Winter 12-15-2022

Bioinformatics and cancer genomics approaches to advance precision medicine and elucidate tumor mutational landscapes

Kelsy Cotto
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Genetics Commons

Recommended Citation
Cotto, Kelsy, "Bioinformatics and cancer genomics approaches to advance precision medicine and elucidate tumor mutational landscapes" (2022). Arts & Sciences Electronic Theses and Dissertations. 2735.
https://openscholarship.wustl.edu/art_sci_etds/2735

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Bioinformatics and Cancer Genomics Approaches to Advance Precision Medicine and Elucidate Tumor Mutational Landscapes

by

Kelsy C. Cotto

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2022
St. Louis, Missouri
# Table of Contents

**List of Figures**

vi

**Acknowledgments**

vii

**Abstract of the Dissertation**

xi

**Chapter 1: Introduction**

1.1 Introduction to precision medicine and next-generation sequencing 1
   
   1.1.1 Library preparation for NGS data 3
   
   1.1.2 Next-generation sequencing approaches 4
   
   1.1.3 Alignment and variant calling of NGS data 6
   
   1.2 The “Druggable Genome” as a concept 8
   
   1.3 The use of next-generation sequencing in cancer research 10
   
   1.4 The current mutational landscape of breast cancer 11
   
   1.5 Splice-associated variants are under-identified in current analysis workflows 13
   
   1.6 Summary 14

**Chapter 2: DGIdb 3.0: a redesign and expansion of the drug-gene interaction database**

2.1 Preamble 16

2.2 Summary 16

2.3 Introduction 17

2.4 Results
   
   2.4.1 Expanded Content 18
   
   2.4.2 New Features and Enhancements 21
   
   2.4.3 User Interface (UI) Updates 24
   
   2.4.4 Usage and Accessibility 26
   
   2.5 Conclusion and Future Directions 27
   
   2.6 Availability 28
   
   2.7 Supplementary Data 28
   
   2.8 Funding 29

**Chapter 3: Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts**

3.1 Preamble 30

3.2 Summary 30

3.3 Introduction 31

3.4 Results 32
3.4.1 Integration with crowdsourced efforts 32
3.4.2 New and updated sources 35
3.4.3 Drug grouping improvements 37
3.4.4 New Query Score and updated Interaction Score 38
3.4.5 Inclusion of interaction directionality 41
3.4.6 Search feature improvements 42
3.4.7 Monthly data releases 43
3.4.8 Improved transparency and details on licensing of sources 44
3.4.9 Application framework updates 44
3.5 Conclusions and Future Directions 45
3.6 Data Availability 46
3.7 Supplementary Data 46
3.8 Acknowledgments 47
3.9 Funding 47

Chapter 4. ORegAnno 3.0: a community-driven resource for curated regulatory annotation 48
4.1 Preamble 48
4.2 Summary 48
4.3 Introduction 49
4.4 Results 50
  4.4.1 Overview 50
  4.4.2 Updates 54
  4.4.3 New Records 55
  4.4.4 Data access 57
  4.4.5 Applications 57
4.5 Acknowledgements 61
4.6 Supplementary Data 61
4.7 Funding 61

Chapter 5. Identification of recurrent regulatory mutations in breast cancer 62
5.1 Preamble 62
5.2 Summary 62
5.3 Introduction 63
5.4 Results 66
  5.4.1 Molecular subtyping of samples using genomic data 66
  5.4.2 Validation of the custom-capture reagent 67
  5.4.3 Discovery of recurrently mutated promoter regions 69
5.5 Discussion 72
5.6 Methods 72
5.6.1 Sample Procurement
5.6.2 Sequencing
5.6.3 Sequence alignment and somatic event detection
5.6.4 PAM50 molecular subtype classification
5.6.5 Advanced filtering of somatic variants
5.6.5 Subsetting variants to promoter regions
5.6.6 Identification of recurrently mutated promoter regions

Chapter 6. RegTools: Integrated analysis of genomic and transcriptomic data for the
discovery of splice-associated variants in cancer

6.1 Preamble
6.2 Summary
6.3 Introduction
6.4 Results

6.4.1 The RegTools tool suite supports splice-associated variant discovery by the integration of genome and transcriptome data
6.4.2 Pan-cancer analysis of 35 tumor types identifies somatic variants that alter canonical splicing
6.4.3 RegTools identifies splice-associated variants missed by other splice-associated variant predictors and annotators
6.4.4 Orthogonal validation of RegTools using clinical data and verified splice-associated mutations
6.4.5 Pan-cancer analysis reveals novel splicing patterns within known cancer genes and potential cancer drivers

6.5 Discussion
6.6 Methods

6.6.1 Software implementation
6.6.2 Benchmarking
6.6.3 Using RegTools to identify cis-acting, splice-associated variants
6.6.4 Dataset Description
6.6.5 Sample processing for cohorts with bulk transcriptome data
6.6.6 Mice used for MCB6C experiments
6.6.7 Mouse bladder organoid culture for mouse injection
6.6.8 Mouse injection with MCB6C organoid cells
6.6.9 Harvesting MCB6C tumors for single cell RNA-seq analysis
6.6.10 Single-cell RNA-seq analysis of MCB6C cells
6.6.11 Long read sequencing of HCC1395 cell line
6.6.12 Candidate junction filtering
6.6.13 Statistical filtering of candidate events
6.6.14 Annotation with GTEx junction data and other splice prediction tools
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>DGIdb 3.0 content by source</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison of the DGIdb 2.0 and 3.0 response time by search and result size</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>The DGIdb interaction search interface</td>
<td>25</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of main components of DGIdb</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>DGIdb 4.0 content by source</td>
<td>36</td>
</tr>
<tr>
<td>3.3</td>
<td>Overview of DGIdb's new Query scores and Interaction scores</td>
<td>40</td>
</tr>
<tr>
<td>3.4</td>
<td>Current content of the ORegAnno database</td>
<td>51</td>
</tr>
<tr>
<td>3.5</td>
<td>Comparison of the genomic coverage captured by ORegAnno and the ENSEMBL Regulatory Track</td>
<td>53</td>
</tr>
<tr>
<td>3.6</td>
<td>Capture reagent using ORegAnno sites improves coverage of regulatory regions in human hepatocellular carcinoma</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Breakdown of the number of samples classified as each molecular subtype for all 458 samples, irrespective of original cohort</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Breakdown of the number of samples classified as each molecular subtype split by original cohort</td>
<td>67</td>
</tr>
<tr>
<td>5.3</td>
<td>Coverage data for each of the 154 TCGA samples with WRS data</td>
<td>68</td>
</tr>
<tr>
<td>5.4</td>
<td>Coverage data for each of the 116 TCGA samples with WGS data</td>
<td>69</td>
</tr>
<tr>
<td>5.5</td>
<td>Lollipop of recurrent mutations in the bidirectional promoter of RMRP and CCDC107</td>
<td>71</td>
</tr>
<tr>
<td>6.1</td>
<td>Flexible, streamlined discovery of cis-acting splice-associated variants with RegTools modules and cis-splice-effects identify workflow</td>
<td>83</td>
</tr>
<tr>
<td>6.2</td>
<td>Splice-associated variants often lead to the expression of multiple alternative junctions</td>
<td>88</td>
</tr>
<tr>
<td>6.3</td>
<td>Comparison of RegTools with other tools that identify potential splice-associated variants</td>
<td>90</td>
</tr>
<tr>
<td>6.4</td>
<td>Intronic SNV in Trp53 associated with an exon skipping event</td>
<td>97</td>
</tr>
<tr>
<td>6.5</td>
<td>Pan-cancer analysis of cohorts from TCGA and MGI reveals genes recurrently disrupted by variants that cause non-canonical splicing patterns</td>
<td>99</td>
</tr>
<tr>
<td>6.6</td>
<td>Several SNVs in B2M are associated with alternate acceptor and alternate donor usage</td>
<td>102</td>
</tr>
<tr>
<td>7.1</td>
<td>Comparison of record counts for drug claims, gene claims, and interaction claims between DGIdb version 4.2.0 and 5.0</td>
<td>129</td>
</tr>
<tr>
<td>7.2</td>
<td>Comparison of normalization rates for drug claims, gene claims, and interaction claims between DGIdb version 4.2.0 and 5.0</td>
<td>130</td>
</tr>
</tbody>
</table>
Acknowledgments

There truly are not enough words to properly thank everyone who has supported me throughout this journey of obtaining my PhD. However, I will do my best to succinctly thank those who have been instrumental in my success over these last several years.

To Malachi and Obi, thank you both for welcoming me into your lab as an undergraduate student in 2015 and again as a graduate student in 2016. You allowed me, with no computational experience, to jump headfirst into a wide variety of projects and encouraged me to learn new techniques constantly. I think you both provided such a balanced mix of being compassionate and understanding while also challenging me to achieve more than I thought was possible. I have learned and grown so much, and you have both enabled every thought, learning exercise, and opportunity. I could not imagine better PIs to train under or a better environment to do my thesis work in. More than anything, I hope you know that your empathy, brilliance, and mentoring have only reinforced my passion for science and career aspirations. I really cannot thank you both enough.

To my committee, Barak, Allegra, Ron, and Gavin, I could not ask for a better committee. You all provided the perfect mix of scientific expertise to my dissertation projects and without a doubt made all my projects better through the insight that you provided. You all challenged to think outside of the lens that I was used to using when it came to projects and become a more well-rounded scientist. I cannot thank you enough for the support and guidance you have provided me with throughout my graduate career.

To the members of the Griffith lab, I cannot express how much of my success is due to the support you all have provided and how much I have learned from you. From the first day I started, I felt like each of you were invested in my success and you all have encouraged me to
dream bigger. I cannot thank you enough for that. It is so hard to imagine finding an environment as special as this one and it makes leaving difficult. I want to extend special thanks to some of the OG lab members who made the Griffith lab an environment I wanted to join, Katie Campbell, Erica Barnell, Zach Skidmore, Rob Lesurf, Alex Wagner, and Ben Ainscough. I also want to thank the lab members who were along for the end of my PhD and helped me finish strong, Megan Richters, Huiming Xia, Chris Yoon, Sharon Freshour, Felicia Gomez, Yang-Yang Feng, Kartik Singhal, and My Hoang.

To my friends, thanks for being with me every step of this journey, through both the highs and lows that we all endured. I truly could not have done this without you, especially Tomas, Leo, Rachel, and Marlie. I appreciate every meal, drink, laugh, and adventure we have shared over these past few years.

To my family, Mom, Rae, Nana, and Grandpa, thanks for every ounce of support you have given me over these past six years and throughout my life. Thanks for believing in me even when my training has taken me far away from you all. I appreciate every phone call, text, and visit from each of you over the years. I love you and I could not have done this without you.

I would like to acknowledge my advisors in DBBS and MCB, including Stacy Kiel, Jason Weber, Heather True. In addition, I would like to thank Tim Ley for his mentorship through the Cancer Biology Pathway and the Siteman Cancer Center for fellowship support and additional training through the Cancer Biology Pathway. I would also like to thank Tim Schedl, for his mentorship through the Precision Medicine Pathway. Also, I would like to thank Robert Mercer and Josh Rubin for mentorship throughout my graduate career, even though they did have a formal mentoring role on my committee.
Most importantly, I would like to thank the patients and their families for their selfless contribution to the advancement of science. Thank you for allowing me into your lives so that we, together, may help others and improve patient care and outcomes.

Kelsy C. Cotto

Washington University in St. Louis

August 2022
Dedicated to Stephen Mullis and Joann Senecal.

Thank you for always believing in me.
ABSTRACT OF THE DISSERTATION

Bioinformatics and cancer genomics approaches to advance precision medicine and elucidate tumor mutational landscapes

by

Kelsy Clara Cotto

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Cell Biology

Washington University in St. Louis, 2022

Professor Malachi Griffith, Chair

Professor Obi Griffith, Co-Chair

Next-generation sequencing of DNA and RNA continues to be integrated into pre-clinical and clinical research. However, challenges still remain that impede the translation of findings into an improved understanding of human diseases or clinically actionable alterations. The projects described in this dissertation start with compiling efforts from experts who have identified druggable genes within the human genome, followed by in-depth analyses which characterize the pan-cancer landscape of splice-associated mutations and noncoding, regulatory mutations across multiple subtypes of breast cancer. In Chapters 2 and 3, substantial efforts were made to update the content and the user experience for the drug-gene interaction database (DGIdb, dgidb.org). Chapter 2 describes the substantially expanded comprehensive catalog of druggable genes and anti-neoplastic drug-gene interactions included in DGIdb. Along with these content updates, there were major overhauls of the DGIdb codebase, including an updated user interface, preset interaction search filters, consolidation of interaction information into interaction groups, greatly improved search response times, and upgrades to the underlying web application framework. In addition, we expanded the API to add new endpoints, allowing users to extract more detailed
information about queried drugs, genes, and drug-gene interactions, including listings of PubMed IDs, interaction type, and other interaction metadata. The updates described in Chapter 3 focus on the integration of DGIdb with crowdsourced efforts, leveraging the Drug Target Commons for community-contributed interaction data, Wikidata to facilitate term normalization, and export to NDEx for drug-gene interaction network representations. Seven new sources were added since the previous major version release, and of the previously aggregated sources, 15 were updated. This update also included improvements to the process of drug normalization and grouping of imported sources. Other notable updates included the introduction of a more sophisticated Query Score for interaction search results, an updated Interaction Score, the inclusion of interaction directionality, and several additional improvements to search features, data releases, licensing documentation, and the application framework. In Chapters 4 and 5, we discuss how comprehensive sequencing approaches were used to discover noncoding, regulatory mutations within 458 breast cancer samples. After extensive filtering, our analysis revealed significant mutation clustering within the noncoding space of RMRP and WDR74, as has been noted in previous studies, as well as ~130 other genes not previously reported. Additionally, noncoding splice-associated mutations were discovered using RegTools. In Chapter 6, we assessed the landscape of splice-associated mutations within patient tumor cohorts from The Cancer Genome Atlas (TCGA) and Washington University clinical cohorts. We developed and employed RegTools to identify significant splice-associated mutations and discovered 235,778 events where a variant significantly increased the splicing of a particular junction across 158,200 unique variants and 131,212 unique junctions. To characterize these somatic variants and their associated splice isoforms, we annotated them with the Variant Effect Predictor (VEP), SpliceAI, and Genotype-Tissue Expression (GTEx) junction counts and compared our results to other tools.
that integrate genomic and transcriptomic data. We identified novel splice-associated variants and previously unreported patterns of splicing disruption in known cancer drivers, such as TP53, CDKN2A, and B2M, as well as in genes not previously considered cancer-relevant, such as RNF145. This dissertation describes studies that address challenges, including the accessibility of information to researchers, discovery of noncoding regulatory mutations, and identification of splice-associated mutations with an open-source tool, in order to advance the dissemination of knowledge within the bioinformatics and cancer genomics communities, elucidate novel mechanisms of tumor biology, and identify potential therapeutic targets for cancer therapy.
Chapter 1: Introduction

1.1 Introduction to precision medicine and next-generation sequencing

According to the President's Council of Advisors on Science and Technology in the United States, precision medicine is defined as “the tailoring of medical treatment to the individual characteristics of each patient and the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment” (President’s Council of Advisors on Science, 2008). One example of precision medicine is the identification of variants that are characteristic of a patient’s disease and utilizing tailored therapy protocols, specific to those variants, to improve outcomes. Over the last few decades, technological advancements and changes in the healthcare infrastructure have permitted the integration of precision medicine into clinical practice, which has improved the treatment for many patients, with cancer being a prominent example.

Some initial research that was a necessary precursor for the field of precision medicine was derived from The Human Genome Sequencing project, of which a draft of the human genome was completed in 2000 and required global collaboration from over 20 groups (Lander et al., 2001). Using a shotgun sequencing strategy, the project identified 30,000-40,000 protein-coding genes and 1.4M single nucleotide polymorphisms (SNPs) (Lander et al., 2001). This project was a revolutionary milestone; however, the work required to fully analyze the information contained within the sequence proved to be more extensive than the sequencing itself, resulting in minimal immediate clinical benefits (Collins and McKusick, 2001). Based on the core philosophy of shotgun sequencing, Lynx Therapeutics (which later merged with Solexa
and was eventually bought by Illumina) launched the first of the Next Generation Sequencing (NGS) technologies in 2000, called massively parallel signature sequencing (MPSS) (Reinartz et al., 2002; Zhang et al., 2011). In the next several years, other NGS technologies arose such as pyrosequencing (Roche(454)), polymerase-based sequencing by-synthesis (Solexa/Illumina), and ligation-based sequencing (SOLiD) (Mardis, 2008; Margulies et al., 2005). Illumina has since emerged as the market leader by virtue of having the lowest cost of sequencing per gigabase (Gb) of data as well as the accuracy and applicability of the data generated. At this time, Illumina holds about 80% of the market share for DNA sequencing. As the development of high-throughput approaches has continued to be optimized, the sequencing cost of the human genome has plummeted from around $3B for the first human genome to $1,000 as of 2015 (Phillips et al., 2015; Sboner et al., 2011). Emerging companies claim $100 genome prices at the time of writing this dissertation, but these claims have yet to be validated.

The reduction in the cost of sequencing has made it possible to develop sequencing-based approaches for the diagnosis and treatment of human diseases. Precision medicine tools developed over the last 15 years have demonstrated success with regard to treating patients based on genetic, environmental, and behavioral factors. Individual cases and small cohorts have provided substantial evidence that clinical practice and patient outcomes are improved when employing precision medicine to treat patients, such as in the case of cystic fibrosis, certain types of cancer, and the field of pharmacogenomics (Ashley, 2016). However, it is currently not feasible to scale this process of precision medicine for every individual. This is due in part to the rapid growth in the volume of clinically relevant information that can potentially impact patient care. For each individual, the relevant DNA or RNA must be sequenced (whether it be whole genome sequencing (WGS), whole exome sequencing (WES), RNA sequencing (RNA-Seq),
etc.), variants must be called, and all aberrations must be evaluated for actionability. This process requires extensive computing power for data processing, large amounts of disk space for storage of genomic data, a high manual burden for variant identification and review, and a cohort of experts for summarization and execution (Mardis, 2010).

While sequencing centers, universities, and companies usually have pipelines to automate alignment and variant calling, these pipelines largely focus on small nucleotide variants (SNVs) and small insertions and deletions (indels) within coding regions for which to assign actionability. Such approaches overlook alterations within intronic regions or non-coding areas of the genome. Additionally, synonymous mutations may be disregarded when in reality such variants could lead to changes in transcription or splicing. Apart from the narrow focus of current analysis pipelines, there is also the issue of how we annotate areas of the genome in regards to being disease-associated or potentially druggable. While individual labs and research groups may identify such associations, the information is located in disjoint publications and databases, making it difficult for researchers to make conclusions. This slows the process of achieving consensus for research that has been done or to translating findings from the research lab into the clinic. Bioinformatic and genomic efforts to summarize annotations across publications and databases and identify novel classes of mutations that are currently overlooked will ultimately help to advance the field of precision medicine.

1.1.1 Library preparation for NGS data

There are three main steps that are involved in generating NGS data. The first step, library construction, typically involves genomic fragmentation, ligation to adapters, and polymerase chain reaction (PCR) amplification, enzymatically polishing the fragment ends, and ligating known sequence adapters onto the fragment ends. Genomic fragmentation involves breaking the
DNA into smaller pieces and can be accomplished using physical or chemical means. Physical fragmentation methods include sonication, nebulization, or enzymatic reactions. Chemical fragmentation relies on hydroxyl radicals to break DNA into fragments. Relative to physical fragmentation, chemical fragmentation can accommodate more material, but can induce false positives through novel mutations or transversion artifacts. After fragmentation, adaptors are ligated to DNA fragments and may include sequences to allow binding to a flow cell, sequencing primer sites, sample indexes, unique molecular identifier (UMI) sequences, etc. These adaptors are ultimately sequenced and may require trimming prior to alignment, depending on the alignment strategy utilized. Fragment amplification via PCR is needed to provide sufficient signal during the sequencing reaction for on-instrument detection. Similarly, libraries for generating RNA-Seq data also can be prepared for NGS. First, RNA is converted to DNA using reverse transcriptase, followed by adapter ligation and surface amplification. The data resulting from RNA-Seq can be analyzed to determine digital expression values for genes in a given tissue or tumor.

1.1.2 Next-generation sequencing approaches

Sequencing is the last step in the data generation section of a standard genomic analysis workflow. The most commonly used platform that harnesses the power of next-generation sequencing to efficiently evaluate genomic samples is Illumina. This platform follows the protocol described in the previous section for library preparation. Amplified reads are then sequenced by adding individual fluorescently tagged and blocked nucleotides to the complementary DNA sequence and exposing the nucleotide to light to produce a characteristic fluorescence which is captured via imaging. These blocked nucleotides can then be unblocked to allow for another blocked nucleotide to bind, thus repeating the process until the entire
complementary sequence is elucidated. This platform has a high accuracy rate, can evaluate 50-300 base pairs per read, and has very high-throughput runs (millions to billions of reads per flow cell). Each run takes approximately 2-3 days to complete for as little as $1,000 per 30x whole genome sample.

Newer platforms have been developed to address several issues that currently exist with NGS. Specifically, PacBio and Oxford Nanopore allow for the sequencing of longer reads compared to NGS-based approaches. PacBio utilizes hairpin adaptors to create a loop of DNA that can be fed through an immobilized polymerase that adds complementary base pairs. While each nucleotide is held in the detection volume by the immobilized polymerase, a light pulse identifies the newly added base. PacBio requires high-quality intact DNA with highly controlled fragmentation and can generate average read lengths of 10-15 kb with their v2 chemistry. Oxford Nanopore sequencing utilizes hundreds of biological transmembrane proteins that translocalize DNA per individual flow cell. Sequence composition is determined via changes in electrical conductivity as a DNA or RNA strand passes through one of the transmembrane pores. This platform averages read lengths of 10-30 kilobase pairs (kb) for genomic libraries (but has generated reads of up to 2.3 megabase pairs (Mb) (Payne et al., 2018) and is inexpensive relative to other technologies. Interestingly, Oxford Nanopore instruments are capable of selective sequencing, or “Read Until”, which refers to the ability of a nanopore sequencer to reject individual molecules whilst they are being sequenced (Payne et al., 2021). Oxford Nanopore boasts the MinION device, which is completely portable, commercially available, and can evaluate 20-100 Mb per run. For labs wishing to generate larger amounts of long-read sequencing data, Oxford Nanopore also has the GridION and PromethION instruments, which are capable of running multiple flowcells in parallel.
For each sequencing platform described above, there are several broad classes of sequencing strategies that can be employed. This includes broad capture of the entire genome (whole genome sequencing), capture of all protein coding exons (whole exome sequencing), or targeted capture of desired loci of the genome (custom-capture sequencing). Other applications of sequencing include the evaluation of transcribed nucleic acid (RNA-Seq) and the evaluation of nucleic acid released into the blood (circulating tumor cells (CTCs) or cell-free DNA (cfDNA)). These methods and applications of sequencing technology continue to expand due to reduced sequencing costs, increased read accuracy and length, and ever-increasing research questions.

1.1.3 Alignment and variant calling of NGS data

Following the generation of raw sequencing data, alignment to the reference genome is the next step within the genomic analysis pipeline. The human reference genome approximates the complete representation of the 3.2 billion base pairs of human DNA. Using a representative assembly prevents the need to build an assembly each time a genome is sequenced; however, there are trade-offs to this approach. Specifically, due to SNPs (and large-scale variants) intrinsic to every individual, the reference genome does not perfectly match any one person. Further, due to repetitive elements (duplications, inverted repeats, tandem repeats), the commonly used reference genomes GRCh37 and GRCh38 are incomplete or incorrect in places, as seen through the completion of the Telomere-to-Telomere CHM13 (T2T-CHM13) genome, which compared to GRCh38, closes all remaining gaps, adds nearly 200 million base pairs of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the human genome for scientific inquiry (Aganezov et al., 2022). Alignment to the reference genome can be performed using various alignment software and, generally speaking, alignment strategies can either optimize
accuracy or processing time. Optimal solutions include either Smith-Waterman (Smith and Waterman, 1981) or Needleman-Wunsch (Needleman and Wunsch, 1970) alignment strategies, which are computationally expensive and process read strands slowly. Alternatively, fast solutions include hash-based algorithms such as Burrows-Wheeler transformation (M. Burrows, 1994), utilized in BWA-MEM (Li and Durbin, 2009), which create shortcuts to reduce alignment time with minimal reduction in accuracy.

The next step in the typical cancer sequencing pipeline is to use paired tumor and normal alignments for germline and somatic variant calling. Germline variant calling consists of identifying SNPs, indels, and structural variants (SVs) that are intrinsic to the normal tissue. An example of a germline variant calling pipeline is the Genome Analysis Tool Kit (GATK) from The Broad Institute (McKenna et al., 2010), which can be used for all types of variants described above. Somatic variant calling is a similar process, but it requires the variant to be exclusively observed in the tumor tissue and not present in the germline (i.e. normal) tissue. Somatic variant calling may involve looking for SNVs, indels, SVs, copy number variants (CNVs), and loss of heterozygosity (LOH), depending on the type of sequencing performed. These different types of variants can be identified by using various software (e.g., Strelka (Kim et al., 2018; Saunders et al., 2012), MuTect (Benjamin et al., 2019; Cibulskis et al., 2013), VarScan (Koboldt et al., 2009, 2012), Manta (Chen et al., 2016), and CNVKit (Talevich et al., 2016)). A common strategy is to apply ensemble or multi-caller approaches and these may be implemented in larger analysis management systems. For example, within the Griffith lab and other labs at Washington University, such variant calling is done through the Genome Modeling System (Griffith et al., 2015a).
1.2 The “Druggable Genome” as a concept

The druggable genome can be defined as the genes or gene products that are known or predicted to interact with drugs, ideally with a therapeutic benefit to patients. This concept was first coined in the late 1990s when a comprehensive review of the drug portfolio of the pharmaceutical industry identified 483 targets and hypothesized that there could be 5,000-10,000 potential targets through an estimate of the number of disease-related genes (Drews, 1996, 2000; Drews and Ryser, 1997). This initial analysis failed to investigate the properties of drugs that defined those 483 targets (Drews, 2000). Utilizing the Investigational Drugs Database and the Pharmaprojects Database, in addition to a comprehensive review of the literature at the time, Hopkins and Groom identified 399 molecular targets which were potentially druggable (Hopkins and Groom, 2002). An updated approach utilizing a similar analysis technique identified 2,917 druggable genes within the Ensembl gene annotations (release 32) (Russ and Lampel, 2005).

Since these initial foundational studies, there has been extensive additional work to identify more potentially druggable genes according to their membership in gene categories associated with druggability (e.g. kinases) or through containing protein domains that are known to interact with existing drugs. Furthermore, as genetic and genomic techniques have advanced over the years, there are an increasing number of genes with known drug interactions. These advances have happened in dozens of research groups around the world, and the information is summarized in disparate individual publications and databases. Assembling all of the information about known drug-gene interactions and potential gene druggability is a major challenge due to the diverse and numerous set of publications and databases that house this information. Each resource may report relevant associations differently, including the use of ENSEMBL Gene IDs vs Gene Symbols, trade vs generic drug names, or by publishing the data in such a way that makes it hard
to easily access (e.g. as a PDF with drug-gene associations or a list embedded in a PDF of potential druggable genes).

The Drug-Gene Interaction database (DGIdb) (Griffith et al., 2013) was created in order to consolidate the existing information from genome-wide studies to help researchers interpret the context of the druggable genome. At its creation, DGIdb integrated data from 13 primary sources that covered disease-relevant genes (Flicek et al., 2011; Maglott et al., 2011), drugs (Wang et al., 2012), drug-gene interactions (Knox et al., 2011; McDonagh et al., 2011; Rask-Andersen et al., 2011; Somaiah and Simon, 2011; Zhu et al., 2010), and potential druggability (Ashburner et al., 2000; Hopkins and Groom, 2002; Kumar et al., 2013; Russ and Lampel, 2005). This included 14,144 drug-gene interactions involving 2,611 genes and 6,307 drugs, and 6,761 genes belonging to one or more of 39 potentially druggable gene categories. In total, 7,668 unique genes had either known or potential druggability.

DGIdb received its first update in 2016 (Wagner et al., 2015). The 2.0 update to DGIdb added 14 new sources, bringing the cumulative number of sources that described drug-gene interactions and druggable gene categories to 27. Through the inclusion of new sources and updates of existing sources, DGIdb 2.0 contained 26,298 unique drug-gene interaction claims involving 2,644 genes and 7,569 drugs, and 7,524 unique genes belonging to one or more of 41 potentially druggable gene categories. A total of 8,419 unique genes had known or potential druggability. While these numbers fall within the early predictions for the number of targets, there is still plenty of ‘dark matter’ that has yet to be identified within the druggable genome (Edwards et al., 2011). Chapters 2 and 3 will discuss further efforts to integrate novel findings with respect to the druggable genome while also highlighting new and improved features of DGIdb to make it more useful to other researchers.
1.3 The use of next-generation sequencing in cancer research

The first NGS-generated cancer genome was sequenced and reported on in 2008 (Ley et al., 2008). Individual cancer genome sequencing studies such as this study highlighted the importance of sequencing a patient’s normal genome in addition to the tumor genome (Ding et al., 2010a), in order to distinguish germline alterations from somatic mutations. These initial small studies were expensive at the time, but as the cost of generating NGS data decreased, other studies began to dramatically advance the field of cancer research through the assessment of the mutational landscapes, changes in expression patterns and signatures, and identification of other molecular features of tumors and their microenvironments (Dellaire et al., 2013; Horak et al., 2016; Mardis, 2019). To date, there have been numerous studies from large consortium efforts which have performed comprehensive characterization and stratification of cancer cohorts using DNA sequencing (e.g. whole genome, exome, custom-capture) and transcriptome sequencing (e.g. bulk RNA-Seq), in addition to methylomics, metabolomics, and proteomics, especially of bulk tumor samples (Alexandrov et al., 2020; Cancer Genome Atlas Network, 2012; Curtis et al., 2012; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; International Cancer Genome Consortium et al., 2010; The Cancer Genome Atlas Network *, 2015).

Somatic variants are acquired over the course of time, due to mutagens and carcinogens (e.g. tobacco, UV damage), errors in DNA replication, along with failure of DNA damage repair mechanisms. As these mutations are acquired, their impact on the cell’s biology is conferred based upon their effect on gene function and expression. Tumor cells contain both driver events as well as other mutations, known as passenger mutations, acquired in prior cell divisions. Driver mutations have been distinguished from passengers through computational analyses of large patient cohorts, where driver genes are recurrently mutated across patients, leading researchers to
discover and more selectively study genes implicated in tumorigenesis (Cancer Genome Atlas Network, 2015; Sjöblom et al., 2006; Wood et al., 2007). Mutations are annotated at the genomic level by the changes in DNA sequence, and are further assessed for their impact on protein sequence, structure, and function (McLaren et al., 2016). Driver genes have been associated with not only cell-autonomous pathways regarding cell survival (e.g. activation of growth factor receptor pathways, inactivation, and resistance to cell death), but also signaling to the microenvironment and immune system (e.g. recruitment of pro-tumor inflammation, activation of angiogenesis, suppressive signaling to anti-tumor immune cells) (Hanahan and Weinberg, 2011).

Clinically, annotation of sequence alterations can be used to identify diagnostic, prognostic, and predictive biomarkers associated with drug mechanisms (Ainscough et al., 2016; Cerami et al., 2012; Gao et al., 2013; Griffith et al., 2017). Some somatic mutations are viable therapeutic targets, while others confer resistance to treatment approaches, leading to the foundation of precision oncology (Griffith et al., 2017; Hyman et al., 2015).

1.4 The current mutational landscape of breast cancer

Breast cancer is the most commonly diagnosed cancer among women, with an estimated 1.7 million new cases and 500,000 deaths attributed to it each year (Cancer Genome Atlas Network, 2012; Torre et al., 2015). The prognosis for this disease varies depending on the stage of the disease at the time of detection. Early stages of breast cancer (stages 0-II) usually have greater than 90% relative survival at 5 years, while stages III and IV have greatly decreased relative survival rates, 72% and 22%, respectively. Beyond staging, prognosis and treatment can vary depending on the molecular features of the tumor. Microarray analysis in the early 2000s was used to define the concept of molecular subtypes within breast cancer and resulted in the
definition of five distinct subtypes: luminal A, luminal B, HER2 enriched, triple-negative/basal-like, and normal-like (Perou et al., 2000; Sorlie et al., 2003). While these subtypes are based on molecular expression signatures, they are often thought about with respect to their expression of the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) (Tofigh et al., 2014). The different subtypes vary significantly with respect to gene expression, tumor presentation, and prognosis (Dai et al., 2015). The genomic mutational landscape of coding regions within these breast cancer subtypes has been thoroughly described (Cancer Genome Atlas Network, 2012; Ding et al., 2010b; Shah et al., 2009, 2012). These studies have stratified subtypes by mutational rate within coding regions, copy number changes, and recurrently mutated genes. While the aforementioned studies focused on coding mutations, there has been increased interest in better understanding the role of noncoding variation in breast cancer. This interest partially derives from the initial discovery of recurrent \textit{TERT} promoter mutations resulting in increased TERT expression in melanoma (Horn et al., 2013; Huang et al., 2013). Several studies have attempted pan-cancer analyses in order to identify these types of mutations, but often the results suffer from low sequence coverage of regulatory regions or do not extend to breast cancer (Fredriksson et al., 2014). While some breast cancer-specific studies have identified some significantly mutated promoters and long non-coding RNAs (lncRNAs), they have often failed to incorporate transcriptome data to assess the impact and relevance of mutations on the expression and RNA splicing of genes within these tumors (Nik-Zainal et al., 2016; Rheinbay et al., 2017; Weinhold et al., 2014). Additionally, while these studies have been able to identify recurrent mutations in large cohorts, their analyses fail to describe subtype-specific recurrent, noncoding mutations and their consequences. Chapters 4 and 5 will discuss work done to address these current shortcomings within breast cancer research.
1.5 Splice-associated variants are under-identified in current analysis workflows

Canonical pre-mRNA splicing involves the removal of introns and ligation of the exons of a gene by the spliceosome, a ribonucleoprotein complex (Chen and Cheng, 2012; Faustino and Cooper, 2003). Assembly of the spliceosome and initiation of splicing is guided by conserved sequences near exon/intron boundaries (Burset et al., 2000; Parada et al., 2014). Mutations in these consensus splice site sequences or more variable regulatory elements known as exon and intron splicing enhancers (ESEs and ISEs, respectively) and silencers (ESSs and ISSs, respectively) have been estimated to account for 15-50% of genetic diseases (Cáceres and Kornblihtt, 2002; Faustino and Cooper, 2003; Matlin et al., 2005). More specifically, these enhancers and silencers can affect the spliceosome’s preference for using canonical splice sites. When the spliceosome does not use a canonical splice site, this is known as alternative splicing (AS). AS has been noted as a hallmark of cancer, with around 15,000 identified AS events that are associated with cancer biology processes (He et al., 2009; Shapiro et al., 2011; Venables et al., 2008). Such events are well documented in breast cancer. For example, several proteins involved in splicing have been found to be deregulated in breast cancer, resulting in increased AS, including serine and arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) (Silipo et al., 2015). Additionally, recurrent mutations within the spliceosome have been found in 5.6% of breast cancers (Maguire et al., 2015). A specific example of AS within breast cancer includes a variant form of BRAF, which lacks exons 4-8, and confers resistance to RAF inhibitors (Poulakakos et al., 2011). Furthermore, a recent pan-cancer analysis from Climente-González et al. revealed that there was a negative correlation between the number of AS changes in a tumor sample and the number of somatic mutations in cancer drivers (Climente-González et al., 2017). This paper predicted that these AS changes
might result in oncogenic processes and could be considered an AS driver of the tumor. Breast cancer was found to have isoform switches, indicative of AS, in multiple genes including TP53 and HER2 (Climente-González et al., 2017). When all breast cancer samples were analyzed for their relative proportion of AS events compared to mutational drivers, 25% of breast cancer samples were found to be AS driven. A more comprehensive understanding of these events, both across and within cancer types, could lead to better prognostic indicators and novel therapies targeting non-canonical protein products. Such an analysis will be discussed within Chapter 6 of this dissertation.

1.6 Summary
Overall, this dissertation will summarize a set of databases, tools, and bioinformatic approaches to address the precision oncology bottleneck. This bottleneck refers to the challenge of interpreting a growing deluge of measurable tumor-specific molecular alterations for clinical relevance. Cancers are characterized by hundreds or thousands of individual alterations, some of which are drivers of tumor progression but most of which are passengers. A subset of these drivers are clinically actionable. Initial studies have focused almost entirely on protein-altering mutations for diagnostic, prognostic, and predictive information. However, the next phase of precision oncology requires extension to those alterations that affect non-coding, splicing, or other kinds of alterations. Tools are needed to harmonize existing disease-relevant data for ease of interpretability and identify novel alterations in other categories of mutations (e.g. promoter region mutations, splice-associated variants, etc.) to alleviate the precision oncology bottleneck while also furthering our understanding of the somatic alterations that drive tumor biology. Furthermore, such tools and the efforts surrounding their development and maintenance should be open-source to allow for advancement of science while providing reproducibility and
transparency. This dissertation will attempt to address these issues through describing databases that harmonize druggability and regulatory studies, analysis of regulatory non-coding variants in breast cancer, and the pan-cancer identification of splice-associated variants using RegTools, a novel software suite.
Chapter 2: DGIdb 3.0: a redesign and expansion of the drug-gene interaction database

2.1 Preamble

The following chapter has been published:


As an author of the published manuscript, and in compliance with the editorial policies at Nucleic Acids Research, the cited publication is included in full in the following chapter. I would like to distinguish my role from the other first-author in this study. I led curation and content update efforts while Alex Wagner led efforts to update the user interface (UI) and backend structure changes. I primarily wrote the manuscript and generated figures with input from Alex and other co-authors. Additionally, I was responsible for manuscript submission. Alex and I were equally responsible for addressing reviewers’ comments during the resubmission process.

2.2 Summary

The drug-gene interaction database (DGIdb, www.dgidb.org) consolidates, organizes, and presents drug-gene interactions and gene druggability information from papers, databases, and web resources. DGIdb normalizes content from 30 disparate sources and allows for user-friendly advanced browsing, searching, and filtering for ease of access through an intuitive web user interface, application programming interface (API), and public cloud-based server image. DGIdb v3.0 represents a major update of the database. Nine of the previously included 24 sources were updated. Six new resources were added, bringing the total number of sources to 30. These updates and additions of sources have cumulatively resulted in 56,309 interaction claims. This
has also substantially expanded the comprehensive catalog of druggable genes and anti-
neoplastic drug-gene interactions included in the DGIdb. Along with these content updates, v3.0
has received a major overhaul of its codebase, including an updated user interface, preset
interaction search filters, consolidation of interaction information into interaction groups, greatly
improved search response times, and upgrading the underlying web application framework. In
addition, the expanded API features new endpoints which allow users to extract more detailed
information about queried drugs, genes, and drug-gene interactions, including listings of PubMed
IDs, interaction type, and other interaction metadata.

2.3 Introduction

The drug-gene interaction database (DGIdb, www.dgidb.org) was first released in 2013 to
consolidate drug–gene interactions and potentially druggable genes into a single resource with a
powerful interface to query these data (Griffith et al., 2013). DGIdb 2.0 was released in 2016 and
featured substantial content updates, a more intuitive user interface, and the inclusion of an
application programming interface (API) (Wagner et al., 2015). Despite the success of the
DGIdb, 2.0 became outdated in several ways. New sources have become available since 2.0’s
release and many existing sources within the database required updates. The notion of grouping
drug claims (an assertion of a drug concept by a constituent resource) to match a canonical drug
source (PubChem compounds), and grouping gene claims (an assertion of a gene concept by a
constituent resource) to a canonical gene source (NCBI Entrez Gene) was described in the 2.0
paper (Brown et al., 2015; Kim et al., 2016). In brief, grouping allows the DGIdb to relate
disparate representations of a gene or drug concept through a core entity, or group. As the
database grew, the grouping success rates dropped and new classes of drugs were not fully
represented in the search results. Additionally, while the DGIdb grouped drug and gene claims,
interactions were still reported at the claim level, with no concept of interaction groups. Furthermore, the initial API that was released alongside 2.0 had limited endpoints that prevented users from extracting additional information about their queried drug–gene interactions, such as PubMed IDs (PMIDs) supporting the interactions and other interaction metadata. Finally, the DGIdb has grown enormously since its initial publication in 2013, and as a result, interaction searches had become significantly slower. This new release addresses these limitations, providing an improved user interface, quicker response times for searches, and new methods for accessing and incorporating the DGIdb into bioinformatics workflows.

2.4 Results

2.4.1 Expanded Content

For 3.0, there has been a substantial expansion of the content of the database through the addition of new sources and the updating of existing sources. Six new sources have been added, which brings the total number of sources represented by the DGIdb to 30. Three of these new sources provide drug-gene interactions from prominent expert-curated databases of clinically actionable variants similar to the CIViC and Clearity Foundation sources, already included in the DGIdb (Supplementary Table S1) (Griffith et al., 2017). These include the Precision Oncology Knowledge Base (OncoKB), cancer genome interpreter (CGI), and The Jackson Laboratory Clinical Knowledgebase (CKB) (Supplementary Table S1) (Chakravarty et al., 2017; Patterson et al., 2016). From these three new sources, 2822 new interaction claims were added. These sources were added as anti-neoplastic interaction sources, resulting in a 36.7% increase in anti-neoplastic drugs. Drug–gene interactions were also added from the United States Food and Drug Administration (FDA) Pharmacogenomics website and the National Cancer Institute (NCI)
Cancer Gene Index, resulting in an additional 276 and 6,231 interaction claims, respectively (Supplementary Table S1). In addition to these sources, the existing druggable gene category sources were expanded with data extracted from a new druggable genome paper that used computational approaches to identify druggable genes from genome-wide association studies (Supplementary Table S2) (Finan et al., 2017). The inclusion of this source into the DGIdb results in an additional 2300 potentially druggable genes, a 58% increase from 2.0 (Figure 2.1). many other gene categories (e.g. kinases, G-protein coupled receptors) that are expected to be good drug targets (Supplementary Figure S1).

![Figure 2.1 DGIdb 3.0 content by source. The number of drug–gene interaction claims (first panel) and druggable gene categories (second panel) are separated into three categories: sources that existed in the DGIdb previously, sources that existed in the DGIdb previously but have been updated for 3.0 or sources that are entirely new to the DGIdb. Abbreviations: CF: Clearity Foundation; CGI = Cancer genome interpreter, CKB = JAX-Clinical Knowledgebase, CMI = Caris Molecular Intelligence, FO = Foundation One, GTP = Guide to pharmacology, MCG = My cancer genome, OncoKB = Precision Oncology Knowledge Base, TALC = Targeted agents in lung cancer, TTD = Therapeutic Target Database, TEND = Trends in the exploration of novel drug targets, GO = Gene Ontology and MSK = Memorial Sloan Kettering.](image)

Notably, this independent definition of the druggable genome almost completely encapsulates As the DGIdb has matured, maintaining current representations of existing sources has become increasingly important relative to identifying new interaction sources. Curated
updates of drug–gene interaction data from My Cancer Genome and the targeted and biologic therapies for non-small-cell lung cancer (TALC) have occurred, resulting in moderate increases in drug–gene interactions from these sources (Simon and Somaiah, 2014; Yeh et al., 2013). To help prevent regularly updated sources from becoming outdated in the DGIdb, we have rewritten and expanded online updaters to keep frequently updated sources current within the database. With these more frequent, incremental updates to the DGIdb, there has been a substantial increase in drug–gene interaction content from IUPHAR’s Guide to Pharmacology and CIViC, a 23 and 200% increase, respectively (Griffith et al., 2017; Southan et al., 2016). Additionally, a major update of DrugBank from version 4.3 to 5.0 was performed, resulting in a 30% increase in interaction claims from this source (Law et al., 2014). Gene Ontology was also updated which provided a moderate increase for the druggable gene categories (Gene Ontology Consortium, 2015). For this release, 18,493 new interaction claims were added, of which 50% were from updated sources and 50% from new sources. In total, there are now 6,106 druggable genes and 29,783 drug–gene interactions, which cover 41,100 genes and 9,495 drugs, within the DGIdb.

The enhancements to the online updaters have also been applied to Entrez Gene, from which 99% of all gene claims made by the DGIdb constituent sources were grouped (Supplementary Figure S2) (Brown et al., 2015). Another major change from 2.0 to 3.0 is that the canonical drug source for the DGIdb has switched from using PubChem compounds to ChEMBL molecules (Bento et al., 2014). This switch has added 1.7 million ChEMBL molecules to the database for potential matching to drug claims. Importantly, switching to ChEMBL has added 195 antibody drugs (e.g. trastuzumab, cetuximab), a drug class that is absent from the PubChem database and frequently requested by users. These antibody drugs matched to 539 distinct drug claims from the constituent sources of the DGIdb. With ChEMBL as the canonical
drug source and the improvements to the grouping strategy below, 80.2% of all drug claims now group. Many of the resources we pull from strive to be as comprehensive as possible, and sometimes include broad classes of drug or therapy (e.g. ‘hormone therapy’, ‘mtor inhibitors’, ‘chemotherapy’, ‘radiation’, ‘antibiotics’, etc.), which account for a large percentage of the remaining drug claims.

2.4.2 New Features and Enhancements

A major feature added in 3.0 is a user-selectable series of preset filters. These allow users to select drug-gene interactions that align with some of the most common search use cases. While it was possible for a user to filter on some of the following attributes in 2.0, these concepts have now been structured as stand-alone filters that can help focus a user's search results with a single click. The filters currently include FDA-approved drugs, anti-neoplastic drugs, immunotherapy drugs, clinically actionable genes, genes included in the druggable genome definition, and drug-resistant genes. FDA approval status is extracted from ChEMBL v23 (current as of this writing) (Bento et al., 2014). Anti-neoplastic drugs are defined by inclusion in an anti-neoplastic drug-gene interaction source (e.g. My Cancer Genome), or as a drug with an anti-neoplastic attribute from its constituent source. Immunotherapy drugs are defined as any drug with an attribute of ‘immunosuppressive agent’, ‘immunomodulatory agent’ or ‘immunostimulant’. Clinically actionable genes are genes that constitute the DGIdb ‘clinically actionable’ gene category, and by definition are used to inform clinical action (e.g. the Foundation One diagnostic gene panels). Similarly, druggable genome genes are genes listed in the DGIdb ‘druggable genome’ gene category (Cheng et al.; Edwards et al., 2011; Finan et al., 2017; Gene Ontology Consortium, 2015; Hopkins and Groom, 2002; Kumar et al., 2013; Russ and Lampel, 2005; Southan et al., 2016). Drug resistance genes are defined by the Gene Ontology as genes that confer drug
resistance or susceptibility (GO identifier 0042493), and are maintained in the DGIdb through the ‘drug resistance’ gene category (Gene Ontology Consortium, 2015). To incorporate these new filters, we have redesigned the data model, expanded the definition of the druggable genome, and restructured the user interface (UI) to include a redesigned search form and results interface. The UI also now features drug, gene, and interaction views to leverage the new grouping strategies and preset filters.

Since the last release, several changes have been made to the drug grouping strategy to improve overall grouping percentages. We have added support for fuzzy searching when grouping if direct matches to drug groups are not found, enabling grouping of drug claims with slightly different means of joining multi-word terms (whitespace, dashes, and underscores). Code has also been added to identify and remove aliases that are highly ambiguous to improve the exact matching of drug claims.

In addition to improved drug grouping, we have added interaction grouping, linking together interaction claims from multiple sources that describe the exact same drug-gene interaction. The grouping efficiency for interactions is 75.2%. The successful grouping of an interaction claim is dependent on the associated gene claims and drug claims successfully grouping. As a result, interaction grouping percentages are closely related to the grouping percentages for drugs and genes (Supplementary Figure S2). The introduction of interaction groups has led to a noticeable improvement in response times on the UI for interaction searches, even when query sizes and result sizes grow larger (Figure 2.2). Before interaction groups were created, searches in 2.0 had to query almost twice as many database tables as in 3.0. Interaction groups allow for more efficient queries, which leads to a 14-fold reduction in response times when searching the DGIdb.
We have expanded and added additional API endpoints that now allow users to extract more information on the drug-gene interactions provided by the DGIdb. These include endpoints to list all the interaction groups, gene groups, and drug groups in the DGIdb, as well as endpoints to view detailed information about an individual interaction, gene, or drug group. Drug/gene/interaction group endpoints include various metadata about the group including its constituent claims. The information presented in the interaction search results endpoint has been expanded to include the Entrez and Chembl ID of the gene and drug involved in an interaction, as well as a list of publications that support each interaction. Users can also apply the preset filters to their interaction search through the API. This brings the interaction search endpoint in sync with the user interface. Moreover, the information presented in the interaction search results and the new individual endpoints for interactions, genes, and drugs more closely mirrors the views available via the user interface. These new endpoints allow users to more efficiently export all of the data available in the DGIdb.
2.4.3 User Interface (UI) Updates

The 3.0 release of the DGIdb features a dramatic overhaul of the UI that reflects many of the backend changes detailed above. The search interactions page now allows the user to apply the new preset filters in addition to the existing advanced filters (e.g. source databases, gene categories, and interaction types) directly on the search form (Figure 2.3). On performing a search, the user is redirected to the redesigned search interactions results page, enabled by a new concept of interaction groups instead of individual interaction claims. This new view displays search results as lists of visually distinct panels for each search term mapped to gene or drug groups. A list of uniquely matched terms is shown in one tab whereas a second tab summarizes the list of ambiguously matched search terms and unmatched terms. Each panel displays a table of interaction groups, along with summaries of the respective interaction types, sources, supporting publication PMIDs, and a ranking. Users interested in a particular interaction can navigate from the interaction search results to the corresponding interaction group view. Similarly, detailed information about an interacting gene or drug can be found in the associated gene or drug group view (Supplementary Figure S3).
Figure 2.3 The DGIdb interaction search interface. (A) A search field accepts drug or gene identifiers, depending on which tab is selected and provides auto-completion suggestions for search terms. (B) Preset options are provided to filter search results based on the attributes of mapped interactors. (C) Advanced filters allow the user to further filter based on source database, gene category or interaction type. (D) The interaction search results (Unique Matches) view shows results for search terms that were matched within the DGIdb. (E) The interaction search results (Ambiguous or Unmatched) view shows search terms that were either ambiguously matched or unmatched within the DGIdb. (F) Additional side panels provide a brief summary of unmatched and ambiguously matched terms.
These new group views feature (i) a summary tab, which details all relevant information collected from the various source claims (e.g. aliases, FDA approval status, supporting publications); (ii) an interactions tab (for gene and drug group views), which lists summary panels of each interaction in the DGIdb for this gene or drug; and (iii) a claims tab with detailed information about each claim supporting the group.

2.4.4 Usage and Accessibility

The utility of the DGIdb as a resource is reflected in its substantial web and especially API traffic. The website receives $\sim 1700$ unique users and $\sim 2800$ sessions a month, an increase of 42 and 39%, respectively, since the initial release of the DGIdb. Additionally, 42% of unique visitors within the past year were returning users. Within the last year, the DGIdb API has been used by 3689 unique IP addresses for a total of 939,429 requests, as seen in Supplementary Figure S4. At the time of publication, 2.0 was used by a number of bioinformatics tools in the development of their platforms. DGIdb has since been integrated into four additional tools, bringing the total number of known platforms that utilize the DGIdb to 10. These platforms are GeneCards (http://www.genecards.org), CVE (Mock et al., 2017), rDGIdb (Thurnherr et al., 2016), PANDA (Hart et al., 2015), iCAGES (Dong et al., 2016), BioGPS (Wu et al., 2009), Omics Pipe (Fisch et al., 2015), GEMINI (Paila et al., 2013), StationX (www.stationxinc.com) and IHLDB.rf (www.lungcancerdatabase.com). These integrations highlight the accessibility and utility of the DGIdb API, which has been expanded to include new endpoints in 3.0.

To handle the increased usage of the DGIdb, the database backend has undergone significant updates. One of the most notable is the upgrade from Rails 3 to the Rails 5 framework. We also added 22 new tables to the PostgreSQL database schema consisting of 2,326,676 records supporting the described changes and added 4,065 new lines of code to the
repository, an expansion of the codebase by 18.9%. This activity can be seen through the commit history to the GitHub repository in Supplementary Figure S5. For access to this code and data, please see the availability section below.

The open-source DGIdb software was previously available under the GNU General Public License 3.0 (GPL3) and has been re-released under the more permissive MIT license. The data contained within the DGIdb are available under the licenses assigned by their host sources which makes it possible to integrate the DGIdb into any workflow. Since 2.0, the database has moved from being hosted on a local server to being hosted on Amazon Web Services. With this change in server host, the availability of the DGIdb has been expanded through the release of two Amazon Machine Images, a development environment and production environment. By providing this production environment for users, we now support a quick-start solution for private, cloud-based applications.

2.5 Conclusion and Future Directions
DGIdb 3.0 has undergone significant changes. The number of drug–gene interactions and druggable genome definitions has been substantially expanded through the updating of existing sources and inclusion of new sources. To improve searches, we have included preset filters that allow searches for commonly requested gene or drug classes. To utilize these filters, a more intuitive UI was created and new fields to the database schema were added. Additionally, there are new ways to both access the data within the DGIdb and to deploy local or cloud instances of the database.

While the DGIdb remains a powerful resource for querying drug–gene interactions, we anticipate future changes that will improve the user experience and the content within it. One change is that claim information will be linked to the licensing terms from constituent sources.
through the various methods of accessing the DGIdb including the web interface, API endpoints and data downloads. Second, while grouping statistics for claims have improved in the DGIdb 3.0, there are still a significant percentage of claims that remain ungrouped. To address this, the drug grouper will need to be optimized to handle the exceptions that currently prevent successful grouping. Third, much of the user interface is a series of static renders by the server. A potential future enhancement is a more reactive, client-side web application that would allow for more dynamic visualization and exploration of the data. Finally, other databases (e.g. CIViC) have had success utilizing a community curation model. In an effort to enable community feedback to address the complexity of grouping and representing drug–gene interaction data, the DGIdb will be moved to a similar model. Expected changes include users creating login accounts, suggesting sources or claims for inclusion and verification that a source or claim has enough support to be included within the DGIdb. These changes will require significant code updates but will greatly increase the utility of the DGIdb.

2.6 Availability

DGIdb is an open-access database and web interface (www.dgidb.org) with open source code available on GitHub (https://github.com/griffithlab/dgi-db). We also provide data downloads for drug claims, gene claims, and interaction claims on the website in addition to a SQL data dump (http://dgidb.org/downloads). Information about the API and its endpoints can also be found on the website (http://dgidb.org/api).

2.7 Supplementary Data

Supplementary Data are available at NAR Online. (This includes supplementary tables and figures).
2.8 Funding

National Human Genome Research Institute [R00HG007940 to M.G.]; National Cancer Institute [K22CA188163 to O.L.G., T32CA113275, F32CA206247 to A.H.W, T32CA113275 to K.C.C., U01CA209936]; National Institute of General Medical Sciences [5R25GM103757]. Funding for open access charge: National Institutes of Health [R00HG007940].
Chapter 3: Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts

3.1 Preamble

The following chapter has been published:


* denotes co-first authors

As an author of the published manuscript, and in compliance with the editorial policies at Nucleic Acids Research, the cited publication is included in full in the following chapter. I would like to distinguish my role from the other first-authors in this study. All co-first authors contributed to updating content within DGIdb and writing of the manuscript. Furthermore, I generated figures and identified new sources for inclusion within DGIdb 4.0. Sharon, Susanna, and I were equally responsible for addressing reviewers’ comments during the resubmission process. In addition, for the previous version of DGIdb (v3.0), I led curation and content update efforts while Alex Wagner led efforts to update the user interface (UI) and backend structure changes. I primarily wrote the v3.0 manuscript and generated figures with input from Alex and other co-authors.

3.2 Summary

The Drug-Gene Interaction Database (DGIdb, www.dgidb.org) is a web resource that provides information on drug-gene interactions and druggable genes from publications, databases, and
other web-based sources. Drug, gene, and interaction data are normalized and merged into conceptual groups. The information contained in this resource is available to users through a straightforward search interface, an application programming interface (API), and TSV data downloads. DGIdb 4.0 is the latest major version release of this database. A primary focus of this update was integration with crowdsourced efforts, leveraging the Drug Target Commons for community-contributed interaction data, Wikidata to facilitate term normalization, and export to NDEx for drug-gene interaction network representations. Seven new sources have been added since the last major version release, bringing the total number of sources included to 41. Of the previously aggregated sources, 15 have been updated. DGIdb 4.0 also includes improvements to the process of drug normalization and grouping of imported sources. Other notable updates include the introduction of a more sophisticated Query Score for interaction search results, an updated Interaction Score, the inclusion of interaction directionality, and several additional improvements to search features, data releases, licensing documentation, and the application framework.

3.3 Introduction

Originally released in 2013, the Drug–Gene Interaction database (DGIdb) (Griffith et al., 2013) serves as a central aggregator of information on drug-gene interactions and druggability from multiple diverse sources. The subsequent major updates to DGIdb 2.0 (Wagner et al., 2015) (in 2016) and 3.0 (Cotto et al., 2017) (in 2018) included improvements to the user interface and search response times, the addition of an API, the introduction and improvement of gene and drug grouping methods, and the expansion of source content through the inclusion of new sources and updates of existing sources. Since the release of DGIdb 3.0, many of the existing sources have been substantially updated and new sources have become available. Here we
describe changes made for our most recent major version release, DGIdb 4.0. In this release, we have made an effort to integrate crowdsourced data and sources in several areas, including the addition of the crowdsourced Drug Target Commons (Tanoli et al., 2018) as a drug-gene interaction source, and the use of the open, community-curated Wikidata (Vrandečić and Krötzsch, 2014) resource for drug normalization. We also illustrate the value of our integration efforts in downstream community tools, through the incorporation of our data into NDEx (Pratt et al., 2017). To keep content offered by DGIdb current, we have developed additional automatic update routines for multiple sources and implemented a new background job management system (Sidekiq, sidekiq.org) for routine job scheduling. Finally, DGIdb 4.0 focuses on numerous improvements of search results, including new and updated scores for interaction search results and improved drug normalization routines.

### 3.4 Results

#### 3.4.1 Integration with crowdsourced efforts

A primary focus of the DGIdb 4.0 release is the inclusion and utilization of crowdsourced efforts in several aspects of our database. The utility of our database begins with importing relevant drug, gene, and drug-gene interaction records (called claims) from outside resources. We normalize and sort these claims into conceptual groups, and make these concepts searchable via a web application and API. We also export data for bulk download and use with external resources (Figure 3.1). In this update, we extend these features by integrating with crowdsourced drug-gene interaction claims, normalizing drug terms, and integrating with external resources.
Figure 3.1 Overview of main components of DGIdb. Data sources are imported from outside resources (over 40 as of DGIdb 4.0), normalized and grouped with internal processes to prepare records to be displayed in DGIdb, and exported to TSV for download and integration with other resources. Process management is handled by Sidekiq for automation of importing, normalization and grouping, and exporting. A subset of new data sources are highlighted in green, a subset of updated pre-existing data sources are highlighted in blue. The updated sources highlighted in this figure are some of the sources that have been updated through manual curation. Information on additional sources and their status in DGIdb 4.0 can be found in Figure 3.2 and Supplementary Table S1. New features and technologies from DGIdb 4.0 are indicated with green dots, pre-existing features and technologies that have been updated are indicated with blue dots. The drug-gene network graph shown in the bottom right is an example of the data visualizations available on NDEx. Abbreviations: CMI = Caris Molecular Intelligence, FO = Foundation One, DTC = Drug Target Commons and HPA = Human Protein Atlas.
For drug–gene interaction claims, we have added Drug Target Commons as a new source in DGIdb 4.0 (Figure 3.1). Drug Target Commons provides an extensive curated database of crowdsourced drug-gene interactions, from which we added a total of 23,879 interaction claims. This represents ∼24% (23,879/100,273) of the total interaction claims in DGIdb.

For drug normalization, we now use a Wikidata normalizer in addition to a ChEMBL (Mendez et al., 2019) normalizer from the thera-py python package (Figure 1; additional detail in Drug Grouping Improvements section). Wikidata serves as a source of collaborative, crowdsourced drug concepts, and has allowed us to improve normalization in cases where ChEMBL normalization failed. For example, concepts representing the terms annamycin, N-methyl scopolamine and Debio 1347 are all found in Wikidata but not ChEMBL.

Finally, we have integrated DGIdb with the Network Data Exchange (NDEx) (Pratt et al., 2017), a community resource that allows sharing and publishing of biological data in a network-based format. For DGIdb, export of DGIdb data to the NDEx platform provides a resource for the visual representation of relationships and interactions between drugs and genes present in our database, allowing users to visually explore a global network of drugs and gene interactions of interest. NDEx TSVs are generated monthly and automatically uploaded to the NDEx server to keep the DGIdb network in NDEx up-to-date (Figure 3.1). NDex is the latest in a number of community resources that have integrated DGIdb. Existing data clients include GeneCards (Stelzer et al., 2016), BioGPS (Wu et al., 2009), CancerTracer (Wang et al., 2020a), Gene4Denovo (Zhao et al., 2020), SL-BioDP (Deng et al., 2019), TargetDB (De Cesco et al., 2020), and OncoGemini (Nicholas et al., 2021), among others.
3.4.2 New and updated sources

In an effort to ensure that DGIdb offers diverse and contemporary information, we have updated and added several sources to DGIdb 4.0. In addition to the previously mentioned Drug Target Commons (Tanoli et al., 2018), we now also include COSMIC (Tate et al., 2019) as a new source of drug-gene interaction data (Supplementary Table S1). COSMIC also serves as an additional source of curated Drug Resistance gene category claims. Other new gene category sources include the Tempus xT (Beaubier et al., 2019) panel of actionable cancer therapy target genes, a list of the top priority genes from the Illuminating the Druggable Genome (IDG) Initiative (Rodgers et al., 2018), the Human Protein Atlas (Uhlen et al., 2017), the Oncomine (Williams et al., 2018) clinical cancer biomarker assay, and understudied targets of the IDG program from Pharos (Nguyen et al., 2017) (Supplementary Table S1). From these new sources, we have added 23,916 new drug-gene interaction claims and 8,478 new druggable gene category claims (Figure 3.2). In total, we have added two new sources of drug-gene interactions and five new sources of druggable gene category claims. DGIdb 4.0 now has 100,273 interaction claims and 33,577 druggable gene category claims. In total, there are now 10,606 druggable genes and 54,591 drug-gene interactions, which cover 41,102 genes and 14,449 drugs, within the DGIdb.
Figure 3.2 DGIdb 4.0 content by source. The number of drug-gene interaction claims (first panel) and druggable gene categories (second panel) are separated into three categories: sources that are new, sources that existed in the DGIdb previously but have been updated for 4.0, or sources that existed previously but have not been updated. Abbreviations: CF = Clearity Foundation, CGI = Cancer Genome Interpreter, CMI = Caris Molecular Intelligence, DTC = Drug Target Commons, FO = Foundation One, GO = Gene Ontology, GTP = Guide to Pharmacology, HPA = Human Protein Atlas, IDG = Illuminating the Druggable Genome, JAX-CKB = JAX-Clinical Knowledgebase, MCG = My Cancer Genome, MSK = Memorial Sloan Kettering, OncoKB = Precision Oncology Knowledge Base, TALC = Targeted Agents in Lung Cancer, TDG = The Druggable Genome, TEND = Trends in the Exploration of Novel Drug targets and TTD = Therapeutic Target Database.

We have also updated multiple sources including large, well-curated sources such as DrugBank (Wishart et al., 2018), Guide to Pharmacology (Armstrong et al., 2020), Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2019), OncoKB (Chakravarty et al., 2017), PharmGKB (Whirl-Carrillo et al., 2012), and the Therapeutic Target Database (Wang et al., 2020b) (Figure 3.2, Supplementary Table S1). To facilitate routine updates, the importer for PharmGKB has been updated to an online importer that can be run periodically using DGIdb's new automated job scheduling system (see Technology Improvements). Similarly, Pharos (Nguyen et al., 2017), one of our new sources, has been implemented as an online updater. Of the 41 sources in DGIdb, 12 sources are now imported using the online updater format, including Entrez (Brown et al., 2015), the core source of gene concepts for gene grouping in DGIdb, and Ensembl (Yates et al., 2020), a key source of gene
aliases. In DGIdb 4.0, the number of genes imported from Entrez has increased from 41,102 to 42,851 and Ensembl has been updated from version 90_38 to version 101_38. We have also migrated several older sources from our original domain-specific language (DSL) importers to the improved TSV importer style implemented in DGIdb 3.0.

In DGIdb 4.0, the database structure and presentation model was updated to allow sources to be imported with multiple source types. This change enables merging of sources that were previously duplicated for each independent claim type (drug, gene, interaction, druggable gene category). For example, we previously imported both interaction claims and druggable gene category claims from Guide to Pharmacology (Armstrong et al., 2020) with two separate importers which created two separate sources (GuideToPharmacologyInteractions and GuideToPharmacologyGenes, respectively). With this update, the import of interaction claims and druggable gene category claims is now handled by one importer and only a single source (GuideToPharmacology) is created. This is intended to simplify and unify claim sources to aid in downstream interpretation. Additionally, supporting sources that have multiple source types enables easy extension to collect more informative claim type information. For example, some claims from CIViC (Griffith et al., 2017) can be imported with the additional categories of drug resistance and clinically actionable. This change results in an additional 150 druggable gene category claims being imported from CIViC. Overall, these changes will increase the efficiency and accuracy of the process of importing and updating sources in DGIdb 4.0 compared to previous versions and will make it easier for users to evaluate individual sources.

### 3.4.3 Drug grouping improvements

Another notable change in the DGIdb 4.0 update is the improvement to drug grouping and normalization. Previously, we grouped drug claims using a rule-based pairwise association
approach. This process was cumbersome, requiring a lengthy and complete re-grouping of all claims whenever we updated sources in order to generate consistent groupings. We revised this approach by creating a normalization component independent of the claims aggregated by DGIdb, that could be run on a per-source basis. When redesigning this part of DGIdb, we took steps to enable reuse of this normalizer as a modular component for other resources. To this end, we leveraged and contributed to a drug normalization service from the Variant Interpretation for Cancer Consortium (the ‘thera-py’ Python Package; source code online at https://github.com/cancervariants/therapy-normalization). Among our contributions to thera-py was a normalizer for the Wikidata (Vrandečić and Krötzsch, 2014) resource, further enabling community contributions to assist in concept normalization both for DGIdb and other resources reliant upon the VICC normalization services.

Drug claims from DGIdb were normalized using the ChEMBL and Wikidata normalizers from thera-py. Rules were written to formalize grouper behavior based upon match characteristics of a query. Briefly, these rules prioritize matches to primary labels over aliases, exact case over case-insensitive, and ChEMBL over other normalizers. An algorithm for constructing a merged drug concept from normalizer results was specified, enabling a standardized set of aliases for a given concept identifier. Pseudocode for this algorithm is provided (see Supplementary Data), and all implemented code is available on our public repository (see Data Availability).

### 3.4.4 New Query Score and updated Interaction Score

One of the main features added in DGIdb 4.0 is the concept of a relative Query Score for interaction search results. Previously, interaction search results displayed only a static Interaction Score based on evidence of an interaction (i.e. the number of publications and sources supporting
an interaction claim). This Interaction Score did not take into account whether the gene and drug
involved in a given interaction were also part of a large number of other interactions and, thus,
had a low specificity that should be penalized. In addition, when searching for a set of genes or
drugs, the Interaction Score does not prioritize results with overlapping interacting drugs or
genes, which might be of more interest to the user, particularly in drug discovery and pathway
applications.

DGIdb 4.0 now provides a Query Score that is relative to the search set and considers the
overlap of interactions in the result set. For interaction searches using a gene list, the Query
Score is calculated from the Evidence Scores (publications and sources), the number of genes
from the search set that interact with the given drug, and the degree to which the drug has known
interactions with other genes (Figure 3.3). Similarly, for interaction searches using a drug list, the
Query Score depends on the Evidence Scores (publications and sources), the number of drugs
from the search set that interact with the given gene, and the degree to which the gene has known
interactions with other drugs (Figure 3.3). In effect, this means that genes and drugs with many
overlapping interactions in the search set will rank more highly, with the caveat that drugs or
genes involved in many interactions, in general, will have lowered scores (Figure 3.3).
Figure 3.3 Overview of DGIdb’s new Query scores and Interaction scores. (A) Schematic of how each of the new scores is calculated within DGIdb. Gene and drug queries both return a Query Score that is dependent on the search terms. Each interaction has an Interaction Score that is calculated independently of other search terms. (B) Example of a Query Score changing based on the terms searched. In the first panel, only MEK1 and MEK2 were searched and the Query Score for the interaction between MEK1 and Cobimetinib was 8.09. In the second panel, BRAF and KRAS were added to the search query. These both interact with Cobimetinib and thus raise the Query Score to 16.18. (C) Example of Interaction Score. The panel on the left shows the interaction between ZEB1 and Salinomycin. This is the only interaction for Salinomycin and thus it has a high Interaction Score. The panel on the right shows the interaction between ZEB1 and Doxorubicin. Doxorubicin is involved in 103 interactions within DGIdb and thus has a much lower Interaction Score. Note that over time, as sources are updated and new claims are added, both Query Scores and Interaction Scores may change.
Our static Interaction Score previously introduced in DGIdb 3.0 (Cotto et al., 2017) has been adjusted in DGIdb 4.0. The Interaction Score now mirrors the Query Score, except it is unaffected by the queried gene or drug sets, instead relying only on Evidence Scores and the degree to which both the gene and drug are involved in other interactions (Figure 3). Interaction Scores follow a long-tail distribution, indicative of many highly promiscuous drugs and genes, and relatively few well-supported, highly specific drug-gene interactions (Haupt et al., 2013) (Supplementary Figure S1).

The introduction of the relative Query Score provides users with a score that gives a more intuitive ranking of drugs or genes based on the search set of interest, allowing the prioritization of drugs or genes that have overlapping, specific interactions with the search set. Similarly, the improvements to the static Interaction Score provide a more nuanced scoring system that takes into consideration the number of interactions for a drug-gene pair, in addition to the previous Evidence Scores, giving a more informative static Interaction Score.

As sources are updated and additional interaction claims are added to DGIdb, Interaction Scores and Query Scores are subject to change as a result of the changing measure to which Drug and Gene concepts interact with one another. Query Scores are always variable, dependent upon the set of genes or drugs searched.

3.4.5 Inclusion of interaction directionality

We have also added information on the directionality of interaction types to the interaction search results. Each interaction type in DGIdb now has an indicated directionality of activating, meaning the interaction type mechanism has an overall activating effect; inhibitory, meaning the interaction type mechanism has an overall inhibitory effect; or n/a, meaning the directionality is unclear for the interaction type mechanism (Supplementary Table S2). Determination of the
directionality for an interaction type was made from mechanistic definitions provided by drug–gene interaction sources in which the interaction type was observed. Where these definitions were not available, we instead relied upon community definitions of these interaction types. These interaction directionalities are included on the UI for interaction search results in parentheses next to the interaction type(s) listed for each interaction result (Supplementary Figure S2). While the directionality may be obvious for some interaction types (e.g. activators are activating), some interaction types are not immediately apparent (e.g. chaperones are activating) to those less familiar with mechanisms of drug–gene interactions. Inclusion of directionality can make it easier for users to distinguish interactions that are more relevant for their purposes. For example, a user interested in exploring drugs that inhibit a particular gene will look for drug-gene interactions with inhibitory directionality. Users are also able to limit their interaction search to only interaction types of a desired directionality. Detailed information on each interaction type, the definition of each interaction type, and the directionality of each interaction are also available on the DGIdb Interaction Types page (https://dgidb.org/interaction_types).

3.4.6 Search feature improvements

In DGIdb 4.0, we have introduced several updates related to searching, search results, and information available on the user interface. Among these are the addition of the option to search only cancer-specific sources (meaning sources that report claims relating to cancer only), or disease-agnostic sources (meaning sources that report claims relating to any disease, including cancer) for both interaction searches and druggable gene category searches. Cancer-related searches are a major use-case for DGIdb and cancer-specific sources are well-represented among all sources, with 13 cancer-specific drug-gene interaction sources and five cancer-specific
druggable gene category sources. However, DGIdb is not a cancer-specific resource and is intended to be utilized for non-cancer related research as well. For drug-gene interactions, there are 4,955 interactions supported by cancer-specific sources only, 48,341 interactions supported by disease-agnostic sources only, and 1,295 interactions are supported by both cancer-specific and disease-agnostic sources. Similarly, for druggable gene categories, 233 genes have categories supported by cancer-specific sources only, 17,168 genes have categories supported by disease-agnostic sources only and 2804 genes have categories supported by both cancer-specific and disease-agnostic sources. These numbers show that although a sizable portion of the sources included in DGIdb are cancer-specific, those types of sources only represent a small proportion of the overall data.

Other improvements to search result features include the addition of linkouts to specific interaction evidence, where available. These will allow users to browse to the primary source for an interaction claim which might provide additional information and context not captured by DGIdb. Also, while we introduced drug and gene filters to the interaction search view in the last major update, we have had several requests to define how these filters are implemented. To address this lack of transparency on the UI, we have now added a link to the FAQ page where these filters are now defined.

### 3.4.7 Monthly data releases

DGIdb 3.0 implemented online updaters that imported data from dynamic sources (such as CIViC (Griffith et al., 2017), Guide to Pharmacology (Armstrong et al., 2020), OncoKB (Chakravarty et al., 2017), etc.) periodically, usually monthly. As a result, the static TSV data releases available on our Downloads page would quickly become outdated. For DGIdb 4.0, we have implemented monthly data releases of these TSVs to coincide with monthly runs of the
online updaters, to ensure that TSVs available for download reflect the most up to date information in our database. The Downloads page now makes available the current Gene, Drug, Interaction and Category TSVs as well as previous monthly TSVs since the release of DGIdb 4.0. These serve as de facto snapshots of the data in DGIdb over time.

3.4.8 Improved transparency and details on licensing of sources

In DGIdb 4.0, we have made a significant effort to update and improve the information we provide on licensing of sources imported into our database through manual curation of data license descriptions and references for every source. This information is now readily available on the sources page. Since DGIdb 3.0, several existing sources have made changes to their licensing, making data from some sources more broadly available and data from other sources more restricted. Notably, PharmGKB (Whirl-Carrillo et al., 2012) has moved to a more permissive Creative Commons Attribution-ShareAlike 4.0 International License and DrugBank (Wishart et al., 2018) has adopted a custom non-commercial license. In contrast, OncoKB (Chakravarty et al., 2017) has restricted API access to registered/approved non-commercial research use only, and JAX-CKB (Patterson et al., 2016) has restricted API access to negotiated licenses only. Both resources continue to provide access to a portion of their data for free through their respective web clients.

3.4.9 Application framework updates

To handle increased web traffic and integration with other tools, we have upgraded DGIdb to Rails 6 (from Rails 5), upgraded to Ruby 2.6.5 (from Ruby 2.3), upgraded to PostgreSQL 12 (from PostgreSQL 9.6), and upgraded the server to the latest Ubuntu LTS release (20.04). In
addition to the new features and performance benefits these upgrades bring, they will ensure that we continue to remain on supported software versions that receive regular security updates.

In order to keep DGIdb's underlying source data current, we had previously implemented an automated job scheduling framework using DelayedJob to schedule monthly runs of online updaters. In this release, we switched to using Sidekiq. In contrast to DelayedJob, Sidekiq offers a convenient user interface which makes identifying job failure reasons and rescheduling of failed jobs easier. Furthermore, the addition of Airbrake (https://airbrake.io/), an online tool for exception tracking, gives error reviews and notifies the development team of these errors in real-time (for instance, via email).

To ease future implementation of fixes and new features, we moved testing to a GitHub continuous integration (CI) workflow which allows us to continuously test newly committed code for errors against multiple versions of Ruby and PostgreSQL.

3.5 Conclusions and Future Directions

With our most recent release, DGIdb has received significant improvements to source content, functionality such as searching and grouping, and underlying application technology. We have significantly expanded the number of records in our database through the addition of new sources and updates of existing sources. Furthermore, we have improved our ability to maintain regular content updates through the implementation of additional online importers for several sources and the use of Sidekiq for automatic job processing. We have revised our process for drug grouping and normalization to be batched by resource and to leverage continual improvement through community contributions to the VICC thera-py normalizers and the Wikidata public-domain crowdsourcing platform. Finally, several updates have been made to inform users of the relevance of search results through information presented on the UI. We have
implemented more sophisticated notations of relative and static interaction scores, improved the relevance of interaction source linkouts wherever possible, and included the concept of directionality for interaction results.

Although the updates in DGIdb 4.0 have improved the usability and content of our resource, we expect there will still be a need for future improvements. One technology improvement on our roadmap is converging the public-facing API with the internal code that powers the web views. Ultimately, we want the APIs available for general use to be the same ones powering our HTML pages. This would provide an even more fully featured API to end users while reducing our overall maintenance burden by eliminating redundant code. We are also evaluating the addition of information on gene-gene relationships. As always, we plan to continue updating sources to online updaters where possible, and migrating TSV-based sources from the legacy DSL importers to the TSV importers introduced in DGIdb 3.0.

3.6 Data Availability
DGIdb is an open access database and web interface (www.dgidb.org) with open source code available on GitHub (https://github.com/griffithlab/dgi-db) under the MIT license. We also provide data downloads for drug claims, gene claims, and interaction claims on the website in addition to a SQL data dump (http://dgidb.org/downloads). Information about the API and its endpoints can also be found on the website (http://dgidb.org/api).

3.7 Supplementary Data
Supplementary Data are available at NAR Online. (This includes supplementary tables and figures).
3.8 Acknowledgments

We want to thank the creators and maintainers of the seven new resources added to DGIdb and the many previously incorporated resources, as well as our growing community of users for notifying us of minor and major issues and for their suggestions on new features and improvements to DGIdb. We would also like to thank the members of the NDEX team, and in particular Dexter Pratt and Rudolf Pillich for their efforts in integrating DGIdb data into the NDEX resource.

3.9 Funding

National Human Genome Research Institute [K99HG010157 to A.H.W., A.C.C., R00HG007940 to M.G.]; National Cancer Institute [U01CA209936, U01CA248235, U24CA237719 to M.G., O.L.G., S.K., A.C.C., A.H.W., J.F.M.]. Funding for open access charge: departmental funding.
Chapter 4. ORegAnno 3.0: a community-driven resource for curated regulatory annotation

4.1 Preamble
The following chapter has been published:


As an author of the published manuscript, and in compliance with the editorial policies at Nucleic Acids Research, the cited publication is included in full in the following chapter. I would like to elucidate my role in this manuscript. I was responsible for curating new records for inclusion within ORegAnno 3.0. Additionally, I assisted the lead author with manuscript writing and editing as well as design of the custom capture reagent.

4.2 Summary
The Open Regulatory Annotation database (ORegAnno) is a resource for curated regulatory annotation. It contains information about regulatory regions, transcription factor binding sites, RNA binding sites, regulatory variants, haplotypes, and other regulatory elements. ORegAnno differentiates itself from other regulatory resources by facilitating crowd-sourced interpretation and annotation of regulatory observations from the literature and highly curated resources. It contains a comprehensive annotation scheme that aims to describe both the elements and outcomes of regulatory events. Moreover, ORegAnno assembles these disparate data sources and annotations into a single, high quality catalogue of curated regulatory information. The current
release is an update of the database previously featured in the NAR Database Issue, and now contains 1,948,307 records, across 18 species, with a combined coverage of 334,215,080 bp. Complete records, annotation, and other associated data are available for browsing and download at http://www.oreganno.org/.

### 4.3 Introduction

The Open Regulatory Annotation database (ORegAnno) was first released about a decade ago (Montgomery et al., 2006), with the intention to collect and synthesize a catalogue of regulatory elements. It remains unique in the field because of its focus on collecting high quality, curated regulatory records from the literature. Moreover, ORegAnno relies on a thriving community of scientists who are interested in contributing to the resource, as well as utilizing its data. Since the last release of ORegAnno in early 2008 (Griffith et al., 2008), the amount and types of published regulatory data have grown exponentially. This relates in part to high-throughput studies from the ENCODE consortium and others, who have performed an enormous number of ChIP-seq, DNase-seq, FAIRE-seq and other experiments aiming to identify biochemically available and transcriptionally active regions of genomes (ENCODE Project Consortium, 2012). While these efforts are excellent resources for identifying candidate regulatory regions, ENCODE efforts have suggested that as much as 80% of the genome could be functional (ENCODE Project Consortium, 2012). This controversial finding has been the focus of much attention in the community, with several commentaries pointing out that these types of high-throughput data are prone to overestimates due to experimental and statistical methods that result in a high number of false positive calls (Doolittle, 2013; Eddy, 2013; Graur et al., 2013). Moreover, they do not necessarily provide a comprehensive understanding of all of the elements involved in gene regulation. For example, knowing the region of DNA that is bound by a transcription factor does
not directly indicate whether the expression of any genes are altered, nor whether an alteration results in up- versus down-regulation. Validation of the genomic regions identified by ENCODE and others requires a large number of low-throughput experimental data paired with careful manual curation. Additionally, much of the available evidence supporting gene regulation is dispersed across various experiments, specialized datasets, and individual publications, making it cumbersome to obtain regulatory information that has been released by the community across this broad set of sources. The current version of ORegAnno seeks to address these issues by cataloging a large number of new, curated, high quality regulatory records that are derived from published literature and other data resources.

4.4 Results

4.4.1 Overview

The current version of ORegAnno now has a total of 1,948,307 unique records. These records cover a combined 334,215,080 bp across 18 species (Figure 4.1A and B). The vast majority of these records are mapped to human and mouse genomes, with 1,452,466 records in human (261,660,516 bp in the GRCh38/hg38 genome assembly version) and 415,808 records in mouse (57,253,973 bp in the GRCm38/mm10 genome assembly version).
Figure 4.1 Current content of the ORegAnno database. Content statistics are divided by species (A and B), regulatory type (C and D), and data source (E and F).
As a measure of the success of our community-based participation, ORegAnno currently has 1,044 registered users. Aside from the principal authors of this paper, 13,301 records have been contributed by members of the broader community (The Open Regulatory Annotation Consortium). ORegAnno continues to have a robust verification system to ensure that contributed records are accurate and appropriately annotated. A set of trusted consortium members have been granted a ‘validator’ status, allowing them to review and up- or down-vote records. This results in individual record scores that are visible to all users. Moreover, when a record is negatively scored, it will typically be assigned a deprecated status. ORegAnno additionally includes an ontology for summarizing the experimental evidence that supports the regulatory elements and outcome in each record. Together, these features allow users to filter records according to various quality criteria.

The ORegAnno database has served as a repository for publishing regulatory sites derived from experimental data (Wederell et al., 2008), and it has been incorporated into other resources including the Babelomics (Medina et al., 2010), cisRED (Sleumer et al., 2009), ConTra (Hooghe et al., 2008), GRASP (Eicher et al., 2015), i-cisTarget (Imrichová et al., 2015), LASAGNA-Search 2.0 (Lee and Huang, 2014), the UCSC Genome Browser (Rosenbloom et al., 2015) and more. Similarly, the annotated information included in ORegAnno has been used to construct gene regulation networks for the development of other tools and the analysis of gene expression data (Baitaluk et al., 2012; Chu et al., 2014; Fazekas et al., 2013; Komurov et al., 2010; Türei et al., 2015). ORegAnno records were used in the REC-set design for a capture sequence reagent (Bainbridge et al., 2011), and as part of the definition for regulatory sites of the human genome (tier 2) in the Genome Modeling System (Griffith et al., 2015a), an analysis information management system at the McDonnell Genome Institute of Washington University.
that has been used to process over 4800 human whole genome samples, over 40,000 exomes, and over 1400 transcriptomes. Similarly, ORegAnno has been adapted into the information systems of other research centers including the Broad Institute and Cancer Research UK, where it has been used in the analysis of several high impact studies (Baca et al., 2013; Li et al., 2013; Stransky et al., 2011; Turnbull et al., 2012).

Because ORegAnno focuses on curated regulatory information, the total genomic coverage found in ORegAnno is smaller than that identified by resources such as ENCODE or the ENSEMBL regulatory tracks (Zerbino et al., 2015), which are largely a summary of ENCODE data (Figure 4.2). This trade off is part of an effort to ensure that ORegAnno represents a high-quality curated set of regulatory elements, with the aim of maintaining a low number of false positive records.

Figure 4.2 Comparison of the genomic coverage captured by ORegAnno and the ENSEMBL Regulatory Track. A Venn diagram demonstrates coverage overlaps for human genome assembly version GRCh38/hg38, with sets sized to scale. The ENSEMBL Regulatory Track is divided into two main sets, a track overview set and the transcription factor binding site (TFBS) set.
4.4.2 Updates

Older records, including those that were added through crowd-sourcing efforts via the web, have been updated to ensure that only accurate and up-to-date gene symbols are being used. This was accomplished through a combination of automatically updating symbols using NCBI Gene or ENSEMBL identifiers, as well as by manually checking incorrect and missing data. In addition, previously missing identifiers from NCBI Gene or ENSEMBL have been added where possible, allowing for future automated updates to ensure the accuracy of these gene lists. These updates have resulted in 423 automated changes and 13,174 manually curated changes (13,597 total) affecting 10,386 records.

For all ORegAnno records (existing and new), genomic coordinates have been updated and expanded using liftOver (Kent et al., 2003). This involved converting older genomic coordinates to newer assembly versions, as well as converting coordinates from new versions to older assemblies. Thus, each record may now be associated with multiple genomic coordinates (from multiple assembly versions). For example, since the last version of ORegAnno was published in 2008, the human genome assembly version GRCh38/hg38 was released. All existing ORegAnno human records having genomic coordinates based on assembly versions GRCh36/hg18 or GRCh37/hg19 now have additional updated coordinates using GRCh38/hg38. Similarly, new records that were entered using GRCh38/hg38 coordinates have received additional coordinates based on GRCh37/hg19 and GRCh36/hg18. This allows users to access the genomic coordinates of regulatory regions for the assembly versions that best suit their purposes. Finally, new types of transcriptional regulation have been defined in the current release (Figure 4.1C and D). These includes microRNA and small non-coding RNA binding sites, as well as operons that function to regulate multiple genes under a single promoter.
4.4.3 New Records

ORegAnno has maintained a focus on incorporating records derived from high quality, manually curated evidence for gene regulation. These typically include experimental evidence demonstrating that binding of a regulatory element to a specific region of DNA or RNA alters corresponding gene expression levels. In total, the current release of ORegAnno contains 2010 unique records covering 112,582 bp derived directly through literature curation, including 661 records that have been added since the previous ORegAnno release.

Highly validated external databases that had been incorporated into earlier ORegAnno releases have been updated. This includes 1874 new records covering an additional 3,591,656 bp derived from VISTA Enhancers (Visel et al., 2007) (2196 total records covering 3,996,796 total bp), 2934 new records covering an additional 863,201 bp derived from the Yeast Regulatory Map (MacIsaac et al., 2006) (7320 total records covering 899,449 total bp), as well as 2051 new transcription factor binding site records covering an additional 29,405 bp derived from REDfly (Gallo et al., 2011) (2695 total records covering 913,486 total bp). Previously, ORegAnno had imported records from FlyReg (Bergman et al., 2005), which has since been merged into REDfly.

New records have been created by importing data from external databases that were not found in previous ORegAnno releases. This includes 1,093,443 records covering 11,780,604 bp imported from the JASPAR CORE database (Mathelier et al., 2014), which contains a curated, non-redundant set of experimentally obtained transcription factor binding sites in eukaryotes. 783,742 records covering 300,003,052 bp were imported from the PAZAR database (Portales-Casamar et al., 2009), which included only records with curated evidence of transcription factor binding and regulatory sequence annotation across various species. 11,451 records covering
4,194,677 bp were derived from RegulonDB (Salgado et al., 2013), a database of transcriptional regulation in Escherichia coli K-12, and includes manually curated records that have been complemented with high throughput datasets and comprehensive computational predictions. We combined conserved miRNA target site predictions from miRanda-mirSVR (Betel et al., 2010) with experimentally-validated miRNA-target interaction data from miRTarBase (Hsu et al., 2014), leading to the addition of 3,072 new ORegAnno records covering 44,353 bp. 131 records covering 1216 bp were derived from NFI-RegulomeDB (Gronostajski et al., 2011), a database with curated binding sites for the NFI (Nuclear Factor I) family of transcription factors using data from the published literature. Finally, 51 transcription factor binding site records covering 7503 bp were created from the PCNE database of phylogenetically conserved noncoding elements (Hufton et al., 2009).

Because of the open and accessible design of the ORegAnno database and website, ORegAnno has been used for submitting published experimental data. Since the previous ORegAnno release, four datasets derived from high throughput studies have been submitted, and were subsequently curated to ensure that only regulatory regions with a high degree of evidence were retained. These include RELA (p65) ChIP−PET binding sites in human monocytes (Lim et al., 2007) (489 records covering 52,886 bp), ESR1 binding sites in human MCF-7 breast cancer cells (Lin et al., 2007) (1234 records covering 165,538 bp), Esr1 binding sites in mouse liver (Gao et al., 2008) (5568 records covering 2,378,460 bp), and Foxa2 binding sites in mouse liver (Wederell et al., 2008) (11 475 records covering 8,236,933 bp). In all of these cases, DNA sequences were filtered according to signal strength and proximity to signal peak to reduce false positive calls. A summary of the number of records and genomic coverage contributed by each data source is shown in Figure 4.1E, F and Supplementary Table S1.
4.4.4 Data access

The ORegAnno database continues to be accessible under an open-source license (GNU Lesser General Public License), in order to encourage development and participation from the community. Monthly ORegAnno database summaries are automatically performed and provide fundamental regulatory information from ORegAnno in a tab-delimited text file that is available for free download, without the need to register with the ORegAnno website (http://www.oreganno.org/). The ORegAnno website back end code has been updated to improve security and performance, and to accommodate the new data types, dataset sources, and the increased number of records that have been added since the previous release. New search functionality has been added, including the ability to browse records by transcription factor/regulatory element of interest. Source code for the ORegAnno website is available at https://java.net/projects/oreganno/.

The regulatory regions and associated annotation for all supported species have been submitted to the UCSC Genome Browser (Rosenbloom et al., 2015) as updates to existing ORegAnno tracks. This updates existing tracks with a more comprehensive collection of putatively regulatory elements, and additionally provides new tracks on several genome assembly versions.

4.4.5 Applications

Recently, there has been immense focus on the role of regulatory regions in cancer. In particular, recurrent somatic mutations in the TERT promoter have been identified in various cancer types (Horn et al., 2013; Huang et al., 2013; Killela et al., 2013; Vinagre et al., 2013), and are
associated with increased expression of TERT. Although the importance of TERT up-regulation in cancer has been well-established for nearly two decades (Zhang et al., 1999), it is only in recent years that we've identified the regulatory mechanism driving TERT up-regulation in such cases. While additional efforts have identified a small number of other recurrent regulatory mutations in cancer (Fredriksson et al., 2014; Melton et al., 2015; Weinhold et al., 2014), this number is far smaller than the recurrent protein-coding mutations that have been identified. This is likely due to several factors, including that most cancer survey projects have focused primarily on coding regions by using exome capture reagents to enrich for these regions, and that the TERT promoter region, as with many other genes, has a high GC content making both PCR amplification and sequencing challenging.

Previous identification of coding regions of the genome made it possible to perform exome targeted sequencing of these regions in a large number of cancer cases at a relatively low cost. Similarly, we've used ORegAnno and other sources to design a ‘regulome’ capture reagent for targeted sequencing. The high quality, relatively small coverage of literature-curated transcription factor binding sites, regulatory polymorphisms, and NFI-RegulomeDB (Gronostajski et al., 2011) sites identified in ORegAnno, in conjunction with regulatory regions defined by FunSeq (Khurana et al., 2013), and 500 bp regions upstream of each gene transcription start site, were used to define the ‘regulome’ region. As a proof of principle, we then applied ‘regulome’ capture-sequencing to ten normal/tumor pairs of hepatocellular carcinoma (HCC). Overall coverage of the regulatory region defined in the capture reagent was higher in whole regulome sequencing (WRS) samples versus whole genome sequencing (WGS) samples of the same tissues, with median average read depths of 29× in WGS normal, 49× in WGS tumor, 60× in WRS normal and 68× in WRS tumor (Figure 4.3A, Supplementary Table
S2). This improved coverage allowed us to reliably identify the canonical somatic TERT promoter mutation C228T in six of the ten cases, an illustrative example of which is shown in Figure 4.3B.
Figure 4.3 Capture reagent using ORegAnno sites improves coverage of regulatory regions in human hepatocellular carcinoma. (A) Coverage across the entire 'regulome' is visualized as a heatmap for each of the ten HCC cases. WRS samples have greater sequencing read depth across the targeted region, compared with WGS samples. (B) An illustrative IGV (Robinson et al., 2011) screenshot is shown of the TERT promoter for one HCC case. A canonical C228T somatic mutation is observed in the WRS data, but cannot be reliably called in the WGS data.
4.5 Acknowledgements

We thank the Open Regulatory Annotation Consortium for their continuing contributions to ORegAnno through the identification of relevant publications and manual curation efforts. We additionally thank the publishers of experimental data and external regulatory databases that were included in the current release of ORegAnno.

4.6 Supplementary Data

Supplementary Data are available at NAR Online.

4.7 Funding

American Cancer Society [IRG-58-010-58- to O.L.G.]; Edward Mallinckrodt, Jr. Foundation (to S.B.M.); National Institutes of Health [K99HG007940 to M.G., R01HG008150 and R01MH101814 to S.B.M., K22CA188163 to O.L.G.]. Funding for open access charge: Departmental funding (to O.L.G.).

Conflict of interest statement. None declared.
Chapter 5: Identification of recurrent regulatory mutations in breast cancer

5.1 Preamble
This work has not yet been submitted. However, the following contents are in the process of being assembled for journal submission. The planned author list is as follows:
Cotto KC, Danos, AM, Lesurf R, Park M, Hunt K, Keyomarsi K, Mardis ER, Griffith M, Griffith OL.

5.2 Summary
Many studies have attempted pan-cancer analyses in order to identify recurrent coding mutations, but few have explored non-coding space, and often the results do not extend to breast cancer. While some breast cancer specific studies have identified significantly mutated promoters, they have often failed to incorporate transcriptome data to assess the impact and relevance of mutations on gene expression within tumors. In order to address this issue, we have compiled a data set consisting of 458 breast cancer tumor/normal pairs. Mutations have been identified with whole genome, exome, or custom capture sequencing and expression with transcriptome sequencing or microarray. The custom (regulome) capture reagent covers regions assembled from regulatory databases, 5’ untranslated regions, 500 bases upstream and downstream of transcription start sites, and 50,000 bases upstream and downstream of 178 genes that have been implicated as being important in breast cancer. While this custom capture region is similar in size to an exome, it has advantages over WGS and exome sequencing. It provides deep sequence coverage of regulatory regions, including GC-rich promoter regions, where WGS often produces insufficient coverage. After extensive filtering, our analysis revealed significant
mutation clustering within the noncoding space of \textit{RMRP} and \textit{WDR74}, as has been noted in previous studies, as well as \textasciitilde130 other genes not previously reported. The significance of these mutations will be presented based on recurrence, transcriptome changes, and molecular subtype relationships.

\section*{5.3 Introduction}

Breast cancer is the most commonly diagnosed cancer among women, with an estimated 1.7 million new cases and 500,000 deaths due to it each year (Cancer Genome Atlas Network, 2012; Torre et al., 2015). The prognosis for this disease varies depending on the stage of the disease at the time of detection. Early stages of breast cancer (stages 0-II) usually have greater than 90\% relative survival at five years, while stages III and IV have a significantly decreased relative survival, 72\% and 22\%, respectively. Beyond staging, prognosis and treatment can vary depending on the molecular features of the tumor. Microarray analysis in the early 2000s was used to define the concept of molecular subtypes within breast cancer and resulted in the definition of five distinct subtypes: luminal A, luminal B, HER2 enriched, triple-negative/basal-like, and normal-like (Perou et al., 2000; Sorlie et al., 2003). While these subtypes are based on molecular expression signatures, they are often thought about with respect to their expression of the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) (Tofigh et al., 2014). The different subtypes vary significantly with respect to gene expression, tumor presentation, and prognosis (Dai et al., 2015). The genomic mutational landscape of coding regions within these breast cancer subtypes has been thoroughly described (Cancer Genome Atlas Network, 2012; Ding et al., 2010b; Shah et al., 2009, 2012). These studies have stratified subtypes by mutational rate within coding regions, copy number changes, and recurrently mutated genes. While the aforementioned studies focused on coding mutations,
there has been increased interest in better understanding the role of noncoding variation in breast cancer. This interest partially derives from the initial discovery of recurrent TERT promoter mutations resulting in increased TERT expression in melanoma (Horn et al., 2013; Huang et al., 2013). Several studies have attempted pan-cancer analyses in order to identify these types of mutations, but often the results suffer from low sequence coverage of regulatory regions or do not extend to breast cancer (Fredriksson et al., 2014). In attempts to define the non-coding genomic space of breast cancer, studies have begun to systematically investigate the existence of recurrent regulatory somatic mutations by genome-wide analysis in breast cancer. These studies have identified some genes, such as FOXA1, TBC1D12, and NEAT1, as having recurrent mutations within their promoter regions (Nik-Zainal et al., 2016; Rheinbay et al., 2017; Weinhold et al., 2014). While some breast cancer-specific studies have identified some significantly mutated promoters and lncRNAs, they have often failed to incorporate transcriptome data to assess the impact and relevance of mutations on the expression and RNA splicing of genes within these tumors (Nik-Zainal et al., 2016; Rheinbay et al., 2017; Weinhold et al., 2014). Additionally, while these studies have been able to identify recurrent mutations in large cohorts, their analyses fail to describe subtype-specific recurrent, noncoding mutations and their consequences.

We have assembled and generated a data set consisting of 458 breast cancer cases with matched tumor/normal pairs. This data set contains samples from The Cancer Genome Atlas (TCGA), and collaborations with the American College of Surgeons Oncology Group (ACOSOG), MD Anderson Cancer Center, and McGill University. Based on clinically available information, this cohort consists of 22.4% luminal A, 19% luminal B, 16.4% HER2-enriched,
21% basal-like, 0.8% normal-like, and 20.4% unknown with regards to molecular subtype. Having this mix of breast cancer subtypes is important due to them having dissimilar phenotypes and varying rates of protein coding mutations. To improve sensitivity within difficult to sequence GC-rich promoter regions, we designed a custom capture reagent that covers the ‘regulome’, areas of the genome that we believe contain important regulatory sequences (Lesurf et al., 2016). This regulome consists of regions assembled from regulatory databases, 5’ untranslated regions, 500 bases upstream and downstream of transcription start sites, and 50,000 bases upstream and downstream of 178 genes that have been implicated as being important in breast cancer (Bahcall, 2013; Bainbridge et al., 2011; Betel et al., 2010; Dreos et al., 2015; Griffiths-Jones, 2006; Gronostajski et al., 2011; Hsu et al., 2014; Lesurf et al., 2016; Visel et al., 2007). This custom capture region is similar to an exome capture reagent in terms of the genomic space that it seeks to cover, 61 megabase pairs (Mb) vs 39-64 Mb, respectively. However, it has advantages over whole genome sequencing (WGS) and whole exome sequencing (WES), particularly with respect to coverage in GC-rich promoter regions. We tested this custom capture sequencing reagent, hereafter referred to as performing ‘whole regulome sequencing’ (WRS), on ten normal/tumor pairs of hepatocellular carcinoma (HCC) and found that the overall coverage of the defined regions was higher by 20-30X when compared to WGS (Lesurf et al., 2016). This improved coverage with the custom capture reagent allowed us to identify the canonical somatic \textit{TERT} promoter mutation C228T in 6 of 10 cases; these same mutations were unable to be reliably called in the WGS data (Lesurf et al., 2016). While the other 4 samples had enough coverage to be able to call this mutation, there was no evidence of it.

Using the ‘regulome’ custom-capture reagent in addition to WGS, WES, RNA-Seq, and microarray data, we were able to efficiently study regulatory mutations within our 458 matched
tumor/normal breast cancer samples. Here we focus on non-coding mutations in promoter regions, defined as 400 base pairs (bp) upstream to 250 bp downstream of the annotated transcription start sites (TSS). Our analysis revealed significant mutation clustering within the noncoding space of \textit{RMRP, WDR74}, as noted in previous studies, as well as \textasciitilde130 other genes not previously reported.

5.4 Results

5.4.1 Molecular subtyping of samples using genomic data

While we did have some subtype information from clinical data for our cohort, we used a 50-gene predictor (PAM50) to classify each of our samples based on molecular subtype. Figures 5.1 and 5.2 show the subtype breakdown for the entire cohort and each individual cohort, respectively. We have higher numbers of both basal-like and HER2 enriched subtypes due in part to the purposeful curation of basal subtypes within this study and the ACOSOG study being a HER2 enriched subtype focused study.

![Subtype Classification for the entire cohort](image)

Figure 5.1 Breakdown of the number of samples classified as each molecular subtype for all 458 samples, irrespective of original cohort.
Figure 5.2 Breakdown of the number of samples classified as each molecular subtype split by original cohort.

5.4.2 Validation of the custom-capture reagent

While the custom-capture reagent was originally validated in the manuscript that describes its design, we wanted to confirm that using this additional method of sequencing allowed us to capture more of the regulatory regions of the genome that we were interested in analyzing within this study. To quantify how well the custom-capture reagent covered its regions of interest compared to WGS, we used samtools (Li et al., 2009) to collect coverage metrics for TCGA samples that had either WRS or WGS data (Figures 5.3 and 5.4).
Figure 5.3 Coverage data for each of the 154 TCGA samples with WRS data. Read depth data for the regions of interest from the designed custom-capture reagent with each WRS sample represented as an individual line.
Figure 5.4 Coverage data for each of the 116 TCGA samples with WGS data. *Read depth data for the regions of interest from the designed custom-capture reagent with each WGS sample represented as an individual line.*

### 5.4.3 Discovery of recurrently mutated promoter regions

Discovery of regions with an excess of mutations requires careful estimation of background mutation frequencies, which can be influenced by multiple genomic factors. For coding regions, established methods take into account patient-specific coverage information, patient-specific
overall mutation rate, genomic covariates of mutation rates, and clustering of mutations
(Alexandrov et al., 2013; Lawrence et al., 2013). For the identified promoter regions, the tool we
employed, MutEnricher, uses a similar strategy. Whereas patient-specific background mutation
rates in coding regions are estimated on the basis of silent coding and nearby non-coding
mutations, for non-coding regions we used all mutations, because it is unclear which are non-
functional. This conservative approach overestimates the background rate. MutEnricher
identified promoters with either an overall excess of mutations above expectation or an unusual
clustering of mutations. The latter may reflect events in specific transcription factor binding
sites, whose signal may otherwise be diluted when the larger promoter region is considered.

MutEnricher identified 139 recurrently mutated hotspot regions using the mutations from
the promoter regions that we defined (Appendix 2, Table 1). Some of these regions have been
previously described such as RMRP/CCDC107 and WDR74 (Rheinbay et al., 2017; Weinhold et
al., 2014). Specifically, within our own data, we identified 11 mutations in the non-coding space
of the bidirectional promoter region of RMRP and CCDC107 (Figure 5.3).
Figure 5.3 Lollipop plot of recurrent mutations in the bidirectional promoter of RMRP and CCDC107. 11 clustered mutations are shown in the shared promoter region of RMRP (blue) and CCDC107 (red).

Novel findings within our analysis include LINC01410, FRG1DP, HLA-DRB1, IGHV3-71, AZGP1, and many others. Some of these genes such as HLA-DRB1 and IGHV3-71 are involved in the immune system. Recurrent mutations within the promoter regions of such genes could affect how cancer specific antigens (neoantigens) are presented to the immune system for clearing. Other genes, such as AZGP1, have been associated with breast cancer but have not had non-coding regulatory mutations described. We did perform statistical analysis to determine
whether or not any of these recurrently mutated promoter regions were associated with a specific subtype but ultimately failed to find such correlations.

5.5 Discussion

We performed a comprehensive analysis of promoters in a large cohort of 458 patients with primary breast cancer and discovered 139 significantly mutated promoter regions. These regions show recurrent mutations at a specific base or at nearby bases – suggesting that they target-specific elements within the promoter (for example, transcription factor binding sites). Furthermore, these mutated promoter regions appear to be molecular subtype independent, suggesting that these mutations might drive breast cancer in a subtype agnostic manner.

Ultimately, appropriate clinical treatment will depend on the ability to recognize all functionally important mutations in each patient—including in regulatory elements, such as promoters. Identifying the targets of regulatory events will require systematic analysis of large cohorts of patients across cancer types. Promoter mutations may help explain activation or inactivation of known cancer genes in patients lacking coding mutations and may lead to discovery of new cancer genes. Completing our understanding of all alterations – coding and non-coding – in cancer genes will be an important foundation for cancer precision medicine.

5.6 Methods

5.6.1 Sample Procurement

We have assembled and generated a data set consisting of 458 breast cancer cases with matched tumor/normal pairs. This data set contains samples from The Cancer Genome Atlas (TCGA) (n=268), and collaborations with the American College of Surgeons Oncology Group
(ACOSOG) (n=42), MD Anderson Cancer Center (n=72), and McGill University (n=114). Based on clinically available information, this cohort consists of 22.4% luminal A, 19% luminal B, 16.4% HER2-enriched, 21% basal-like, 0.8% normal-like, and 20.4% unknown with regards to molecular subtype.

5.6.2 Sequencing

Genomic DNA was isolated by the Siteman Cancer Center Tissue Processing Core using the DNeasy Blood and Tissue Kit (Qiagen, Venlo, Netherlands). Library construction and sequencing were performed as previously described with a few exceptions described below (Griffith et al., 2015b). Single indexed libraries were constructed according to the manufacturer’s recommendations using the Illumina TruSeq Nano Kit (Illumina Inc, San Diego, CA) for whole genome sequencing (WGS) on the Illumina HiSeq X (2x150 bp reads). Genomic DNA was fragmented using the Covaris E210 DNA Sonicator (Covaris, WoBurn, MA). Dual indexed whole exome sequencing (WES) libraries were constructed/pooled according to the manufacturer’s recommendations using one of three kits/approaches: (1) the Paired-End Sample Prep Kit (Illumina Inc, San Diego, CA) for sequencing on the HiSeq 2500 (2x125 bp reads) (2) Kapa Auto Illumina (Kapa Biosystems, Woburn, MA) for sequencing on the HiSeq 2500 V4 1Tb (2x125 bp reads) and (3) Kapa Auto Illumina (Kapa Biosystems, Woburn, MA) for sequencing on the HiSeq 4000 (2x150 bp reads). Samples were pooled and captured using the NimbleGen SeqCap EZ Human Exome Library v3.0 Kit (Roche NimbleGen, Madison, WI). WRS was performed using the custom capture reagent described above. Total RNA was isolated by the Siteman Cancer Center Tissue Processing Core using Qiagen RNeasy kits. Single-indexed RNA sequencing (RNAseq) libraries were prepared using the Illumina TruSeq Stranded Total RNA kit with 500 ng of starting material according to the manufacturer’s recommendations. Sequencing
was performed on either the Illumina HiSeq 2500 V4 1 Tb platform (2x125 bp reads) or the Illumina HiSeq 4000 platform (2x150 bp reads).

Custom-capture sequencing, or “whole regulome sequencing (WRS)”, was performed on breast TCGA samples in which there was still tumor and normal tissue available (n=154) and on samples received through our collaboration with McGill University (n=114). Whole exome sequencing (WES) and RNA-sequencing data (RNA-Seq) was unarchived for TCGA samples with the additional WRS. We obtained unaligned sequencing data for TCGA breast cancer samples that had WGS, WES, and RNA-seq data available (n=116). The Griffith lab had established collaborations with the American College of Surgeons Oncology Group (n=42) and MD Anderson Cancer Center (n=72) for other breast cancer projects within the lab. The McGill University samples that underwent WRS also had pre-existing microarray data. Hence, the data set used for this project is a mix of WGS, WRS, WES, RNA-Seq, and microarray data.

5.6.3 Sequence alignment and somatic event detection

The Genome Modeling System (GMS) was used for all analysis, including the somatic variant detection and RNA-seq analysis (Griffith et al., 2015a). WGS, WES, and WRS data was processed through SpeedSeq v0.1.0 (Chiang et al., 2015; Faust and Hall, 2014), which aligns reads by BWA-MEM (Li and Durbin, 2009) to the human reference genome (NCBI build 38, GRCh38) and marks duplicates using SAMBLASTER v0.1.22 (Faust and Hall, 2014). RNA reads were aligned to GRCh38 with TopHat v2.0.8 (Kim et al., 2013; Trapnell et al., 2009). Somatic variants were called using several variant callers by comparing primary tumor to matched normal pairs. Single nucleotide variants (SNVs) were detected by taking the union of VarScan2 (Koboldt et al., 2009, 2012), Strelka (Kim et al., 2018; Saunders et al., 2012), Mutect (Benjamin et al., 2019; Cibulskis et al., 2013), and SomaticSniper (Larson et al., 2012). Small
insertions and deletions (indels) were detected by GATK Somatic Indel Detector (McKenna et al., 2010), VarScan2, Strelka, and Mutect. Variants were annotated by the GMS transcript variant annotator against Ensembl v74 and compared to the database of curated mutations (DoCM) (Ainscough et al., 2016) and COSMIC mutations (Sondka et al., 2018; Tate et al., 2019). All SNVs and indels were manually reviewed for removal of false positives according to standard procedures (Barnell et al., 2018). Somatic CNAs were detected by CopyCat v0.1 [https://github.com/chrisamiller/copyCat], and structural variations were predicted by Manta v0.29.6 (Chen et al., 2016).

5.6.4 PAM50 molecular subtype classification

Using each RNA-Seq data, the molecular subtype of each sample was determined using the Prediction Analysis of Microarray 50 (PAM50) (Nielsen et al., 2010; Paquet and Hallett, 2015; Perou et al., 2000; Pu et al., 2020; Sorlie et al., 2003).

5.6.5 Advanced filtering of somatic variants

After samples were processed with the Somatic Validation pipeline, any somatic variants or indels that were called underwent additional filtering. One method of filtering was using allele frequencies from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015). SNVs and indels were filtered against this dataset at a cutoff of 0.001. This means that any SNV or indel from this dataset that matches to the same genetic alteration within the 1000 Genomes dataset will be filtered out if it occurs at a rate of 0.001 or higher within the general population. Another method of filtering involved the use of gnomAD, a database containing mutational frequency data for 123,136 exome sequences and 15,496 whole-genome sequences at the time of us utilizing it (v1) (Lek et al., 2016). SNVs and indels were against gnomAD at a cutoff of 0.001.
Next, SNVs and indels were subjected to a panel of normal (PoN) filter. This involved evaluating the evidence present for each SNV or indel within all normal samples in order to identify if it is a possible sequencing or mapping artifact or previously uncharacterized germline variant that was not removed by previous filtering steps.

5.6.5 Subsetting variants to promoter regions

In order to focus analysis on regions that were most likely to have a functional, regulatory consequence, mutations were limited to a ‘promoter region’, defined as 400 base pairs (bp) upstream and 250 bp downstream of each annotated transcription start site (TSS) (Rheinbay et al., 2017). TSSs were pulled from the transcript annotations obtained from the Ensembl genome annotation database (Zerbino et al., 2018). Coding regions, miRNAs, and snoRNAs were removed from this analysis.

5.6.6 Identification of recurrently mutated promoter regions

Once mutations within promoter regions were defined, statistical analysis was performed in order to identify which regions harbor a significant recurrence of mutations among patients. This analysis was based on previously described methods implemented into a tool called MutEnricher (Rheinbay et al., 2017; Soltis et al., 2020). This method calculates a cumulative beta-binomial probability for each patient within each region which describes the probability of seeing at least one mutation in that region for that patient. Patient-specific probabilities are calculated in order to get the distribution of observing $x$ number of patients with at least one mutation within the promoter region. This distribution is then transformed into a $p$ value. Additionally, each promoter region with at least 3 mutations were evaluated to see if mutations within it are clustering more than expected by chance. The clustering analysis performed using a modification
of the CLUMPS algorithm, as described by Rheinbay et al. (Kamburov et al., 2015; Rheinbay et al., 2017). For regions with at least three mutations, the two $p$ values were be combined with Fisher’s method; otherwise, only the burden test $p$ value was used. The Benjamini-Hochberg FDR procedure was used to correct for multiple-hypothesis testing (Benjamini and Hochberg, 1995).
Chapter 6. RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer

6.1 Preamble
The following chapter is currently under review and has been posted as a preprint through bioRxiv (https://www.biorxiv.org/content/10.1101/436634v5). All supplemental files are available through the bioRxiv link:


* denotes co-first authors

I would like to distinguish my role in this manuscript from the other first-author in this study. Yang-Yang and I both oversaw the study, contributed to the code base, wrote the manuscript, and made figures for the manuscript. I analyzed the TCGA data, which contributed to most of the pan-cancer analysis. Additionally, I reworked the statistics script to work with scRNA data and tested RegTools on scRNA and long-read RNA-Seq data. I also oversaw the revisions for the last 2 rounds of review.

6.2 Summary
Somatic mutations in non-coding regions and even within exons may have unidentified regulatory consequences that are often overlooked in analysis workflows. Here we present
RegTools (www.regtools.org), a free, open-source software package designed to integrate analysis of somatic variants from genomic data with splice junctions from bulk or single cell transcriptomic data to identify variants that may cause aberrant splicing. RegTools was applied to over 9,000 tumor samples with both tumor DNA and RNA sequence data. We discovered 235,778 events where a variant significantly increased the splicing of a particular junction, across 158,200 unique variants and 131,212 unique junctions. To characterize these somatic variants and their associated splice isoforms, we annotated them with the Variant Effect Predictor (VEP), SpliceAI, and Genotype-Tissue Expression (GTEx) junction counts and compared our results to other tools that integrate genomic and transcriptomic data. While many events were corroborated by the aforementioned tools, the unbiased nature of RegTools allowed us to identify novel splice-associated variants and previously unreported patterns of splicing disruption in known cancer drivers, such as TP53, CDKN2A, and B2M, as well as in genes not previously considered cancer-relevant, such as RNF145.

6.3 Introduction

Alternative splicing of messenger RNA allows a single gene to encode multiple gene products, increasing a cell’s functional diversity and regulatory precision. However, splicing malfunction can lead to imbalances in transcriptional output or even the presence of novel oncogenic transcripts (Chabot and Shkreta, 2016). The interpretation of variants in cancer is frequently focused on direct protein-coding alterations (Vogelstein et al., 2013). However, most somatic mutations arise in intronic and intergenic regions, and exonic mutations may also have unidentified regulatory consequences (Jung et al., 2015; Soemedi et al., 2017; Supek et al., 2014; Venables et al., 2008). For example, mutations can affect splicing either in trans, by acting on
splicing effectors, or in cis, by altering the splicing signals located on the transcripts themselves (Climente-González et al., 2017).

Increasingly, we are identifying the importance of splice-associated variants in disease processes, including in cancer (Chen and Weiss, 2015; Xiong et al., 2015). However, our understanding of the landscape of these variants is currently limited, and few tools exist for their discovery. One approach to elucidating the role of splice-associated variants has been to predict the strength of putative splice sites in pre-mRNA from genomic sequences, such as the method used by the SpliceAI tool (Fairbrother et al., 2002; Jaganathan et al., 2019; Wang et al., 2004; Yeo and Burge, 2004). With the advent of efficient and affordable RNA-seq, we are also seeing the complementary approach of evaluating alternative splicing events (ASEs) directly from RNA sequencing data. Various tools exist which allow the identification of significant ASEs from transcript-level data within sample cohorts, including SUPPA2 and SPLADDER (Kahles et al., 2016; Trincado et al., 2018). Many of these tools have also evaluated the role of trans-acting splice-associated mutations (Kahles et al., 2018). However, few tools are directed at linking specific aberrant RNA splicing events to specific genomic variants in cis to investigate the splice regulatory impact of these variants. Those few relevant tools that do exist have significant limitations that preclude them from broad applications. The sQTL-based approach taken by LeafCutter (Li et al., 2018) and other tools (Li et al., 2016; Monlong et al., 2014) is designed for relatively frequent single-nucleotide polymorphisms. It is thus ill-suited to studying somatic variants, or any case in which the frequency of a particular variant is very low (often unique) in a given sample population. Recent tools that have been created for large-scale analysis of cancer-specific data, such as MiSplice and Veridical, ignore certain types of ASEs, are tailored to specific analysis strategies and hypotheses, or are otherwise inaccessible to the end-user due to
issues such as lack of documentation, difficulty with installation and integration with existing pipelines, limited computing efficiency, or license restrictions (Jayasinghe et al., 2018; Shirley et al., 2018; Viner et al., 2014). To address these needs, we have developed RegTools, a free, open-source (MIT license) software package that is well-documented, easy to install and use, and designed to efficiently and flexibly identify potential cis-acting splice-associated variants in tumors (www.regtools.org). At the highest level, RegTools contains three sub-modules: a variants module to annotate variant calls with respect to their potential splicing relevance, a junctions module to analyze aligned RNA-seq data and associated splicing events, and a cis-splice-effects module that integrates genomic variant calls and transcriptomic sequencing data to identify potential splice-associated variants. Each sub-module contains one or more commands, which can be used individually or integrated into regulatory variant analysis pipelines.

To demonstrate the utility of RegTools in identifying potential splice-associated variants from tumor data, we analyzed a combination of data available from the McDonnell Genome Institute (MGI) at Washington University School of Medicine and The Cancer Genome Atlas (TCGA) project. In total, we applied RegTools to 9,173 tumors across 35 cancer types. We contrasted our results with other tools that integrate genomic and transcriptomic data to identify potential splice-associated, specifically Veridical, MiSplice, and SAVNet (Jayasinghe et al., 2018; Shiraishi et al., 2018; Viner et al., 2014). Novel junctions identified by RegTools were compared to data from The Genotype-Tissue Expression (GTEx) project to assess whether these junctions are present in normal tissues (GTEx Consortium, 2013). Variants significantly associated with novel junctions were processed through VEP and Illumina’s SpliceAI tool to compare our findings with splicing consequences predicted based on the variant information alone (Jaganathan et al., 2019; McLaren et al., 2016). With this additional analysis, we were able
to more easily identify both variants in known cancer drivers, whose splicing consequences have not been previously reported in the literature, and potentially novel cancer drivers, whose disruption relies on splice-associated mutations.

6.4 Results

6.4.1 The RegTools tool suite supports splice-associated variant discovery by the integration of genome and transcriptome data

RegTools is a suite of tools designed to aid users in a broad range of splicing-related analyses. The variants module contains the annotate command. The variants annotate command takes a VCF of somatic variant calls and a GTF of transcript annotations as input. RegTools does not have any particular preference for variant callers or source of reference transcript annotations. Each variant is annotated by RegTools with known overlapping genes and transcripts, and is categorized into one of several user-configurable “variant types”, based on position relative to the edges of known exons. The variant type annotation depends on the stringency for splicing-relevance that the user sets with the “splice variant window” setting. By default, RegTools marks intronic variants within 2 bp of the exon edge as “splicing intronic”, exonic variants within 3 bp as “splicing exonic”, other intronic variants as “intronic”, and other exonic variants simply as “exonic.” RegTools considers only “splicing intronic” and “splicing exonic” as important. To allow for discovery of an arbitrarily expansive set of variants, RegTools allows the user to customize the size of the exonic/intronic windows individually (e.g. -i 50 -e 5 for intronic variants 50 bp from an exon edge and exonic variants 5 bp from an exon edge) or even consider all exonic/intronic variants as potentially splice-associated (e.g. -E or -I) (Figure 6.1A).
Figure 6.1 Flexible, streamlined discovery of cis-acting splice-associated variants with RegTools modules and cis-splice-effects identify workflow. A) By default, variants annotate marks variants within 3 bp on the exonic side and 2 bp on the intronic side of an exon edge as potentially splicing-relevant. This “splice variant window” can be modified individually for the exonic side and intronic side using the “-e” and “-i” options, respectively. With cis-splice-effects identify, for each variant in the splice variant window, a “splice junction region” is determined by finding the largest span of sequence space between exons which flank the exon associated with the splicing-relevant variant. The splice junction region can also be set manually to contain the entire sequence space n bases upstream and downstream of the variant using the “-w” option. Junctions overlapping the splice junction region are associated with the variant. Using the -E option considers all exonic variants as potentially splicing-relevant, but is otherwise the same. The -I option considers all intronic variants and also limits the splice junction region to the intronic region in which the variant is found, excluding the flanking exons.

B) Cis-splice-effects identify and the underlying junctions annotate command annotate splicing events based on whether the donor and acceptor site combination is found in the reference transcriptome GTF. In this example, there are two known transcripts (shown in blue) which overlap a set of junctions from RNAseq data (depicted as junction supporting reads in red). Comparing the observed junctions to the reference junctions in the first transcript (top panel), RegTools checks to see if the observed donor and acceptor splice sites are found in any of the reference exons and also counts the number of exons, acceptors, and donors skipped by a
particular junction. Double arrows represent matches between observed and reference acceptor/donor sites while single arrows show novel splice sites. These steps are repeated for the rest of the relevant transcripts, keeping track of whether there are known acceptor-donor combinations. Junctions with a known donor but novel acceptor or vice-versa are annotated as “D” or “A”, respectively. If both sites are known but do not appear in combination in any transcripts, the junction is annotated as “NDA”, whereas if both sites are unknown, the junction is annotated as “N”. If the junction is known to the reference GTF, it is marked as “DA”. C) The cis-splice-effects identify command relies on the variants annotate, junctions extract, and junctions annotate submodules. This pipeline takes variant calls and RNA-seq alignments along with genome and transcriptome references and outputs information about novel junctions and associated potential cis splice-associated sequence variants. RegTools is agnostic to downstream research goals and its output can be filtered through user-specific methods and thus can be applied to a broad set of scientific questions.

The junctions module contains the extract and annotate commands. The junctions extract command takes an alignment file containing aligned RNA-seq reads, infers the exon-exon boundaries based on the CIGAR strings (Li et al., 2009), and outputs each “junction” as a feature in BED12 format. The junctions annotate command takes a file of junctions in BED12 format (such as the one output by junctions extract), a FASTA file containing the reference genome, and a GTF file containing reference transcriptome annotations and generates a TSV file, annotating each junction with: the number of acceptor sites, donor sites, and exons skipped, and the identities of known overlapping transcripts and genes. We also annotate the “junction type”, which denotes if and how the junction is novel (i.e. different compared to provided transcript annotations). If the donor is known, but the acceptor is not or vice-versa, it is marked as “D” or “A”, respectively. If both are known, but the pairing is not known, it is marked as “NDA”, whereas if both are unknown, it is marked as “N”. If the junction is not novel (i.e. it appears in at least one transcript in the supplied GTF), it is marked as “DA” (Figure 6.1B).

The cis-splice-effects module contains the identify command, which identifies potential splice-associated variants from sequencing data. The following are required as input: a VCF file containing variant calls, an alignment file containing aligned RNA-sequencing reads, a reference
genome FASTA file, and a reference transcriptome GTF file. The identify pipeline internally relies on variants annotate, junctions extract, and junctions annotate to output a TSV containing junctions proximal to putatively splice-associated variants. The identify pipeline can be customized using the same parameters as in the individual commands. Briefly, cis-splice-effects identify first performs variants annotate to determine the splicing relevance of each variant in the input VCF. For each variant, a “splice junction region” is determined by finding the largest span of sequence space between the exons that flank the exon associated with the variant. From here, junctions extract identifies splicing junctions present in the RNA-seq BAM. Next, junctions annotate labels each extracted junction with information from the reference transcriptome as described above and its associated variants based on splice junction region overlap (Figure 6.1C).

For our analysis, we annotated the pairs of associated variants and junctions identified by RegTools, which we refer to as “events”, with additional information such as whether this association was identified by a comparable tool, whether the junction was found in GTEx, and whether the event occurred in a cancer gene according to the Cancer Gene Census (CGC) (Figure 6.1C) (GTEx Consortium, 2013; Sondka et al., 2018). Finally, we created IGV sessions for each event identified by RegTools that contained a bed file with the junction, a VCF file with the variant, and an alignment (BAM) file for each sample that contained the variant (Thorvaldsdóttir et al., 2013). These IGV sessions were used to manually review candidate events to assess whether the association between the variant and junction makes sense in a biological context.

RegTools is designed for broad applicability and computational efficiency. By relying on well-established standards for sequence alignments, annotation files, and variant calls and by
remaining agnostic to downstream statistical methods and comparisons, our tool can be applied to a broad set of scientific queries and datasets. Moreover, performance tests show that cis-splice-effects identify can process a typical candidate variant list of 1,500,000 variants and a corresponding RNA-seq BAM file of 82,807,868 reads in just ~8 minutes (Appendix 1, Supplementary Figure 1).

6.4.2 Pan-cancer analysis of 35 tumor types identifies somatic variants that alter canonical splicing

RegTools was applied to 9,173 samples over 35 cancer types. 32 of these cohorts came from TCGA while the remaining three were obtained from other projects being conducted at MGI. Cohort sizes ranged from 21 to 1,022 samples. In total, 6,370,631 somatic variants (Appendix 1, Supplementary Figure 2A) and 2,387,989,201 junction observations (Appendix 1, Supplementary Figure 2B) were analyzed by RegTools. By comparing the number of initial variants to the number of statistically significant variants, we see that RegTools produces a prioritized list of potential splice-associated variants (Appendix 1, Supplementary Figure 3). Additionally, when analyzing the junctions within each sample, we found that junctions present in the reference transcriptome are frequently seen within GTEx data while junctions not present in the reference GTF are rarely seen within GTEx (Appendix 1, Supplementary Figure 4). These represent potentially novel tumor associated junctions. 235,778 significant variant junction pairings were found for junctions that use a known donor and novel acceptor (D), novel donor and known acceptor (A), or novel combination of a known donor and a known acceptor (NDA), with novel here meaning that the junction was not found in the reference transcriptome (Appendix 1, Supplementary Figure 2C, Supplemental Files 1 and 2). While our analysis
primarily focuses on variants in relation to novel splice events because of the potential importance of these events within tumor processes, we also wanted to assess how often a variant was significantly associated with a known junction. 5,157 variant junction pairings were found for canonical splice junctions (DA junctions) (Supplemental Files 3 and 4). This finding indicates that while splice-associated variants usually result in a novel junction occurring, they also may alter the expression of known junctions. Generally, significant events were evenly split among each of the novel junction types considered (D, A, and NDA). The number of significant events increased as the splice variant window size increased, with both the E and I results being comparable in number. Notably, hepatocellular carcinoma (HCC) was the only cohort that had whole genome sequencing (WGS) data available and, as expected, it exhibited a marked increase in the number of significant events for its results within the “I” splice variant window. This observation highlights the low sequence coverage of intronic regions that occurs with whole exome sequencing (WES) which subsequently leads to underpowered discovery of potential splice-associated variants within introns.
Figure 6.2 Splice-associated variants often lead to the expression of multiple alternative junctions. 

A) A single splice-associated variant can result in either one or more than one alternatively spliced junctions. Depicted is a variant resulting in a single novel transcript product (orange), a variant resulting in two novel transcript products that both use alternate donor sites (purple), and a variant resulting in multiple junctions of different types (green).

B) Stacked bars reveal how often significant splice-associated variants are associated with only one junction, multiple junctions of the same type, or multiple junctions of different types.

C) Bar chart showing how often each of the described junction combinations occurs when a single splice-associated variant results in multiple junction types across each of the RegTools splice variant windows used.
Variants were analyzed across tumor types for how often they result in either single or multiple novel junctions (Figure 6.2A). While a single variant resulting in a single novel junction is most commonly observed (72.27-83.78%), a single variant also commonly results in multiple junctions being created, either of the same type (6.56-10.94%) or of different types (9.66-16.79%) (Figure 6.2B). Variants that are associated with multiple novel junctions of different types were further investigated to identify how often a particular junction type occurred with another (Figure 6.2C). Most commonly, we observed an alternate donor or acceptor site being used in conjunction with an exon skipping event. These events were particularly common within the default window (2 intronic bases or 3 exonic bases from the exon edge), as a SNV or indel within these positions has a high probability of disrupting the natural splice site, thus causing the splicing machinery to use a cryptic splice site nearby or skip the exon entirely. The next most common event was an alternate donor site and an alternate acceptor site both being used as the result of a single variant. The combination of a novel acceptor site and novel donor site being used in conjunction with an exon-skipping event occurred the least and occurrence of this type of event remains fairly low, even as the search space increases within the larger splice variant windows. This finding indicates the low likelihood of a single variant resulting in simultaneous disruption of a splice acceptor and donor as well as complete skipping of an exon. Overall, this analysis highlights that there is evidence that a single variant can lead to multiple novel junctions being expressed. Tools, such as SpliceAI, that only allow for a single junction to be predicted or associated with a variant therefore may not be completely describing the effect of the variant in question in up to ~27% of cases.
6.4.3 RegTools identifies splice-associated variants missed by other splice-associated variant predictors and annotators

To evaluate the performance of RegTools, we compared our results to those of SAVNet, MiSplice, Veridical, VEP, and SpliceAI (Jaganathan et al., 2019; Jayasinghe et al., 2018; McLaren et al., 2016; Shiraishi et al., 2018; Viner et al., 2014). These tools vary in their inputs and methodology for identifying splice-associated variants (Figure 3A).

![Figure 3. Comparison of RegTools with other tools that identify potential splice-associated variants. A) UpSet plot comparing splice altering variants identified by RegTools to those identified by other splice variant predictors and annotators. Each tool and its total number of variant predictions are shown on the left sidebar graph. The numbers of variants specific to each tool or shared between different combinations of tools are indicated by the bar graph along the top and the individual or connected dots. B) Conceptual diagram of contrasting approaches employed by various splice variant identification tools/methods. An italicized tool name indicates that the source only considers genomic data for making its calls, as opposed to a combination of genomic and transcriptomic data.](image)

Both VEP and SpliceAI only consider information about the variant and its genomic sequence context and do not consider information from a sample’s transcriptome. A variant is considered to be splice-associated according to VEP if it occurs within 1-3 bases on the exonic side or 1-8 bases on the intronic side of a splice site. SpliceAI does not have restrictions on where the
variant can occur in relation to the splice site but by default, it predicts one new donor and acceptor site within 50 bp of the variant, based on reference transcript sequences from GENCODE. Like RegTools, SAVNet, MiSplice, and Veridical integrate genomic and transcriptomic data in order to identify splice-associated variants. MiSplice only considers junctions that occur within 20 bp of the variant. Additionally, SAVNet, MiSplice, and Veridical filter out any transcripts found within the reference transcriptome. SAVNet, MiSplice, and Veridical employ different statistical methods for the identification of splice-associated variants. In contrast to RegTools, none of the mentioned tools allow the user to set a custom window in which they wish to focus splice-associated variant discovery (e.g. around the splice site, all exonic variants, etc.). These tools have different levels of code availability. MiSplice is available via GitHub as a collection of Perl scripts that are built to run via Load Sharing Facility (LSF) job scheduling. To run MiSplice without an LSF cluster, the authors mention code changes are required. Veridical is available via a subscription through CytoGnomix’s MutationForecaster. Similar to RegTools, SAVNet is available via GitHub or through a Docker image. However, SAVNet relies on splicing junction files generated by STAR (Dobin et al., 2013) whereas RegTools can use RNA-Seq alignment files from HISAT2 (Kim et al., 2015), TopHat2 (Kim et al., 2013), or STAR, thus allowing it to be integrated into bioinformatics workflows more easily. To demonstrate the time that it would take to generate the needed STAR splicing junction files and then run SAVNet, we benchmarked RegTools and SAVNet using LUAD samples from TCGA (Appendix 1, Supplementary Figure 5). On average, it took SAVNet 3.2 times as long as RegTools to to run on the same samples when taking into account the unalignment and realignment required to generate the necessary starting files.
In their recent publications, SAVNet (Shiraishi et al., 2018), MiSplice (Jayasinghe et al., 2018), and Veridical (Shirley et al., 2018; Viner et al., 2014) also analyzed data from TCGA, with only minor differences in the number of samples included for each study. VEP and SpliceAI results were obtained by running each tool on all starting variants for the 35 cohorts included in this study. In order to efficiently compare these data, an UpSet plot (Figure 6.3A) was created (Conway et al., 2017). Only 343 variants are identified as splice-associated by all six tools. Comparatively, MiSplice and SAVNet find few splice-associated variants, potentially indicating that these tools are overlooking the complete set of variants that have an effect on splicing. In contrast, Veridical identifies by far the most splice-associated variants across all tools, with 94.54 percent of its calls being found by it alone. SpliceAI and VEP called a large number of variants, either alone or in agreement, that none of the tools that integrate transcriptomic data from samples identify. This highlights a limitation of using tools that only focus on genomic data, particularly in a disease context where transcripts are unlikely to have been annotated before. RegTools addresses these short-comings by identifying what pieces of information to extract from a sample’s genome and transcriptome in a very basic, unbiased way that allows for generalization. Other tools either only analyze genomic data, focus on junctions where either the canonical donor or acceptor site is affected (missing junctions that result from complete exon skipping), or consider only those variants within a very narrow distance from known splice sites. RegTools can include any kind of junction type, including exon-exon junctions that have ends that are not known donor/acceptor sites according to the GTF file (N junction according to RegTools), any distance value to make variant-junction associations, and any window size in which to consider variants. Due to these advantages, RegTools identified events missed by one
or multiple of the tools to which we compared (Figure 6.3B; Appendix 1, Supplementary Figures 6 and 7).

### 6.4.4 Orthogonal validation of RegTools using clinical data and verified splice-associated mutations

The large number of events that RegTools identified as significant compared to other tools could suggest a high percentage of false positives within our results. Therefore, we tested RegTools against multiple datasets to further validate this tool suite. The first dataset that we compared against was the 10 splice-site-creating mutations that Jayasinghe et al. validated using mini-gene functional assays (Jayasinghe et al., 2018). They selected 11 mutations that their tool, MiSplice, originally identified from TCGA data. These mutations were then compared to wild-type sequences using a pCAS2.1 splicing reporter mini-gene functional assay and 10 were validated through sequencing of alternatively spliced products. These 10 mutations were run through RegTools using corresponding aligned transcriptomic reads for each sample. RegTools was able to identify an association between all mutations and an aberrant splice junction (Supplementary File 5).

The next dataset that we used to validate RegTools was MutSpliceDB (Palmisano et al., 2021). This is a public database that contains manually reviewed RNA based evidence of effects of splice site variants on splicing. Currently, data is curated from TCGA and the Cancer Cell Line Encyclopedia (Barretina et al., 2012; Chang et al., 2017; Ghandi et al., 2019). When we accessed MutSpliceDB, there were 211 entries. Out of these 211 entries, 208 were found to affect splicing either through intron inclusion or exon skipping. We used the mutations provided and the corresponding RNA alignments in order to process each of these mutations through
RegTools. We were able to validate all 211 manually reviewed splice site variants (Supplementary File 6).

We also validated RegTools using clinical sequencing projects that allowed us to directly test the effects of somatic variants between clinical tumors within the same individual cases. The first dataset utilized is from Schaettler and Richters et al. (2022) which investigated the impact of spatial heterogeneity on genomic characteristics of gliomas and brain metastases (Schaettler et al., 2022). For this study, tumor tissue was surgically resected from 30 patients. Immediately following resection, each sample was dissociated into multiple (2-4) spatially distinct tumor regions that then underwent WES and RNA sequencing. We ran RegTools to identify splice-associated mutations within each distinct tumor region. A benefit of the heterogeneity of these samples and the multisector approach that was used is that we were able to compare splicing events directly within the tumor. This allowed us to validate associations within other tumor regions based on whether or not the variant was also present within those regions. Through this approach, we were able to validate 134 out of 146 splice-associated mutations in samples in which multiple sectors shared the same variant and aberrant junction. Conversely, we were also able to find 142 splice-associated variants out of 212 considered in the default splice variant window in which one sector contained a variant and novel splice junction but other regions in which both the variant and associated junction were absent (Supplementary File 7). This provides a form of biological validation that is otherwise difficult to observe in tumor cells and is more representative of true splicing biology than the typical mini-gene assay approach.

Another dataset that we employed was treatment-matched naive and post-treatment recurrence samples of small cell lung cancer (SCLC). By applying RegTools to these samples, we were able to identify splice-associated variants that persisted from the treatment-naive sample
to the recurrence sample (0%-35.96%). Additionally, we were able to identify samples where a splice-associated variant was lost due to treatment or arose post-treatment, either through the growth of a previously existing subclone or the emergence of a novel splice-associated mutation (64.04-100%) (Supplementary File 8).

To further validate RegTools, we wanted to validate some of its findings with long-read sequencing data in order to confirm the full-length structure of alternatively spliced isoforms. For this analysis, we used a well-described breast cancer cell line, HCC1395. For a normal comparator, we used HCC1395’s matched lymphoblastoid cell line, HCC1395BL. For each of these samples, whole genome, exome and RNA-seq were performed. For HCC1395, Oxford Nanopore Technologies long-read sequencing was performed using both the Direct RNA Sequencing Kit and Direct cDNA Sequencing Kit. After applying RegTools to the bulk genomic and transcriptomic data and obtaining candidate splice-associated variants, we were able to use the Nanopore long-read data to validate 20% of novel junctions observed within the short-read data and confirm the resulting novel transcript sequences (Supplementary File 9).

A final dataset that we used to evaluate RegTools was a single cell RNA (scRNA) dataset from a study investigating the mechanisms of response to immune checkpoint blockade (ICB) using MCB6C, a transplantable organoid model of urothelial carcinoma with features of human basal-squamous urothelial carcinoma (Sato et al., 2018). This model had also been subjected to WES of DNA isolated from tumor cells and matched normal cells from the tail of the mouse originally used to create the tumor. Analysis of the tumor/normal WES DNA was performed to identify somatic variants. We then identified single cells from three conditions and surveyed their expressed transcripts for evidence of the somatic variants. Each single cell was then classified as either tumor or normal, based on somatic variant expression, and separated into
corresponding alignment files. More specifically, to identify a tumor cell, we used the following criteria: two or more somatic variants detected with $>20X$ total coverage, $>5$ variant reads, and $>10\%$ variant allele fraction (VAF). To identify a normal cell, we used the following criteria: no variants detected and two or more of the variant positions with $>20X$ total coverage. Using these criteria, we defined 5,587 tumor cells and 17,022 normal cells for a total of 22,609 single cells. We processed these cells through an updated version of RegTools modified to support single cell data, treating each cell as an individual sample. This approach allowed us to greatly increase our power for determining tumor-associated splice-associated mutations due to all mutations being tumor-specific and each cell representing an independent readout of the splicing machinery. We were able to identify over 300 splice-associated mutations that had multiple cells of support, including within Trp53 and Bin1 (Figure 6.4; Appendix 1, Supplemental Figure 8).
Figure 6.4 Intronic SNV in *Trp53* associated with an exon skipping event. A) Schematic of a single nucleotide variant (mm10, chr11:g.69589711T>G) within an intron of *Trp53*. This variant is significantly associated with an exon skipping event causing the formation of an NDA junction. This result was found using the default splice variant window parameter (i2e3). B) UMAP projection of single cells from MCB6C organoid derived tumors with high confidence tumor cells (orange) and high confidence normal cells (blue) highlighted. C) UMAP projection of single cells from MCB6C organoid derived tumors overlaid with Log2 expression of *Trp53*. D) Zoomed view of cells containing the *Trp53* exon skipping event. E) Violin plot comparing the normalized junction score of the novel exon skipping event in cells with and without the variant.
Through the application of RegTools to the aforementioned datasets, we were able to identify high-quality, validated splice-associated mutations. Additionally, we were able to utilize well-designed clinical and scRNA datasets to more stringently identify tumor associated splice-associated mutations. These results demonstrate the broad utility of RegTools and its ability to robustly identify splice-associated somatic variants.

6.4.5 Pan-cancer analysis reveals novel splicing patterns within known cancer genes and potential cancer drivers

While efforts have been made to associate variants with specific cancer types, there has been little focus on identifying such associations in splice-associated variants, even those in known cancer genes. TP53 is a rare example whose splice-associated variants are well characterized in numerous cancer types (Surget et al., 2013). As such, we further analyzed significant events to identify genes that had recurrent splice-associated variants. Within each cohort, we looked for recurrent genes using two separate metrics: a binomial test p-value and the fraction of samples (see Methods). For ranking and selecting the most recurrent genes, each metric was computed by pooling across all cohorts. For assessing cancer-type specificity, each metric was then also computed using only results from a given cancer cohort. Since the mechanisms underlying the creation of novel junctions versus the disruption of existing splicing patterns may be different, analysis was performed separately for D/A/NDA junctions (Figure 6.5, Supplementary File 10) and DA junctions (Appendix 1, Supplementary Figure 9, Supplementary File 11), which allowed multiple test correction in accordance with the noise of the respective data.
We identified 6,954 genes in which there was at least one variant predicted to influence the splicing of a D/A/NDA junction. The 99th percentile of these genes, when ranked by either metric, are significantly enriched for known cancer genes, as annotated by the CGC (p=1.26E-19, ranked by binomial p-values, p=2.97E-24, ranked by fraction of samples; hypergeometric test). We also identified 3,643 genes in which there was least one variant predicted to influence the splicing of a DA (known) junction. The 99th percentile of these genes, when ranked by either metric, are also significantly enriched for known cancer genes, as annotated by the CGC (p=1.00E-04, ranked by binomial p-values, p=3.56E-07, ranked by fraction of samples; hypergeometric test). We also performed the same analyses using either the TCGA or MGI cohorts alone. The TCGA-only analyses gave very similar results to the combined analyses, with the 99th percentile of genes found in the D/A/NDA and DA analyses again being enriched for cancer genes (Appendix 1, Supplementary Figures 10 and 11). Due to small cohort sizes, in the MGI-only analyses, we identified only 329 and 208 genes in the D/A/NDA and DA analyses, respectively. The 99th percentile of genes from these analyses, respectively, were not significantly enriched for cancer genes (Supplementary Figures 12 and 13).

When analyzing D, A, and NDA junctions, we saw an enrichment for known tumor suppressor genes among the most splice disrupted genes, including several examples where
splice disruption is a known mechanism such as *TP53, PTEN, CDKN2A*, and *RB1*. Specifically, in the case of *TP53*, we identified 428 variants that were significantly associated with at least one novel splicing event. One such example is the intronic SNV (GRCh38, chr17:g.7673609C>A) that was identified in an OSCC sample and was associated with an exon skipping event and an alternate acceptor site usage event, with 23 and 41 reads of support, respectively (Appendix 1, Supplemental Figure 14). The cancer types in which we find splice disruption of TP53 and other known cancer genes is in concordance with associations between genes and cancer types described by CGC and CHASMplus (Sondka et al., 2018; Tokheim and Karchin, 2019). Our recovery of known drivers, many of which with known susceptibilities to splicing dysregulation in cancer, indicates the ability of our method to identify true splicing effects that are likely cancer-relevant.

Another cancer gene that we found to have a recurrence of splice-associated variants was *B2M*. Specifically, we identified six samples with intronic variants on either side of exon 2 (Figure 6.6). While mutations have been identified and studied within exon 2, we did not find literature that specifically identified intronic variants near exon 2 as a mechanism for disrupting *B2M* (Bicknell et al., 1996). These mutations were identified by VEP to be either splice acceptor or splice donor variants and were also identified by Veridical. MiSplice was able to predict one of the novel junctions for each variant but failed to predict additional novel junctions due to the limitation of that tool to only predict one novel acceptor and donor site per variant.
Figure 6.6 Several SNVs in *B2M* are associated with alternate acceptor and alternate donor usage. A) IGV snapshot of three intronic variant positions (GRCh38 - chr15:g.44715421A>G, chr15:g.44715422G>T, chr15:g.44715702G>C) found to be associated with usage of an alternate acceptor and alternative donor site that leads to the formation of novel transcript products. This result was found using the default splice variant window parameter (i2e3). B) Zoomed in view of the variants identified by RegTools that are associated with alternate acceptor and donor usage. Two of these variant positions flank the acceptor site and one variant flanks the donor site of the area that is being affected. C) Sashimi plot visualizations for samples containing the identified variants that show alternate acceptor usage (red) or alternate donor usage (orange).

Notably, 4 out of the 6 samples that these variants were found in are MSI-H (Microsatellite instability-high) tumors (Bonneville et al., 2017). Mutations in B2M, particularly within colorectal MSI-H tumors, have been identified as a method for tumors to become incapable of HLA class I antigen-mediated presentation (Kloor et al., 2005). Furthermore, in a study of patients treated with immune checkpoint blockade (ICB) therapy, defects to B2M were observed in 29.4% of patients with progressing disease (Sade-Feldman et al., 2017). In the same study, B2M mutations were exclusively seen in pretreatment samples from patients who did not respond to ICB or in post-progression samples after initial response to ICB (Sade-Feldman et al., 2017). There are several genes that are responsible for the processing, loading, and presentation of antigens, and have been shown to be mutated in cancers (Seliger et al., 2000). However, no proteins can be substituted for B2M in HLA class I presentation, thus making the loss of B2M a particularly robust method for ICB resistance (Güssow et al., 1987). We also observe exonic variants and variants further in intronic regions that disrupt canonical splicing of B2M. These findings indicate that intronic variants that result in alternative splice products within B2M may be a mechanism for immune escape within tumor samples.

We also identify recurrent splice-associated variants in genes not known to be cancer genes (according to CGC), such as RNF145. RegTools identified a recurrent single base pair deletion that results in an exon skipping event of exon 8 (Appendix 1, Supplementary Figure 15).
This gene is a paralog of *RNF139*, which has been found to be mutated in several MSI-H cancer types (Wang et al., 2017). This variant-junction association was found in STAD, UCEC, COAD, and ESCA tumors, all of which are considered to be MSI-H tumors (Bonneville et al., 2017). After analyzing the effect of the exon skipping event on the mRNA sequence, we concluded that the reading frame remains intact, possibly leading to a gain of function event. Additionally, the skipping of exon 8 leads to the removal of a transmembrane domain and a phosphorylation site, S352, which could be important for the regulation of this gene (Hornbeck et al., 2015). Based on these findings, *RNF145* is a promising candidate for a novel cancer driver.

While most of our analysis focused on splice-associated variants that resulted in D, A, NDA junctions, we also investigated variants that shifted the usage of known donor and acceptor pairs. Through this analysis, we identified *CDKN2A*, a tumor suppressor gene that is frequently mutated in numerous cancers (Zhao et al., 2016), to have several variants that led to alternate donor usage (Supplementary Figure 16). When these variants are present, an alternate known donor site is used that leads to the formation of the transcript ENST00000579122.1 instead of ENST00000304494.9, the transcript that encodes for p16ink4a, a known tumor suppressor. The transcript that results from use of this alternate donor site is missing the last twenty-eight amino acids that form the C-terminal end of p16ink4a. Notably, this removes two phosphorylation sites within the p16 protein, S140 and S152, which could disrupt the association of p16ink4a with CDK4 (Gump et al., 2003). This finding highlights the importance of including known transcripts in alternative splicing analyses, as variants may alter splice site usage in a way that results in a known but, potentially oncogenic transcript product.
6.5 Discussion

Splice-associated variants are often overlooked in traditional genomic analysis. To address this limitation, we created RegTools, a software suite for the analysis of variants and junctions in a splicing context. By relying on well-established standards for analyzing genomic and transcriptomic data and allowing flexible analysis parameters, we enable users to apply RegTools to a wide set of scientific methodologies and datasets. To ease the use and integration of RegTools into analysis workflows, we provide documentation and example workflows via (regtools.org) and provide a Docker image with all necessary software installed.

In order to demonstrate the utility of our tool, we applied RegTools to 9,173 tumor samples across 35 tumor types to profile the landscape of this category of variants. From this analysis, we report 133,987 variants that cause novel splicing events that were missed by VEP or SpliceAI. Only 1.4 percent of these mutations were previously discovered by similar attempts, while 98.6 percent are novel findings. We demonstrate that there are splice-associated variants that occur beyond the splice site consensus sequence, shift transcript usage between known transcripts, and create novel exon-exon junctions that have not been previously described. Specifically, we describe notable findings within \( B2M \), \( RNF145 \), and \( CDKN2A \). These results demonstrate the utility of RegTools in discovering novel splice-associated mutations and confirm the importance of integrating RNA and DNA sequencing data in understanding the consequences of somatic mutations in cancer. To allow further investigation of these identified events, we make all of our annotated result files (Supplemental Files 1-4) and recurrence analysis files (Supplemental Files 10-11) available.

Understanding the splicing landscape is crucial for unlocking potential therapeutic avenues in precision medicine and elucidating the basic mechanisms of splicing and cancer
The exploration of novel tumor-specific junctions will undoubtedly lead to translational applications, from discovering novel tumor drivers, diagnostic and prognostic biomarkers, and drug targets, to identifying a previously untapped source of neoantigens for personalized immunotherapy. While our analysis focuses on splice-associated variants within cancers, we believe RegTools will play an important role in answering a broad range of questions across different disease states and biological processes by helping users extract splicing information from transcriptome data and linking it to somatic or germline variant calls. The computational efficiency of RegTools and increasing availability and size of genomic and transcriptomic datasets may also allow for an improved understanding of splice regulatory motifs that have proven difficult to define, such as exonic and intronic splicing enhancers and silencers. Any group with paired DNA and RNA-seq data for the same samples stands to benefit from the functionality of RegTools.

6.6 Methods

6.6.1 Software implementation

RegTools is written in C++. CMake is used to build the executable from source code. We have designed the RegTools package to be self-contained in order to minimize external software dependencies. A Unix platform with a C++ compiler and CMake is the minimum prerequisite for installing RegTools. Documentation for RegTools is maintained as text files within the source repository to minimize divergence from the code. We have implemented common file handling tasks in RegTools with the help of open-source code from Samtools/HTSlib (Li et al., 2009) and BEDTools (Quinlan, 2014) in an effort to ensure fast performance, consistent file handling, and interoperability with any aligner that adheres to the BAM specification. Statistical tests are
conducted within RegTools using the RMath framework. GitHub actions and Coveralls are used to automate and monitor software compilation and unit tests to ensure software functionality. We utilized the Google Test framework to write unit tests.

RegTools consists of a core set of modules for variant annotation, junction extraction, junction annotation, and GTF utilities. Higher level modules such as cis-splice-effects make use of the lower level modules to perform more complex analyses. We hope that bioinformaticians familiar with C/C++ can re-use or adapt the RegTools code to implement similar tasks.

6.6.2 Benchmarking

Performance metrics were calculated for all RegTools commands. Each command was run with default parameters on a single blade server (Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz) with 10 GB of RAM and 10 replicates for each data point (Supplementary Figure 1). Specifically for cis-splice-effects identify, we started with random selections of somatic variants, ranging from 10,000-1,500,000, across 8 data subsets. Using the output from cis-splice-effects identify, variants annotate was run on somatic variants from the 8 subsets (range: 0-17,742) predicted to have a splicing consequence. The function junctions extract was performed on the HCC1395 tumor RNA-seq data aligned with HISAT to GRCh37 and randomly downsampling at intervals ranging from 10-100%. Using output from junctions extract, junctions annotate was performed for 7 data subsets ranging from 1,000-500,000 randomly selected junctions.

Benchmark tests revealed an approximately linear performance for all functions. Variance between real and CPU time is highly dependent on the I/O speed of the write-disk and could account for artificially inflated real time values given multiple jobs writing to the same disk at once. The most computationally expensive function in a typical analysis workflow was junctions extract, which on average processed 33,091 reads/second (CPU) and took an average
of 43.4 real vs 41.7 CPU minutes to run on a full bam file (82,807,868 reads total). The function junctions annotate was the next most computationally intensive function and took an average of 33.0 real/8.55 CPU minutes to run on 500,000 junctions, processing 975 junctions/second (CPU). The other functions were comparatively faster with cis-splice-effects identify and variants annotate able to process 3,105 and 118 variants per second (CPU), respectively. To process a typical candidate variant list of 1,500,000 variants and a corresponding RNA-seq BAM file of 82,807,868 reads with cis-splice-effects identify takes ~ 8.20 real/8.05 CPU minutes (Supplementary Figure 1).

Performance metrics were also calculated for the statistics script and its associated wrapper script that handles dividing the variants into smaller chunks for processing to limit RAM usage. This command, compare_junctions, was benchmarked in January 2020 using Amazon Web Services (AWS) on a m5.4xlarge instance, based on the Amazon Linux 2 AMI, with 64 Gb of RAM, 16 vCPUs, and a mounted 1 TB SSD EBS volume with 3000 IOPS. These data were generated from running compare_junctions on each of the included cohorts, with the largest being our BRCA cohort (1022 sample) which processed 3.64 events per second (CPU).

For the benchmarking comparison between RegTools and SAVNet, we utilized fifty LUAD samples from TCGA. For the purposes of our comparison, we imagined a use case where an individual would start by downloading alignment files from the Genomic Data Commons (GDC) Data Portal. For RegTools CPU and real time measurements, regtools junctions extract, regtools cis-splice-effects associate, and compare_junctions was ran for each sample. For SAVNet’s CPU and real time measurements, alignment files were first unaligned using SamToFastq and then realigned using STAR to get each sample’s splice junction file, which is not available from the GDC Data Portal. Following these steps, SAVNet was then run and the
time was added to that from the unalignment and realignment step. On average, it took SAVNet 3.2 times (real time) as long as RegTools to run on the same samples when taking into account the unalignment and realignment required to generate the necessary starting files.

### 6.6.3 Using RegTools to identify cis-acting, splice-associated variants

RegTools contains three sub-modules: “variants”, “junctions”, and “cis-splice-effects”. For complete instructions on usage, including a detailed workflow for how to analyze cohorts using RegTools, please visit regtools.org.

**variants annotate**

This command takes a list of variants in VCF format. The file should be gzipped and indexed with Tabix (Li, 2011). The user must also supply a GTF file that specifies the reference transcriptome used to annotate the variants.

The INFO column of each line in the VCF is populated with comma-separated lists of the variant-overlapping genes, variant-overlapping transcripts, the distance between the variant and the associated exon edge for each transcript (i.e. each start or end of an exon whose splice variant window included the variant) defined as \( \min(\text{distance}\_\text{from}\_\text{start}\_\text{of}\_\text{exon}, \text{distance}\_\text{from}\_\text{end}\_\text{of}\_\text{exon}) \), and the variant type for each transcript.

Internally, this function relies on HTSlib to parse the VCF file and search for features in the GTF file which overlap the variant. The splice variant window size (i.e. the maximum distance from the edge of an exon used to consider a variant as splice-associated) can be set by the options “-e <number of bases>” and “-i <number of bases>” for exonic and intronic variants, respectively. The variant type for each variant thus depends on the options used to set the splice variant window size. Variants captured by the window set by “-e” or “-i” are annotated as “splicing_exonic” and “splicing_intronic”, respectively. Alternatively, to analyze all exonic or
intronic variants, the “-E” and “-I” options can be used. Otherwise, the “-E” and “-I” options themselves do not change the variant type annotation, and variants found in these windows are labeled simply as “exonic” or “intronic”. By default, single exon transcripts are ignored, but they can be included with the “-S” option. By default, output is written to STDOUT in VCF format. To write to a file, use the option “-o <PATH/TO/FILE>”.

junctions extract

This command takes an alignment file containing aligned RNA-seq reads and infers junctions (i.e. exon-exon boundaries) based on skipped regions in alignments as determined by the CIGAR string operator codes. These junctions are written to STDOUT in BED12 format. Alternatively, the output can be redirected to a file with the “-o <PATH/TO/FILE>”. RegTools ascertains strand information based on the XS tags set by the aligner, but can also determine the inferred strand of transcription based on the BAM flags if a stranded library strategy was employed. In the latter case, the strand specificity of the library can be provided using “-s <INT>” where 0 = unstranded, 1 = first-strand/RF, 2 = second-strand/FR. We suggest that users align their RNA-seq data with HISAT2 (Kim et al., 2015), TopHat2 (Kim et al., 2013), STAR (Dobin et al., 2013), kallisto (Bray et al., 2016), or minimap2 (Li, 2018) as these are the aligners we have tested to date. We have tested RegTools with data from Illumina, Oxford Nanopore Technologies, and 10X Genomics.

Users can set thresholds for minimum anchor length and minimum/maximum intron length. The minimum anchor length determines how many contiguous, matched base pairs on either side of the junction are required to include it in the final output. The required overlap can be observed amongst separated reads, whose union determines the thickStart and thickEnd of the BED feature. By default, a junction must have 8 bp anchors on each side to be counted but this
can be set using the option “-a <minimum anchor length>”. The intron length is simply the end coordinate of the junction minus the start coordinate. By default, the junction must be between 70 bp and 500,000 bp, but the minimum and maximum can be set using “-i <minimum intron length>” and “-I <maximum intron length>”, respectively.

For efficiency, this tool can be used to process only alignments in a particular region as opposed to analyzing the entire BAM file. The option “-r <chr>:<start>-<stop>” can be used to set a single contiguous region of interest. Multiple jobs can be run in parallel to analyze separate non-contiguous regions.

junctions annotate

This command takes a list of junctions in BED12 format as input and annotates them with respect to a reference transcriptome in GTF format. The observed splice-sites used are recorded based on a reference genome sequence in FASTA format. The output is written to STDOUT in TSV format, with separate columns for the number of splicing acceptors skipped, number of splicing donors skipped, number of exons skipped, the junction type, whether the donor site is known, whether the acceptor site is known, whether this junction is known, the overlapping transcripts, and the overlapping genes, in addition to the chromosome, start, stop, junction name, junction score, and strand taken from the input BED12 file. This output can be redirected to a file with “-o /PATH/TO/FILE”. By default, single exon transcripts are ignored in the GTF but can be included with the option “-S”.

cis-splice-effects identify

This command combines the above utilities into a pipeline for identifying variants which may cause aberrant splicing events by altering splicing motifs in cis. As such, it relies on essentially the same inputs: a gzipped and Tabix-indexed VCF file containing a list of variants, an alignment
file containing aligned RNA-seq reads, a GTF file containing the reference transcriptome of interest, and a FASTA file containing the reference genome sequence of interest.

First, the list of variants is annotated. The splice variant window size is set using the options “-e”, “-i”, “-E”, and “-I”, just as in variants annotate. The splice junction region size (i.e. the range around a particular variant in which an overlapping junction is associated with the variant) can be set using “-w <splice junction region size>”. By default, this range is not a particular number of bases but is calculated individually for each variant, depending on the variant type annotation. For “splicing_exonic”, “splicing_intronic”, and “exonic” variants, the region extends from the 3’ end of the exon directly upstream of the variant-associated exon to the 5’ end of the exon directly downstream of it. For “intronic” variants, the region is limited to the intron containing the variant. Single-exons can be kept with the “-S” option. The annotated list of variants in VCF format (analogous to the output of variants annotate) can be written to a file with “-v /PATH/TO/FILE”.

The BAM file is then processed in the splice junction regions to produce the list of junctions. A file containing these junctions in BED12 format (analogous to the output of junctions extract) can be written using “-j /PATH/TO/FILE”. The minimum anchor length, minimum intron length, and maximum intron length can be set using “-a”, “-i”, and “-I” options, just as in junctions extract.

The list of junctions produced by the preceding step is then annotated with the information presented in junctions annotate. Additionally, each junction is annotated with a list of associated variants (i.e. variants whose splice junction regions overlapped the junction). The final output is written to STDOUT in TSV format (analogous to the output of junctions annotate) or can be redirected to a file with “-o /PATH/TO/FILE”.
cis-splice-effects associate

This command is similar to cis-splice-effects identify, but takes the BED output of junctions extract in lieu of an alignment file with RNA alignments. As with cis-splice-effects identify, each junction is annotated with a list of associated variants (i.e. variants whose splice junction regions overlapped the junction). The resulting output is then the same as cis-splice-effects identify, but limited to the junctions provided as input.

6.6.4 Dataset Description

32 cancer cohorts were analyzed from TCGA. These cancer types are Adrenocortical carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG), Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon adenocarcinoma (COAD), Esophageal carcinoma (ESCA), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Lymphoid Neoplasm Diffuse Large B cell Lymphoma (DLBC), Mesothelioma (MESO), Ovarian serous cystadenocarcinoma (OV), Pancreatic adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate adenocarcinoma (PRAD), Rectum adenocarcinoma (READ), Sarcoma (SARC), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Testicular Germ Cell Tumors (TGCT), Thymoma (THYM), Thyroid carcinoma (THCA), Uterine Carcinosarcoma (UCS), Uterine Corpus Endometrial Carcinoma (UCEC), and Uveal Melanoma (UVM). Three cohorts were derived from patients at Washington University in St. Louis. These cohorts are
Hepatocellular Carcinoma (HCC), Oral Squamous Cell Carcinoma (OSCC), and Small Cell Lung Cancer (SCLC).

6.6.5 Sample processing for cohorts with bulk transcriptome data

We applied RegTools to 35 tumor cohorts. Genomic and transcriptomic data for 32 cohorts were obtained from The Cancer Genome Atlas (TCGA). Information regarding the alignment and variant calling for these samples is described by the Genomic Data Commons data harmonization effort (https://gdc.cancer.gov/about-data/gdc-data-processing). Whole exome sequencing (WES) mutation calls for these samples from MuSE (Fan et al., 2016), MuTect2 (Cibulskis et al., 2013), VarScan2 (Koboldt et al., 2012), and SomaticSniper (Larson et al., 2012), were left-aligned, trimmed, and decomposed to ensure the correct representation of the variants across the multiple callers.

Samples for the remaining three cohorts, HCC (Skidmore et al., 2022), SCLC (Wagner et al., 2018), and OSCC (Campbell et al., 2018), were sequenced at Washington University in St. Louis. Genomic data were produced by WES for SCLC and OSCC and whole genome sequencing (WGS) for HCC. Normal genomic data of the same sequencing type and tumor RNA-seq data were also available for all subjects. Sequence data were aligned using the Genome Modeling System (GMS) (Griffith et al., 2015a) using TopHat2 for RNA and BWA-MEM (Li and Durbin, 2009) for DNA. HCC and SCLC were aligned to GRCh37 while OSCC was aligned to GRCh38. Somatic variant calls were made using Samtools v0.1.1 (Li et al., 2009), SomaticSniper2 v1.0.2 (Larson et al., 2012), Strelka V0.4.6.2 (Saunders et al., 2012), and VarScan v2.2.6 (Koboldt et al., 2012) through the GMS. High-quality mutations for all samples were then selected by requiring that a variant be called by two of the four variant callers.
Additional samples from orthogonal projects at Washington University in St. Louis were used for the orthogonal validation analysis. Samples included in this analysis were of the following cancer types: SCLC, GBM, and brain metastases that resulted from lung and breast cancers. The SCLC samples were an extension of the previous SCLC cohort we used for the pan-cancer analysis, with the difference being that the additional samples being aligned to GRCh38. Methods for the sample processing of the GBM and brain metastases samples are described in the original analysis manuscript for these data (Schaettler et al., 2022).

6.6.6 Mice used for MCB6C experiments

5- to 6-week-old black 6 (B6NTac) male mice were purchased from Taconic Biosciences and were housed in a SPF barrier facility under the guidelines of Institutional Animal Care and Use Committee at Washington University. All the in vivo experiments were performed one week after mice were delivered to the animal facility.

6.6.7 Mouse bladder organoid culture for mouse injection

One previously archived frozen vial of singly suspended MCB6C organoid was thawed at least 2 weeks before mouse injection and expanded weekly in culture at least 2 times. For MCB6C organoid culture expansion, growth factor reduced Matrigel was thawed on ice for minimally 1½ hours. Pelleted MCB6C cells were washed and resuspended in 1 ml of Advanced DMEM/F12+++ medium (Advanced DMEM/F12 medium [125634028, Gibco] supplemented with 1% penicillin/streptomycin, 1% HEPEs and Glutamax) and cell concentration was determined by automated cell counter. To establish organoid culture, 50ul Matrigel tabs with 10,000 cells/tab were generated and plated on 6-well suspension culture plates, 6 tabs wells. Tabs were incubated at 37C for 15 minutes until Matrigel was hardened, returned to tissue
culture incubator, and cultured with mouse bladder organoid medium (MBO medium -
Advanced DMEM/F12+++ medium supplemented with EGF, A-83-01, Noggin, R-Spondin, N-
Acetly-L-cysteine and Nicotinamide). Organoids were replenished with fresh MBO medium
every 3-4 days and also one day before mouse injection.

6.6.8 Mouse injection with MCB6C organoid cells

A single cell suspension of MCB6C organoid was generated by TrypLE Express(12605010,
Gibco) digestion organoid Matrigel tabs at 37C for 15 minutes. After digestion, pelleted cells
were washed and resuspended in PBS to determine cell concentration. After cell concentration
was adjusted to 20 million/ml in PBS, organoid cells were mixed with growth factor reduced
Matrigel at 1:1 ratio before injected subcutaneously into the left flank of the mouse (1
million/100ul cells each mouse). Tumor development was monitored using digital calipers to
assess the length, width, and depth of each tumor. For ICB, each mouse was injected
intraperitoneally with 250 ug anti-PD1 (RMP1-14, BioXcell) and 200 ug anti-CTLA-4 (9D9,
BioXcell) day 9 and 12 after organoid implantation. For isotype controls, each mouse was
injected with 250 ug rat IgG2a (2A3, BioXcell) and 200 ug IgG2b (MPC-11, BioXcell). For
CD4+ T cell depletion, each mouse was injected with 250 ug anti-CD4 (GK1.5, BioXcell) day 0
and 7 after organoid depletion. Rat IgG2b (LTF-2, BioXcell) was used as isotype control for
anti-CD4.

6.6.9 Harvesting MCB6C tumors for single cell RNA-seq analysis

Based on 10x Genomics Demonstrated Protocols, 14 days after organoid implantation, tumors
were dissected from euthanized mice, cut into small pieces of ~2-4 mm3, and further processed
into dead-cell depleted single cell suspension following manufacturer’s protocol using Tumor
Dissociation Kit and MACS Dead Cell removal Kit (Miltenyi Biotec). Briefly, tumor tissue pieces were transferred to gentleMACS C tube containing enzyme mix before loading onto a gentleMACS Octo Dissociator with Heaters for tissue digestion at 37°C for 80 minutes. After tissue dissociation was completed, cell suspension was transferred to a new 50 ml conical tube, and supernatant was removed after centrifugation. Cell pellet was resuspended in RPMI 1640 medium, filtered through a prewetted 70-uM cell strained, pelleted, and resuspended in red cell lysis buffer and incubated on ice for 10 minutes. After adding wash buffer, cell suspension was pelleted and resuspended in wash buffer. To remove dead cells, Dead Cell Removal Microbeads were added to resuspend cell pellet (100μl beads per 10^7 cells) using a wide-bore pipette tip. After incubation for 15 minutes at room temperature, the cell-microbead mixture was applied onto a MS column. Dead cells remained in the column and the effluent represented the live cell fraction. The percentage of viable cells was determined by an automated cell counter. Dead cell removal was repeated if the percentage of viable cells did not reach above 90%. Two rounds of centrifugation/resuspension were carefully performed for two rounds in 1xPBS/0.04% BSA using a wide-bore tip. To submit cell samples for single-cell RNA-seq analysis, cell concentration was determined accurately by sampling cell suspension twice and counting each sampling twice and adjusted to 1167 cells/μl.

6.6.10 Single-cell RNA-seq analysis of MCB6C cells

40μl of each cell suspension was submitted to GTAC/MGI for 10x Genomics single cell RNA-seq analysis using the 5’v2 library kit of TotalSeq C antibodies with BCR and TCR V(D)J enrichment. FASTQs and Cell Ranger output was generated. Alignment and gene expression quantification done using CellRanger count (v5.0). Matrices are then imported into Seurat (v4.0.1) for filtering cells, QC, clustering, etc. To filter suspected dying cells, cells were
clustered before filtering and cells clustering based on high mitochondrial gene expression were identified. The cutoff of mitochondrial expression was based on the expression level that captures most of these cells. Doublets were filtered based on high UMI expression, with the top 0.9% of genes removed from each condition in each replicate. Cutoffs for filtering of cells with low feature detection was done by assigning cell type to each cell with CellMatch, identifying cells that did not have enough features for their cell type to be predicted, and identifying average feature expression in these cells. Aftering filtered cells were removed, remaining cells were scaled, normalized, and clustered following Seurat’s vignette.

6.6.11 Long read sequencing of HCC1395 cell line

HCC1395 cell line is described as being of tissue origin: mammary gland; breast/duct. The patient’s cancer was described as: TNM stage I, grade 3, primary ductal carcinoma. The patient received chemotherapy prior to isolation of the tumor (Gazdar et al., 1998). This tumor is considered “Triple-Negative” by classic typing: ERBB2-negative (aka HER2/neu), PR-negative, and ER-negative. Otherwise, it is one of those difficult to classify by expression-based molecular typing but is likely of the “Basal” sub-type (Heiser et al., 2012). For a normal comparator, we used HCC1395’s matched lymphoblastoid cell line, HCC1395BL. The HCC1395BL cell line was created from a B lymphoblast that was transformed by the EBV virus. For each of these samples, whole genome, exome and RNA-seq were performed. Whole-genome sequencing was performed to a target median coverage depth of ~30x for the normal samples and ~50x for the tumor sample. Exome sequencing was performed to a target median coverage depth of ~100x. RNA-seq was performed for both tumor and normal. Additionally, Oxford Nanopore Technologies long-read sequencing was performed using both the Direct RNA Sequencing Kit and Direct cDNA Sequencing Kit. The Direct RNA Sequencing Kit yielded 1.1 million reads
with 1.07 Gb of passed bases and read lengths ranging from ~500 basepairs (bp) to ~8 kilobases (kb). The Direct cDNA Sequencing Kit was run twice. The first run used the RNA from the same mRNA enrichment as the RNA Direct library. This sequencing run yielded 2.48 million reads with 2.36 Gb of passed bases and read lengths ranging from ~500 bp to ~9.6 kb. The second Direct cDNA Sequencing Kit was applied to a new RNA extraction, so a separate mRNA enrichment from the first two runs. This run yielded 6.6 million reads with 4.05 Gb passed bases and read lengths ranging from ~500 basepairs (bp) to ~8 kilobases (kb). These data were aligned to GRCh38 using recommended settings for minimap2 (Li, 2018).

6.6.12 Candidate junction filtering

To generate results for 4 splice variant window sizes, we ran cis-splice-effects identify with 4 sets of splice variant window parameters. For our “i2e3” window (RegTools default), to examine intronic variants within 2 bases and exonic variants within 3 bases of the exon edge, we set “-i 2 -e 3”. Similarly, for “i50e5”, to examine intronic variants within 50 bases and exonic variants within 5 bases of the exon edge, we set “-i 50 -e 5”. To view all exonic variants, we simply set “-E”, without “-i” or “-e” options. To view all intronic variants, we simply set “-I”, without “-i” or “-e” options. TCGA samples were processed with GRCh38.d1.dl1.fa (downloaded from the GDC reference file page at https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files) as the reference fasta file and gencode.v29.annotation.gtf (downloaded via the GENCODE FTP) as the reference transcriptome. OSCC was processed with Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa and Homo_sapiens.GRCh38.79.gtf (both downloaded from Ensembl). HCC and SCLC were processed with Homo_sapiens.GRCh37.dna_sm.primary_assembly.fa and Homo_sapiens.GRCh37.87.gtf (both downloaded from Ensembl).
6.6.13 Statistical filtering of candidate events

We refer to a statistical association between a variant and a junction as an “event”. For each event identified by RegTools, a normalized score (norm_score) was calculated for the junction of the event by dividing the number of reads supporting that junction by the sum of all reads for all junctions within the splice junction region for the variant of interest. This metric is conceptually similar to a “percent-spliced in” (PSI) index, but measures the presence of entire exon-exon junctions, instead of just the inclusion of individual exons. If there were multiple samples that contained the variant for the event, then the mean of the normalized scores for the samples was computed (mean_norm_score). If only one sample contained the variant, its mean_norm_score was thus equal to its norm_score. This value was then compared to the distribution of samples that did not contain the variant to calculate a p-value as the percentage of the norm_scores from these samples, which are at least as high as the mean_norm_score computed for the variant-containing samples. We performed separate analyses for events involving canonical junctions (DA) and those involving novel junctions that used at least one known splice site (D/A/NDA), based on annotations in the corresponding reference GTF. For this study, we filtered out any junctions which did not use at least one known splice site (N) and junctions that did not have at least 5 reads of evidence across variant-containing samples. The Benjamini-Hochberg procedure was then applied to the remaining events. Following correction, an event was considered significant if its adjusted p-value was ≤ 0.05.

6.6.14 Annotation with GTEx junction data and other splice prediction tools

Events identified by RegTools as significant were annotated with information from GTEx, VEP, SpliceAI, MiSplice, and Veridical. GTEx junction information was obtained from the GTEx
Portal. Specifically, the exon-exon junction read counts file from the v8 release was used for data aligned to GRCh38 while the same file from the v7 release was used for the data aligned to GRCh37. Mappings between tumor cohorts and GTEx tissues can be found in Supplemental File 12. We annotated all starting variants with VEP in the “per_gene” and “pick” modes. The “per_gene” setting outputs only the most severe consequence per gene while the “pick” setting picks one line or block of consequence data per variant. We considered any variant with at least one splice-associated annotation to be “VEP significant”. All variants were also processed with SpliceAI using the default options. A variant was considered to be “SpliceAI significant” if it had at least one score greater than 0.2, the developers’ value for high recall of their model. Instructions and scripts to annotate with GTEx and SpliceAI are available at regtools.org and the RegTools GitHub repository. Variants identified by MiSplice (Jayasinghe et al., 2018) were obtained from the paper supplemental tables and were lifted over to GRCh38. Variants identified by SAVNet (Shiraishi et al., 2018) were obtained from the paper supplemental tables and were lifted over to GRCh38. Variants identified by Veridical (Shirley et al., 2018; Viner et al., 2014) were obtained via download from the link referenced within the manuscript and lifted over to GRCh38.

6.6.15 Visual exploration of statistically significant candidate events

IGV sessions were created for each event identified by RegTools that was statistically significant. Each IGV session file contained a bed file with the junction, a vcf file with the variant, and an alignment file for each sample that contained the variant. Additional information, such as the splice sites predicted by SpliceAI, were also added to these session files to enhance the exploration of these events. Events of interest were manually reviewed in IGV to assess whether the association between the variant and junction made sense in a biological context (e.g.
affected a known splice site, altered a genomic sequence to look more like a canonical splice site, or the novel junction disrupted active or regulatory domains of the protein product). An extensive review of literature and visualizations of junction usage in the presence and absence of the variant were also used to identify novel, biologically relevant events.

### 6.6.16 Identification of genes with recurrent splice-associated variants

For each cohort, we calculated a p-value to assess whether the splicing profile from a particular gene was significantly more likely to be altered by somatic variants. Specifically, we performed a 1-tailed binomial test, considering the number of samples in a cohort as the number of attempts. Success was defined by whether the sample had evidence of at least one splice-associated variant in that gene. The null probability of success, $p_{null}$ was calculated as

\[
p_{null} = 1 - (1 - Pr(V \land A))^s
\]

where $s$ is the total number of base positions residing in any of the gene’s splice variant windows, $V$ is the event that a somatic variant occurred at such a base position, and $A$ is the event that this variant was deemed to be significantly associated with at least one junction in our analysis. The joint probability that both $V$ and $A$ occurred was estimated by dividing the total of events across all samples in which each junction was detected by $s$. The value of $s$ was computed based on the exon and transcript definitions in the reference GTF used for performing RegTools analyses on a given cohort.

We also calculated overall metrics, in order to rank genes. For each set of cohorts (e.g. TCGA-only, MGI-only, combined), an overall p-value was computed for each gene according to the above formula, pooling all of the samples across the included cohorts, and the fraction of
samples was simply calculated by dividing the number of samples in which an event occurred within the given gene by the total number of samples, pooled across the included cohorts. The reference GTF used for analyzing the TCGA samples (i.e. gencode.v29.annotation.gtf) was used for all sets of cohorts.

6.7 Code Availability

RegTools is open source (MIT license) and available at https://github.com/griffithlab/regtools/. All scripts used in the analyses presented here are also provided. For ease of use, a Docker container has been created with RegTools, SpliceAI, R, and Python 3 installed (https://hub.docker.com/r/griffithlab/regtools/). This Docker container allows a user to run the workflow we outline at https://regtools.readthedocs.io/en/latest/workflow/. Docker is an open-source software platform that enables applications to be readily installed and run on any system. The availability of RegTools with all its dependencies as a Docker container also facilitates the integration of the RegTools software into workflow pipelines that support Docker images.

6.8 Data Availability

Sequence data for each cohort analyzed in this study are available through dbGaP at the following accession IDs: phs000178 for TCGA cohorts, phs001106 for HCC, phs001049 for SCLC, phs001623 for OSCC, and phs002612.v1.p1 for GBM/Brain metastases. All supplemental files are available through the bioRxiv link at the beginning of this chapter. Statistically significant events for D, A, and NDA junctions across the four variant splicing windows used are available via Supplemental Files 1 and 2. Statistically significant events for DA junctions are available as Supplemental Files 3 and 4. Complete results of gene recurrence analysis are available as Supplemental Files 5 and 6.
6.9 Acknowledgments

We thank the patients and their families for the donation of their samples and participation in clinical trials. We would like to thank Donald Conrad for his initial idea to compare to variant effect predictor tools. Kelsy Cotto was supported by Siteman Cancer Center under fund number #3477-92400 and T32CA113275. Avinash Ramu was supported by the Burroughs Wellcome Fund Institutional Program Unifying Population and Laboratory Based Sciences Award at Washington University. Malachi Griffith was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under Award Number R00HG007940. Malachi Griffith and Obi Griffith were supported by the NIH National Cancer Institute (NCI) under Award Numbers U01CA209936, U01CA231844, U01CA248235 U24CA237719. Malachi Griffith and Megan Richters were supported by the V Foundation for Cancer Research under Award Number V2018-007. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

6.10 Contributions

K.C.C. and Y.-Y.F. were involved in all aspects of this study, including designing methodology, developing and testing the tool software, analyzing and interpreting data, and writing the manuscript, with input from A.R., M.R., S.L.F., Z.L.S., H.X., J.F.M., J.K., K.M.C., T.H.C., E.B.R., D.A., S.D., O.L.G., and M.G. A.R. designed the tool and led software development efforts. Y.L., W.C.C., C.A.M., V.A., G.P.D., R.U., and R.G. provided unpublished tumor datasets and provided critical feedback on the manuscript. O.L.G. and M.G. supervised the study. All authors read and approved the final manuscript.
Chapter 7: Conclusions and Future Directions

In theory, precision oncology is an elegant and obvious method for optimizing treatment protocols for patients. In practice, however, the components and logistics required to effectively actualize precision oncology are extensive. In this research, we described many of the barriers preventing the adoption of precision oncology and presented solutions to some of them. We described our efforts to maintain and update DGIdb with information about potentially druggable genes and known drug-gene interactions (Chapters 2 and 3). We also detail our efforts to implement new features within the UI and API to better address specific use cases and add novel functionality to better characterize the interactions summarized within the database. In addition to focusing on druggability information, we also sought to harmonize efforts that characterized regulatory sequences across the genome in ORegAnno 3.0 (Chapter 4). Using information from this database, we designed a custom-capture reagent that allowed us to better sequence regulatory sequences, ultimately helping us identify 139 recurrently mutated promoter regions within breast cancer (Chapter 5). In order to identify splice-associated variants, we created RegTools and applied it to cancer samples from TCGA, additional clinical cohorts, mouse scRNA data, and long-read sequencing data from cell lines (Chapter 6). From this analysis, we were able to identify pan-cancer and cancer-type specific splice-associated variants while also showing that this class of variants can be identified in scRNA and long-read data.

7.1 Future directions for DGIdb

While substantial updates and improvements were made with the 3.0 and 4.0 updates to DGIdb, new druggability information continues to be discovered as well as potential new use cases for the information within the database. To address these issues, work is already underway for
DGIdb 5.0. In preparation for this update, we have examined the ways in which DGIdb was currently being utilized in research. We have explored the combined 1,428 citations across all our major releases and found that DGIdb was seeing great success as a tool for large-scale annotation and discovery experiments, such as those performed in M Nagel et al. (Nagel et al., 2018) and C Cava et al. (Cava et al., 2020). In addition to these reports, we identified some publications highlighting DGIdb’s potential for use in clinical research applications to profile drugs targeting genetic variations found in actual patient samples, such as those found in Hoefflin R et al. (Hoefflin et al., 2018) and de Witte CJ et al. (de Witte et al., 2020). In this setting, DGIdb could provide annotation of patient-derived gene sets for ‘druggability’ and clinical actionability using our gene categories tool. Similarly, the resource could be potentially invaluable due to its ability to return drug-gene interactions for patient-derived genes of interest with flagged approval when possible. These findings inform our decision-making as we consider additions to DGIdb’s feature-set.

The key focus of our version 5.0 update is to refine the platform for clinical research applications while maintaining the key features and functionality that have made DGIdb successful in the past. To this end, we have rewritten the entire code base from scratch, implemented extensive updates to our server architecture, and changed the way data is displayed on our HTML client. Previous versions of DGIdb made extensive use of the Ruby on Rails library to support the database and coordinate the presentation of data to end users via HTML and direct API calls to a REST interface. In this new version, we have separated DGIdb into two separate components: a front-end client using the React.js library, and a back-end Ruby on Rails server. These two components are connected by a GraphQL communications layer that allows our React client to retrieve and display the underlying data. The GraphQL layer also affords
greater flexibility for user-defined queries. While the previous REST architecture restricts the user to predefined searches, GraphQL permits user customization of return values and nested structures to provide more powerful and efficient lookups. Decoupling the client and server components also affords us greater flexibility in how the data is managed and presented to the user. This allows us to easily update DGIdb’s HTML interface and layout as needed independently of modifications to the backend server architecture.

The separation of DGIdb into two separate components also allows for greater flexibility in the design and presentation of our overall layout. By harnessing the well-supported React library, we have redesigned DGIdb’s HTML interface into a simpler and more interactive user experience. The search pages have been redesigned to be more straightforward to the user while maintaining the bulk query capacities many users are familiar with. Interaction results pages have been refined to make better use of space and highlight more clinically relevant data, including the highest level approval values and associated indications for drug interactions. The redesign of this space has allowed for the addition of a summary panel to provide at-a-glance metrics for entire search data sets. This summary panel allows users to quickly quantify the number, type, and directionality of interactions, and regulatory approval for drugs, with the goal of improving data presentation for clinical applications. The information shown on the interaction results pages for the previous versioning of DGIdb is still available for access on redesigned interior drug and gene record pages.

Additionally, for consideration in clinical applications, our regulatory approval reporting for drugs has been expanded for clarity and confidence purposes. Interaction results now highlight the highest-level regulatory approval values and associated indications for drugs involved in interactions. Individual drug pages now clearly show all reported approval values
and their original sourcing. Additionally, for US-based users, the database Drugs@FDA has been incorporated as an additional source to link normalized drug records to active, on-market new drug applications (NDA) and abbreviated new drug applications (ANDA) to reinforce the notion of approval status. Accordingly, individual drug pages now also highlight these active NDA/ANDA application references in a separate panel for clinical and interpretation pipeline consideration.

Lastly, we are updating our sources, database structure, and grouping methods for this latest update. Four sources have already received significant updates to their latest versioning: ChEMBL (v30), DrugBank (v5.1.9), PharmGKB, and CIViC (v2 API). All additional sources are having their importers updated to ensure compliance with the database’s new structural changes. With these changes so far, we have seen an increase in imported drug claims from 51,541 records in v.4.2.0 to 73,719 records in v.5.0 (+43.02%) (Figure 7.1).

![Figure 7.1 Comparison of record counts for drug claims, gene claims, and interaction claims between DGIdb version 4.2.0 and 5.0](image)

In addition to these updates, slight changes are being made to the underlying data structure, including removing five deprecated tables and adding two new tables related to approval reporting. To ensure the harmonization of drug and therapeutic concepts within DGIdb, modifications have been made to the internal grouping strategy implemented in previous
versions. All drug claims are now subject to normalization via a separate, standalone therapy normalization service. This change has increased normalized therapy concepts from 69.11% in v.4.2.0 to 89.28% in this update (Figure 7.2). Taken together, these changes translate to an increase in searchable records and data quality for the end user.

Figure 7.2 Comparison of normalization rates for drug claims, gene claims, and interaction claims between DGIdb version 4.2.0 and 5.0

7.2 Future directions for identifying non-coding regulatory mutations in breast cancer

While we were able to identify previously found and novel recurrent promoter region mutations, there is still much work to be done with regard to characterizing the genomic landscape beyond known mutations. In particular, more work is needed to characterize the mutations discussed within this thesis and to identify novel mutations such as splice-associated mutations, synonymous mutations that have currently unknown on gene expression, and other variants that could arise from the additional non-coding space not analyzed within this body of work. Due to
differences in expression among the different molecular subtypes, more work is needed to elucidate specific non-coding mutations that associate with specific subtypes of breast cancer.

Novel tools have recently been created that attempt to identify additional non-coding mutations from pan-cancer analyses (Dietlein et al., 2022; Rezaie et al., 2022). The functional impact of these mutations have not been fully characterized much less understood enough to be able to design novel therapeutics against. As sequencing costs decrease and larger datasets of WGS, custom-capture sequencing, RNA-Seq data are created, studies will eventually become more powered to detect novel variants, particularly those in non-coding regions of the genome, and be able to associate them with functional changes within the matched RNA data. The insights gained from such studies will be able to provide prognostic value to breast cancer patients and will be able to guide the design of new targeted therapeutics, particularly for the breast cancer subtypes that do not currently have targeted therapies.

7.3 Future directions based on RegTools
We have demonstrated the functionality of RegTools in identifying specific somatic mutations that lead to exon skipping, novel exon inclusion, alternative acceptor/donor site usage, and intron retention. Novel tumor-specific exon-exon junctions predicted by RegTools can be examined to predict the open reading frame (ORF) that will result given the observed alternative splicing event. For example, if an exon is skipped, one could determine the new open reading frame sequence from that point in the transcript onwards. In some cases, accurate prediction of the new ORF sequence will require either localized assembly of RNA reads surrounding the mutation or full-length assembly of the transcript. Due to the short nature of the cDNA fragments sequenced in RNA-seq experiments and the transcriptome-wide “shotgun” nature of RNA-seq experiments, de novo transcriptome assembly is computationally expensive and error-prone. However,
RegTools provides a region of interest because it identifies a somatic mutation with some localized evidence of a “cis” splicing effect (e.g. a novel exon-exon junction). This focus would allow one to explore more computationally expensive approaches that involve more accurate local RNA assembly at candidate loci. For example, one could identify reads with similarity to the cis-altered locus using an efficient k-mer based approach such as Kallisto and then perform an assembly of these reads using a de novo assembler such as Trinity to predict full-length isoforms for a single locus (Grabherr et al., 2011). With full-length isoform sequences in hand, one could next predict which events are likely to lead to nonsense-mediated decay (NMD) (Hsu et al., 2017). Since NMD may be incomplete, one should retain some of these predictions to explore whether useful neoepitopes from apparently expressed alternative isoforms can be present in spite of an NMD prediction (Hug et al., 2016). Additional complexity arises from the need to determine how much of the alternative isoform’s ORF is new relative to the wild-type transcript sequence. For example, a somatic mutation that causes the skipping of an exon, in which the reading frame is maintained, will lead to a protein sequence with limited novel peptide junction sequences (e.g. nonamers spanning the joined exons). By contrast, a somatic mutation that causes a frameshifting exon skip or alternative splice site usage could lead to a considerable novel protein sequence downstream of the aberrant splicing event. Once the full ORF of the isoform is determined, one could extract the portion that is predicted to differ from the wild-type protein sequence. This portion can be subjected to neoepitope prediction using pVACseq developed in the Griffith lab (https://github.com/griffithlab/pVACtools) (Hundal et al., 2016). HLA typing is needed for neoepitope prediction and can be performed using tools such as Optitype (Szolek et al., 2014) or xHLA (Xie et al., 2017).
7.4 Summary

In conclusion, the projects summarized in this dissertation address some aspects of the precision oncology bottleneck. This dissertation attempts to address these issues by describing databases that harmonize druggability and regulatory studies, analysis of regulatory non-coding variants in breast cancer, and the pan-cancer identification of splice-associated variants using RegTools. Due to the nature of cancer as a disease, individual tumors are characterized by hundreds or thousands of individual alterations, leading to a growing deluge of measurable tumor-specific molecular alterations for clinical relevance. However, only a subset of these drivers are currently clinically actionable. Previous studies have focused almost entirely on protein-altering mutations for diagnostic, prognostic, and predictive information. However, work done within this dissertation reveals that the next phase of precision oncology requires an extension to those alterations that affect non-coding, splicing, or other kinds of alterations, as well as the creation of tools that harmonize existing disease-relevant data for ease of interpretability. It should be reiterated that such tools and the efforts surrounding their development and maintenance should be open-source to allow for the advancement of science while providing reproducibility and transparency. It is the hope that work from this dissertation and studies built off of the foundation that it has laid will help to alleviate the precision oncology bottleneck while also furthering our understanding of the somatic alterations that drive tumor biology. With persistence, the accomplishment of one or both of those goals may lead to the eventual development of novel therapeutics that increase survival rates of cancer, improve the quality of life for cancer patients, or aid in the early detection/prevention of cancer.
Supplementary Figure 1. Benchmarking of each RegTools command. The total CPU time (System Time + User Time) and real time are plotted against the number of entries processed for each available RegTools function using 10 total replicates. For the cis-splice-effects identify/cis-splice-effects associate/variants annotate workflows, the number of entries corresponds to the number of somatic variants, whereas the number of entries in the junctions extract/junctions annotate/compare_junctions workflows corresponds to the number of reads processed from a downsampled BAM file, the number of junctions processed, and the number of candidate variant junction pairings processed, respectively. For compare_junctions, candidate variant junction pairings were compared across the number of samples in that cohort, with the largest being 1,022 samples that comprise our BRCA cohort. LOESS curves are fitted onto each plot.
Supplementary Figure 2. Overview of input data considered and significant events identified by RegTools for each tumor type. A) Summary of initial variants considered for analysis by RegTools per sample per tumor cohort. Each sample’s variant count is plotted and violin plots are overlaid for each cohort. B) Summary of unique exon-exon junction observations for each sample. Each sample’s unique junction count is plotted and violin plots are overlaid for each cohort. C) Summary of significant junction types for each cohort across each of the variant window sizes that were used in this analysis.
Supplementary Figure 3. Summary of variants analyzed by RegTools in each tumor cohort. Summary of the starting number of high-quality variants per sample, the number of initial variants considered for analysis by RegTools for each variant window used per tumor cohort, and the number of significant variants for each variant window used per tumor cohort.
Supplementary Figure 4. Visualization of junctions across cohorts. Summary of the total junction read counts, unique junctions (all types), unique known (DA) junctions, unique known (DA) junctions not found in GTEx, unique D, A, NDA junctions, and unique D, A, NDA junctions not found in GTEx per sample per cohort.
Supplementary Figure 5. Benchmarking of RegTools and SAVNet. The total CPU time (System Time + User Time) and real time are plotted against both RegTools and SAVNet for fifty LUAD samples from TCGA. Each sample is represented as a dotplot and a half violin plot is plotted alongside the dotplot to show the distribution of runtimes.
Supplementary Figure 6: Intronic SNV in CTTN associated with an exon skipping event. A) IGV snapshot of a single nucleotide variant (GRCh38, chr11:g.70407517G>C) within an intron of CTTN in LUAD sample TCGA-86-6851-01A. This variant is associated with an exon skipping event causing the formation of an NDA junction, JUNC00027688, which has 44 reads of support. The variant was identified by RegTools, VEP, and Veridical but no other tools. This result was found using the default splice variant window parameter (i2e3). B) Sashimi plot visualization of the novel junction.
Supplementary Figure 7: Exonic SNV in LZTR1 associated with alternative donor usage. 
A) IGV snapshot of a single nucleotide variant (GRCh38, chr22:g.20995026G>C) within an exon of LZTR1 in LUAD sample TCGA-38-4631-01A. This variant is associated with the formation of an A junction, JUNC00075013, which has 49 reads of support. The variant was identified by RegTools, VEP, and SpliceAI but no other tools. This result was found using the default splice variant window parameter (i2e3). B) Sashimi plot visualization of the novel junction.
Supplementary Figure 8: Intronic SNV in *Bin1* associated with an alternate splice donor event by scRNA-seq analysis. A) Schematic of a single nucleotide variant (mm10, chr18:g.32432427T>C) within an intron of Bin1. This variant is significantly associated with an exon skipping event causing the formation of an NDA junction. This result was found using the default splice variant window parameter (i2e3). B) UMAP projection of single cells from MCB6C organoid derived tumors with high confidence tumor cells (orange) and high confidence normal cells (blue) highlighted. C) UMAP projection of single cells from MCB6C organoid derived tumors overlaid with Log2 expression of Bin1. D) Zoomed view of cells containing the Bin1 alternate donor event. E) Violin plot comparing the normalized junction score of the novel alternate donor event in cells with and without the variant.
Supplementary Figure 9. Pan-cancer analysis of cohorts from TCGA and MGI reveals genes recurrently disrupted by variants which promote splicing of particular canonical junctions. Results of analysis for recurrently disrupted genes in each TCGA and MGI cohort. A) Rows correspond to the 40 most frequently recurring genes, as ranked by binomial p-value. Genes are clustered by whether they were annotated by the CGC as an oncogene (red), an oncogene and tumor suppressor gene (yellow), a tumor suppressor gene (green), or another type of cancer-relevant gene (black, bold). Shading corresponds to $-\log_{10}(p$ value) and columns represent cancer types. Blue marks within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort. B) Rows correspond to the 40 most frequently recurring genes, as ranked by fraction of samples. Shading corresponds to the fraction of samples and columns represent cancer types. Blue dots within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort.
Supplementary Figure 10. TCGA pan-cancer analysis reveals genes recurrently disrupted by variants which cause non-canonical splicing patterns. Results of analysis for recurrently disrupted genes in each TCGA cohort. A) Rows correspond to the 40 most frequently recurring genes, as ranked by binomial p-value. Genes are clustered by whether they were annotated by the CGC as an oncogene (red), an oncogene and tumor suppressor gene (yellow), a tumor suppressor gene (green), or another type of cancer-relevant gene (black, bold). Shading corresponds to −log10(p value) and columns represent cancer types. Blue marks within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort. B) Rows correspond to the 40 most frequently recurring genes, as ranked by fraction of samples. Shading corresponds to the fraction of samples and columns represent cancer types. Blue dots within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort.
Supplementary Figure 11. TCGA pan-cancer analysis reveals genes recurrently disrupted by variants which promote splicing of particular canonical junctions. Results of analysis for recurrently disrupted genes in each TCGA cohort. A) Rows correspond to the 40 most frequently recurring genes, as ranked by binomial p-value. Genes are clustered by whether they were annotated by the CGC as an oncogene (red), an oncogene and tumor suppressor gene (yellow), a tumor suppressor gene (green), or another type of cancer-relevant gene (black, bold). Shading corresponds to $-\log_{10}(p$-value$)$ and columns represent cancer types. Blue marks within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort. B) Rows correspond to the 40 most frequently recurring genes, as ranked by fraction of samples. Shading corresponds to the fraction of samples and columns represent cancer types. Blue dots within cells indicate that the gene was annotated by CHASMplus as a driver within a given TCGA cohort.
Supplementary Figure 12. Analysis of MGI cohorts reveals genes recurrently disrupted by variants which cause non-canonical splicing patterns. Results of analysis for recurrently disrupted genes in each MGI cohort. A) Rows correspond to the 40 most frequently recurring genes, as ranked by binomial p-value. Genes are clustered by whether they were annotated by the CGC as an oncogene (red), an oncogene and tumor suppressor gene (yellow), a tumor suppressor gene (green), or another type of cancer-relevant gene (black, bold). Shading corresponds to $-\log_{10}(p\text{ value})$ and columns represent cancer types. B) Rows correspond to the 3 most frequently recurring genes, as ranked by the fraction of samples. Shading corresponds to the fraction of samples and columns represent cancer types.
Supplementary Figure 13. Analysis of MGI cohorts reveals genes recurrently disrupted by variants which promote splicing of particular canonical junctions. Results of analysis for recurrently disrupted genes in each MGI cohort. A) Rows correspond to the 4 most frequently recurring genes, as ranked by binomial p-value. Genes are clustered by whether they were annotated by the CGC as an oncogene (red), an oncogene and tumor suppressor gene (yellow), a tumor suppressor gene (green), or another type of cancer-relevant gene (black, bold). Shading corresponds to $-\log_{10}(p$ value) and columns represent cancer types. B) Rows correspond to the 4 most frequently recurring genes, as ranked by the fraction of samples. Shading corresponds to the fraction of samples and columns represent cancer types.
Supplementary Figure 14: Intronic SNV in TP53 associated with alternative donor usage.

A) IGV snapshot of a single nucleotide variant (GRCh38, chr17:g.7673609C>A) within an intron of TP53 in an OSCC sample. This variant is associated with an exon skipping event with 23 reads of support and an alternate acceptor site usage with 41 reads of support. This result was found using the default splice variant window parameter (i2e3). B) Sashimi plot visualization of the novel junction.
Supplementary Figure 15: Intronic deletion in RNF145 associated with alternative donor usage. **A)** IGV snapshot of a single nucleotide variant (GRCh38, chr5:g.159169058delA) within an intron of RNF145 in COAD samples. This variant is associated with an exon skipping event with 8 and 6 reads of support for the samples shown. This result was found using the default splice variant window parameter (i2e3). **B)** Sashimi plot visualization of the novel junction.
Supplementary Figure 16: Several SNVs in CDKN2A associated with alternate donor usage.

A) IGV snapshot of three variant positions in CDKN2A found to be associated with usage of an alternate donor site that leads to the formation of an alternate known transcript. This result was found using the default splice variant window parameter (i2e3) for known (DA) junctions. B) Zoomed in view of the variants identified by RegTools that are associated with alternate donor usage. Two of these variant positions flank the donor site that is no longer being used. C) Sashimi plot visualizations for samples containing the identified variants that show alternate donor usage.
References


Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S.,


162


Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D.,


Mardis, E.R. (2010). The 1,000 genome, the 100,000 analysis? Genome Med. 2, 84.


McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,


Nielsen, T.O., Parker, J.S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S.R., Snider,


President’s Council of Advisors on Science (2008). Priorities for Personalized Medicine, Report of the President’s Council of Advisors on Science and Technology, September 2008. President’s Council of Advisors on Science and Technology (U.S.).


Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-


level bioinformatics resource for studying the mechanism and regulation of autophagy. Autophagy 11, 155–165.


Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K.,


