of particular exons. Exon skipping is the most common alternative splicing event in mammals and can vary according to tissue type (Florea, Song, & Salzberg, 2013) and developmental stages (Planells, Gómez-Redondo, Pericuesta, Lonergan, & Gutiérrez-Adán, 2019). Exon skipping can occur when mutations disrupt any of the core sequences specific to splicing: the 5' or 3' splice site, the branchpoint

site, the polypyrimidine tract, or the splicing enhancers or silencers. With the absence of a complete exon, these gene products can lack functional domains or sites and often have altered biological functions, which have been implicated in genetic diseases. The FAS receptor (TNR6) is a cell surface receptor involved with apoptosis with two isoforms whose functions are altered by exon skipping. Inclusion of exon 6 results in the canonical membrane-bound form of Fas that promotes apoptosis when bound to TNFS6. Skipping of exon 6 results in a splice variant that lacks the transmembrane domain and results in a soluble product that inhibits apoptosis by competing with TNFS6 (Izquierdo et al., 2005). Another anti-apoptosis gene, *survivin*, has multiple isoforms, one of which is missing exon 3 (Mahotka, Wenzel, Springer, Gabbert, & others, 1999). The skipped and full-length isoforms are apoptosis inhibitors, but the full-length isoform functions in the cytoplasm, while the skipped isoform functions in the nucleus (Mahotka et al., 2002).

4.2 The genomic landscape of lung cancers

Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer deaths in the world (Sung et al., 2021). While smoking and tobacco usage has fallen in the United States (Sung et al., 2021), smoking rates continue to increase in developing nations (Dela Cruz, Tanoue, & Matthay, 2011), leading to rising lung cancer incidence levels. Men are twice as likely to be diagnosed with lung cancer

than women and the mean age of diagnosis is 70 years old. Lung cancers are classified based on cell of origin, with the two major types being small-cell lung (SCLC) and non-small-cell lung cancers (NSCLC) (Nicholson et al., 2022). NSCLC accounts for 85% of lung cancer cases, with the two most common subtypes being lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) (Nicholson et al., 2022). 80% of lung cancer cases in the Western world are found in patients with a history of smoking (Alberg & Samet, 2003) and all major forms of lung cancers are connected to smoking. NSCLC and SCLC are more associated with smoking patients and men, while LUAD is more likely to be found in women and never smokers. For LUAD and LUSC, the tumor suppressor gene *TP53* is the most commonly mutated gene. In LUAD, the oncogenes *EGFR* and *KRAS* and the tumor suppressor genes *KEAP1*, *STK11*, and *NF1* are commonly mutated. In LUSC, the tumor suppressor gene *CDKN2A* is commonly mutated. Less than a decade ago, half of lung cancers lacked personalized therapies or actionable targets (Herbst, Morgensztern, & Boshoff, 2018). Fortunately, mortality from NSCLC has fallen as new targeted therapies have been approved. *EGFR*, *ALK*, *ROS1*, *RET*, *BRAF*, *MET*, *NTRK1*, and *HER2* are all drivers with inhibitors that have either been approved or are undergoing clinical review. Immunotherapies centered around the programmed cell death protein 1 (PD-1) or programmed cell death ligand 1 (PD-L1) checkpoint pathways have also been successful in improving outcomes (Seo, Kim, Shin, & Kim, 2018) and have become first-line options for NSCLC patients who

lack driver gene mutations (Hanna et al., 2020; Mezquita et al., 2018). This molecular classification is also necessary to understand why lung cancer rates in never-smokers continue to rise. Lung cancers found in never-smokers have a starkly different genetic profile, as they tend to have a 10-fold drop in mutation frequency. *EGFR*, *ALK*, *ROS1*, and *RET* mutations are more likely to be found in never-smokers, while *KRAS*, *TP53*, *NRAS*, and *MAP2K1* mutations are more common in smokers. Genetic profiling has led to the detection of these actionable drivers and increased our understanding of the molecular differences between lung cancer subtypes. As more promising therapies are approved for treatment of lung cancers, classification by molecular features will be absolutely crucial in selecting the most effective treatment course.

4.3 Exon skipping events in cancer and other genetic diseases

Alternative splicing contributes not only to protein diversity, but is also closely linked with genetic diseases. Up to 20% of genetic diseases arise from mutations that alter pre-mRNA splicing. Changes in spliceosome machinery can also lead to disease. The splicing factors splicing factor 3B, subunit 1 (SF3B1), serine/arginine-rich 2 (SRSF2), and U2 small nuclear RNA auxiliary factor 1 (U2AF1) are recurrently mutated in cancers (Dvinge, Kim, Abdel-Wahab, & Bradley, 2016); these mutations

have been discovered in multiple tumor types, including myelodysplastic syndromes (MDS) (Zhang et al., 2015), chronic lymphocytic leukemia (CLL) (L. Wang et al., 2016), lung adenocarcinoma (LUAD) (Imielinski et al., 2012), and breast invasive carcinoma (BRCA) (Maguire et al., 2015). Cancer cells can express altered isoforms of proteins, which can subsequently lead to tumor formation, progression, and drug resistance. With respect to exon skipping, the most prominent example in cancer is *MET* exon 14 (Kong-Beltran et al., 2006). This mutation occurs in approximately 3% of lung adenocarcinomas and 2.3% in other lung cancer subtypes (Cancer Genome Atlas Research Network, 2014) and does not co-occur with other known driver mutations in *KRAS*, *HER2*, and *EGFR* (Frampton et al., 2015). *MET* encodes for a tyrosine kinase receptor that activates cell proliferation, survival, and growth signaling pathways. Exon 14 encodes for the 47-amino acid juxtamembrane domain of MET; this region regulates MET signaling. When the exon is missing, ubiquitination and degradation of MET decrease and often leads to hyperactive MET-mediated signaling (Peschard et al., 2001). These exon skipping events can be caused by multiple mutations occurring in splice acceptor site, splice donor site, or intronic regions surrounding the exon. Recent clinical studies have demonstrated that capmatinib (Wolf et al., 2020) and tepotinib (Paik et al., 2020) successfully target *MET* exon 14 and have been shown to improve outcomes. Other studies have shown other exon skipping events as potentially targetable sites, mostly in lung cancers. *HER2* exon 16 skipping had first been identified as a driver

mutation in breast cancer (Turpin et al., 2016) and has also been detected in gastric and colorectal cancers. The skipping event has been shown to transform lung epithelial cells in vitro and in vivo (Smith et al., 2020) and may even mediate resistance to targeted therapies. Similarly, *EGFR* exon 19 skipping is another major mutation found in lung cancers. Multiple recurrent somatic mutations have been found in *EGFR* and two small molecule inhibitors - gefitinib and erlotinib - have shown promise as potential targeted therapies. Patients with *EGFR* exon 19 skipping have been shown to have better outcomes when treated with gefitinib or erlotinib compared to patients with the *EGFR* L858R hotspot mutation (Jackman et al., 2006).

4.4 Criteria for selecting exon skipping events

Previous studies have found that mutations in the splicing factors U2AF1 and RBM10 are associated with splicing changes in lung adenocarcinomas. We first used juncBASE to quantify the level of alternative splicing in 495 LUAD RNA-seq samples from the The Cancer Genome Atlas (TCGA). JuncBASE calculates a PSI value for each alternative splicing event and can detect novel splicing events, which will likely be missing from current gene annotations. These LUAD samples had matched whole exome mutation calls, which can be linked to these detected splicing events. We defined a splice site alteration as being any variant 3 base pairs into an annotated exon or 30 base pairs into its adjacent intron. Aberrant exon skipping was defined as any

cassette exon splicing event with a PSI 3 standard deviations below the mean. We found 635 candidate exon skipping events that were nearby a somatic splice site mutation. We also identified 11 samples with a *U2AF1* S34F mutation and 28 samples with *RBM10* LOF mutations. Differential splicing analysis revealed 94 splicing events associated with *U2AF1* S34F mutation and 15 splicing events associated with *RBM10* LOF mutations that had significantly increased exon skipping levels. We also found that 106 samples lacked a known oncogenic driver mutation. From these samples, we identified 50 cassette splicing events that were differentially spliced. Additional splice site mutations were provided by Guardant Health. This data was generated from sequencing circulating tumor DNA in blood samples collected from patients with lung adenocarcinoma.

4.5 Guide RNA design

In order to introduce the candidate exon skipping events into a lung cell line, we designed a library of CRISPR/Cas9 guide RNAs (gRNAs) to efficiently target their associated candidate exons. The candidate exon skipping events were first divided into two pools based on their oncogenic potential. The first subpool featured events that did not result in a frameshift after the exon skipping and were not previously detected in normal samples. The second subpool included all skipping events. From these subpools, we used custom python scripts and the CRISPOR gRNA design tool (Haeussler

et al., 2016) to create gRNA sequences that would disrupt the target exon splice sites using the CRISPR/Cas9 system. CRISPOR outputs potential gRNA sequences, complete with PAM sites, and attempts to predict the efficiency and off-target effects of these sequences. We designed a total of 5,461 gRNAs targeting the splice sites and the center of all 794 candidate exons. 6 unique gRNAs were chosen for each splice site of every candidate exon and gRNAs with a cut site closest to the splice site, the highest specificity score, and the highest cutting efficiency score were prioritized (Fig 4.1). gRNAs targeting the middle of each associated exon and the first exon of each candidate gene were included as negative controls. We expect gRNAs that target the center of an exon are less likely to alter the splicing of an exon. gRNAs that target the first exon are likely to result in a loss of function. We also included gRNAs that have no known target in the human genome as another negative control. As described early, *MET* exon 14 skipping was detected as one of the candidate events and confirms that the pipeline is properly identifying potential exon skipping events.

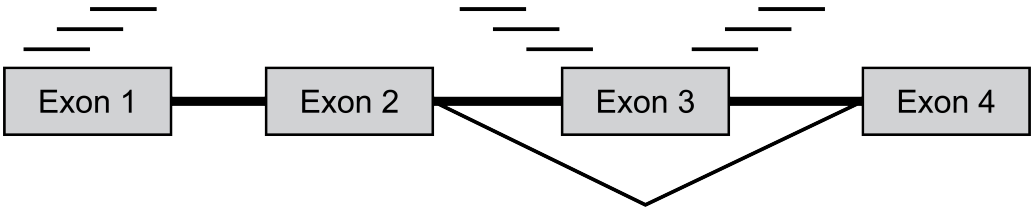


Figure 4.1: **gRNA targeting strategy for ssCRISPR.** 6 gRNAs target the exon of interest. 3 gRNAs also target the first exon as a negative control.

4.6 Conclusion

Using the CRISPR/Cas9 gene editing system to introduce exon skipping events, our ssCRISPR screen is designed to measure the oncogenic potential of these events in lung cancer cell lines. Candidate events from this assay can reveal new mechanisms for tumor formation and could function as new targets for therapeutics. Early functional validation of this assay has already demonstrated gRNAs targeting *HER2* exon 16 results in cells with increased survival in low attachment growth environments, consistent with other transformed cells. While our focus is on discovering novel sites for targeted therapies, this computational pipeline for creating gRNAs can easily be adapted to detect exon skipping events in other tumor types.

4.7 Methods

495 RNA-seq samples were downloaded from the TCGA portal. All samples were run through JuncBASE (v1.2) to generate PSI values for splicing events. To find events with a significant difference in PSI values, a Wilcoxon rank sum test with Benjamini-Hochberg multiple testing correction was performed. Matched somatic mutation calls and genomic metadata for these samples were also downloaded from the TCGA portal. Differentially spliced events were then pooled into candidate groups for gRNA design. CRISPOR (v3.1) was used to generate the sequences for the gRNAs. These sequences

were then altered using custom python scripts in preparation for library design.

Chapter 5

Conclusion

With the introduction of RNA sequencing over a decade ago, the method is now the de facto tool for studying the transcriptome and its links to the genome. New techniques and analysis methods continue to be built on the core technology and have expanded its scope to encompass single-cell gene expression, isoform analysis, translation efficiency, and spatial transcriptomics. Extending its reach to the detection of variants - somatic and splicing - demonstrates another potential use for RNA-seq. Accurate variant detection continues to be vital in our quest to determine the causes of genetic diseases. Methods founded on next generation sequencing technology have already uncovered novel targets that now have successful therapies, but there are an incredible number of diseases that still remain genetically unexplained. In this dissertation, I demonstrated three methods designed to shed light on the molecular mechanisms behind genetic dis-

eases. RNA-VACAY can harness existing RNA-seq data to find somatic mutations that are actively expressed. MESA can generate comprehensive splicing signatures that can be used to identify alternative splicing and splicing variants. ssCRISPR can help researchers design a CRISPR/Cas9 assay to study the effects of exon skipping in a cancer cell line. These methods provide novel biomarkers for researchers to interrogate during drug development. Alternatively, the field of precision medicine has rapidly grown and robust genetic profiling is also playing a pivotal role for clinicians who now rely on the status of key genes in a tumor when making therapeutic decisions. The cost of sequencing platforms has declined significantly in the past decade and the data being generated by NGS is absolutely critical to augmenting existing frameworks for disease management. These tools can provide comprehensive profiles of somatic mutation frequencies and splicing changes, giving clinicians another approach to utilize molecular pathology to better diagnose diseases, predict outcomes, and deliver precise treatment options.

References

Alberg, A. J., & Samet, J. M. (2003, January). Epidemiology of lung cancer. *Chest*, *123*(1 Suppl), 21S–49S.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D. A., Ghosh, S., & Jacobson, A. (2004, November). A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature*, *432*(7013), 112–118.

Ardlie, K. G., DeLuca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., ... Lockhart (2015, May). The Genotype-Tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., ... Predeus, A. V. (2020, February). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci. Rep.*, *10*(1), 2057.

Berget, S. M., Moore, C., & Sharp, P. A. (1977, August). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.*, *74*(8), 3171–3175.

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., ... Graveley, B. R. (2011, February). Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.*, *21*(2), 193–202.

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

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., ... Schultz, N. (2012, May). The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, *2*(5), 401–404.

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., ... Taylor, B. S. (2016, February). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.*, *34*(2), 155–163.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., ... BRIM-3 Study Group (2011, June). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.*, *364*(26), 2507–2516.

Christoforides, A., Carpten, J. D., Weiss, G. J., Demeure, M. J., Von Hoff, D. D., & Craig, D. W. (2013, May). Identification of somatic mutations in cancer through bayesian-based analysis of sequenced genome pairs. *BMC Genomics*, *14*, 302.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012, April). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009, December). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212.

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011, December). Lung cancer: epidemiology, etiology, and prevention. *Clin. Chest Med.*, *32*(4), 605–644.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, *43*(5), 491–501.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., ... Gingeras, T. R. (2012, September). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108.

DNA Technologies Core. (n.d.). *UC rate — UC davis and other UC campuses.* Retrieved from `https://dnatech.genomecenter.ucdavis.edu/uc-prices/`

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016, July). RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer*, *16*(7), 413–430.

Florea, L., Song, L., & Salzberg, S. L. (2013, September). Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res.*, *2*, 188.

Frampton, G. M., Ali, S. M., Rosenzweig, M., Chmielecki, J., Lu, X., Bauer, T. M., ... Miller, V. A. (2015, August). Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. *Cancer Discov.*, *5*(8), 850–859.

Fukuoka, M., Yano, S., Giaccone, G., Tamura, T., Nakagawa, K., Douillard, J.-Y., ... Baselga, J. (2003, June). Multi-Institutional randomized phase II trial of gefitinib for previously treated patients with advanced Non–Small-Cell lung cancer. *J. Clin. Orthod.*, *21*(12), 2237–2246.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., ... Schultz, N. (2013, April). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, *6*(269), l1.

García-Nieto, P. E., Morrison, A. J., & Fraser, H. B. (2019, December). The somatic mutation landscape of the human body. *Genome Biol.*, *20*(1), 298.

Garrison, E., & Marth, G. (2012, July). Haplotype-based variant detection from short-read sequencing.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., ... Concordet, J.-P. (2016, July). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, *17*(1), 148.

Hanna, N. H., Schneider, B. J., Temin, S., Baker, S., Jr, Brahmer, J., Ellis, P. M., ... Masters, G. (2020, May). Therapy for stage IV non-small-cell lung cancer without driver alterations: ASCO and OH (CCO) joint guideline update. *J. Clin. Oncol.*, *38*(14), 1608–1632.

Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature*, *553*(7689), 446–454.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020, February). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93.

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., ... Meyerson, M. (2012, September). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, *150*(6), 1107–1120.

Izquierdo, J. M., Majós, N., Bonnal, S., Martínez, C., Castelo, R., Guigó, R., ... Valcárcel, J. (2005, August). Regulation of fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell*, *19*(4), 475–484.

Jackman, D. M., Yeap, B. Y., Sequist, L. V., Lindeman, N., Holmes, A. J., Joshi, V. A., ... Jänne, P. A. (2006, July). Exon 19 deletion mutations of epidermal growth factor receptor are associated with prolonged survival in Non–Small cell lung cancer patients treated with gefitinib or erlotinib. *Clin. Cancer Res.*, *12*(13), 3908–3914.

Kahles, A., Ong, C. S., Zhong, Y., & Rätsch, G. (2016, June). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics*, *32*(12), 1840–1847.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2020, May). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., ... Greenberg, M. E. (2010, May). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., ... Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, *22*(3), 568–576.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., ... Robinson, P. N. (2014, January). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, *42*(Database issue), D966–74.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., ... Robinson, P. N. (2009, October). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, *85*(4), 457–464.

Kong-Beltran, M., Seshagiri, S., Zha, J., Zhu, W., Bhawe, K., Mendoza, N., ... Yauch,

R. (2006, January). Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res.*, *66*(1), 283–289.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., Mcewen, R., ... Dry, J. R. (2016). VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, *44*(11).

Li, H. (2011, November). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup (2009, August). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018, January). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, *50*(1), 151–158.

Lim, Y., Arora, S., Schuster, S. L., Corey, L., Fitzgibbon, M., Wladyka, C. L., ... Hsieh, A. C. (2021, July). Multiplexed functional genomic analysis of 5' untranslated region mutations across the spectrum of prostate cancer. *Nat. Commun.*, *12*(1), 4217.

Liu, F., Pouponnot, C., & Massagué, J. (1997, December). Dual role of the Smad4/DPC4 tumor suppressor in TGFβ-inducible transcriptional complexes. *Genes Dev.*, *11*(23), 3157–3167.

Long, G. V., Menzies, A. M., Nagrial, A. M., Haydu, L. E., Hamilton, A. L., Mann, G. J., ... Kefford, R. F. (2011, April). Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma. *J. Clin. Oncol.*, *29*(10), 1239–1246.

Maguire, S. L., Leonidou, A., Wai, P., Marchiò, C., Ng, C. K., Sapino, A., ... Natrajan, R. C. (2015, March). SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J. Pathol.*, *235*(4), 571–580.

Mahotka, Wenzel, Springer, Gabbert, & others. (1999, December). Survivin-ΔEx3 and survivin-2b: two novel splice variants of the apoptosis inhibitor survivin with different antiapoptotic properties. *Cancer Res.*.

Mahotka, C., Liebmann, J., Wenzel, M., Suschek, C. V., Schmitt, M., Gabbert, H. E., & Gerharz, C. D. (2002, December). Differential subcellular localization of functionally divergent survivin splice variants. *Cell Death Differ.*, *9*(12), 1334–1342.

Mandelker, D., Gabelli, S. B., Schmidt-Kittler, O., Zhu, J., Cheong, I., Huang, C.-H., ... Amzel, L. M. (2009, October). A frequent kinase domain mutation that changes the interaction between PI3Kalpha and the membrane. *Proc. Natl. Acad. Sci. U. S. A.*, *106*(40), 16996–17001.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008, September). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, *18*(9), 1509–1517.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010, September). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, *20*(9), 1297–1303.

Mezquita, L., Auclin, E., Ferrara, R., Charrier, M., Remon, J., Planchard, D., ... Besse, B. (2018, March). Association of the lung immune prognostic index with immune checkpoint inhibitor outcomes in patients with advanced Non–Small cell lung cancer. *JAMA Oncol*, *4*(3), 351–357.

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

Mulligan, D. (2022). *Mesa schematic.*

Nicholson, A. G., Tsao, M. S., Beasley, M. B., Borczuk, A. C., Brambilla, E., Cooper, W. A., ... Travis, W. D. (2022, March). The 2021 WHO classification of lung tumors: Impact of advances since 2015. *J. Thorac. Oncol.*, *17*(3), 362–387.

Ohshima, K., Hatakeyama, K., Nagashima, T., Watanabe, Y., Kanto, K., Doi, Y., ... Yamaguchi, K. (2017, April). Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors. *Sci. Rep.*, *7*(1), 641.

Oikkonen, L., & Lise, S. (2017, January). Making the most of RNA-seq: Pre-

processing sequencing data with opossum for reliable SNP variant detection. *Wellcome Open Res*, *2*, 6.

Paik, P. K., Felip, E., Veillon, R., Sakai, H., Cortot, A. B., Garassino, M. C., ... Le, X. (2020, September). Tepotinib in Non–Small-Cell lung cancer with MET exon 14 skipping mutations. *N. Engl. J. Med.*, *383*(10), 931–943.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing* (Vol. 40) (No. 12).

Pant, D. C., Dorboz, I., Schluter, A., Fourcade, S., Launay, N., Joya, J., ... Pujol, A. (2019, March). Loss of the sphingolipid desaturase DEGS1 causes hypomyelinating leukodystrophy. *J. Clin. Invest.*, *129*(3), 1240–1256.

PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., ... PCAWG Consortium (2020). Genomic basis for RNA alterations in cancer. *Nature*, *578*(7793), 129–136.

Peschard, P., Fournier, T. M., Lamorte, L., Naujokas, M. A., Band, H., Langdon, W. Y., & Park, M. (2001, November). Mutation of the c-cbl TKB domain binding site on the met receptor tyrosine kinase converts it into a transforming protein. *Mol. Cell*, *8*(5), 995–1004.

Picardi, E., D'Erchia, A. M., Lo Giudice, C., & Pesole, G. (2017, January). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.*, *45*(D1), D750–D757.

Piskol, R., Ramaswami, G., & Li, J. B. (2013, October). Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, *93*(4), 641–651.

Planells, B., Gómez-Redondo, I., Pericuesta, E., Lonergan, P., & Gutiérrez-Adán, A. (2019, March). Differential isoform expression and alternative splicing in sex determination in mice. *BMC Genomics*, *20*(1), 202.

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., ... Morris, D. W. (2013, March). Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*, *8*(3), e58815.

Radenbaugh, A. J., Ma, S., Ewing, A., Stuart, J. M., Collisson, E. A., Zhu, J., & Haussler, D. (2014, November). RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One*, *9*(11), e111516.

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

Riely, G. J., Kris, M. G., Rosenbaum, D., Marks, J., Li, A., Chitale, D. A., ... Ladanyi, M. (2008, September). Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin. Cancer Res.*, *14*(18), 5731–5734.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., ... Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, *46*(8), 912–918.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011, January). Integrative genomics viewer. *Nat. Biotechnol.*, *29*(1), 24–26.

Rusch, V., Baselga, J., Cordon-Cardo, C., Orazem, J., Zaman, M., Hoda, S., ... Dmitrovsky, E. (1993, May). Differential expression of the epidermal growth factor receptor and its ligands in primary non-small cell lung cancers and adjacent benign lung. *Cancer Res.*, *53*(10 Suppl), 2379–2385.

Seo, J.-S., Kim, A., Shin, J.-Y., & Kim, Y. T. (2018, October). Comprehensive analysis of the tumor immune micro-environment in non-small cell lung cancer for efficacy of checkpoint inhibitor. *Sci. Rep.*, *8*(1), 14576.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001, January). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, *29*(1), 308–311.

Shrimal, S., Ng, B. G., Losfeld, M.-E., Gilmore, R., & Freeze, H. H. (2013, November). Mutations in STT3A and STT3B cause two congenital disorders of glycosylation. *Hum. Mol. Genet.*, *22*(22), 4638–4645.

Smith, H. W., Yang, L., Ling, C., Walsh, A., Martinez, V. D., Boucher, J., ... Muller,

W. J. (2020, August). An ErbB2 splice variant lacking exon 16 drives lung carcinoma. *Proc. Natl. Acad. Sci. U. S. A.*, *117*(33), 20139–20148.

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H., & Blencowe, B. J. (2018, October). Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol. Cell*, *72*(1), 187–200.e6.

Sun, S., Zhang, Z., Sinha, R., Karni, R., & Krainer, A. R. (2010, March). SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat. Struct. Mol. Biol.*, *17*(3), 306–312.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021, May). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, *71*(3), 209–249.

Talbot, S. J., & Crawford, D. H. (2004). *Viruses and tumours – an update* (Vol. 40) (No. 13).

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... Forbes, S. A. (2019, January). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, *47*(D1), D941–D947.

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018, March). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, *19*(1), 40.

Turpin, J., Ling, C., Crosby, E. J., Hartman, Z. C., Simond, A. M., Chodosh, L. A., ... Muller, W. J. (2016, November). The ErbB2ΔEx16 splice variant is a major oncogenic driver in breast cancer that promotes a pro-metastatic tumor microenvironment. *Oncogene*, *35*(47), 6053–6064.

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

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008, November). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476.

Wang, L., Brooks, A. N., Fan, J., Wan, Y., Gambe, R., Li, S., ... Wu, C. J. (2016, November). Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell*, *30*(5), 750–763.

Watson, C. T., & Breden, F. (2012, July). The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.*, *13*(5), 363–373.

Wolf, J., Seto, T., Han, J.-Y., Reguart, N., Garon, E. B., Groen, H. J. M., ... Heist, R. S. (2020, September). Capmatinib in MET exon 14–mutated or MET-Amplified Non–Small-Cell lung cancer. *N. Engl. J. Med.*, *383*(10), 944–957.

Xu, F., Wang, W., Wang, P., Jun Li, M., Chung Sham, P., & Wang, J. (2012). A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat. Commun.*, *3*, 1258.

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kübler, K., Grimsby, J., ... Getz, G. (2019, June). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, *364*(6444).

Yung, C. K., O'Connor, B. D., Yakneen, S., Zhang, J., Ellrott, K., Kleinheinz, K., ... the PCAWG Network (2017, July). *Large-Scale uniform analysis of cancer whole genomes in multiple computing environments.*

Zhang, J., Lieu, Y. K., Ali, A. M., Penson, A., Reggio, K. S., Rabadan, R., ... Manley, J. L. (2015, August). Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proc. Natl. Acad. Sci. U. S. A.*, *112*(34), E4726–34.