INVESTIGATING DISEASE MECHANISMS AND DRUG RESPONSE DIFFERENCES IN TRANSCRIPTOMICS SEQUENCING DATA

Edward Ronald Simpson Jr.

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

January 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Yunlong Liu, PhD, Chair

Doctoral Committee

Sarath Janga, PhD

November 3, 2021

Jun Wan, PhD

Huanmei Wu, PhD

Jingwen Yan, PhD

ACKNOWLEDGEMENT

Foremost, I want to express my appreciation for my advisor Dr. Yunlong Liu. Thank you for granting me a seat in your laboratory where I have been able to learn from your example. You taught me how to critically evaluate results, write clearly and directly, and utilize resources to forge a path where none exists. I wish to thank Dr. Richard Kuhn and Dr. Huanmei Wu for their great help when I was struggling; Each of whom granted me a second chance and I fear where I would be without them. Additionally, I would like to thank my committee members Dr. Sarath Janga, Dr. Jingwen Yan and Dr. Jun Wan as well as the incredibly generous editor Dr. Jill Reiter for their hard work and valuable comments. Furthermore, thank you to all faculty members at the School of Informatics and Computing for their excellent mentorship.

Next, I want to mention the people that instilled in me the qualities necessary to become a scientist: My mother, for demonstrating enthusiasm when I asked questions or wanted to try something new, for having the humility to admit when she did not know an answer and guiding me to search for it myself, for having the patience to be a child's research assistant, and for dedicating so much of her time, personal and financial resources to developing my scientific interests; My father, for demonstrating the uncompromisable honesty necessary for objective observation, for teaching me the value of following a protocol and doing a job with precision, and for pushing me to confront difficulties and achieve independence. Finally, I want to thank my wife Victoria for her unwavering friendship and personal sacrifice. Without her caring for me and our two boys I would not have had the financial freedom or personal time required to complete my degree.

Edward Ronald Simpson Jr.

INVESTIGATING DISEASE MECHANISMS AND DRUG RESPONSE DIFFERENCES IN TRANSCRIPTOMICS SEQUENCING DATA

In eukaryotes, genetic information is encoded by DNA, transcribed to precursor messenger RNA (pre-mRNA), processed into mature messenger RNA (mRNA), and translated into functional proteins. Splicing of pre-mRNA is an important epigenetic process that alters the function of proteins through modifying the exon structure of mature mRNA transcripts and is known to greatly contribute to diversity of the human proteome. The vast majority of human genes are expressed through multiple transcript isoforms. Expression of genes through splicing of pre-mRNA plays crucial roles in cellular development, identity, and processes. Both the identity of genes selected for transcription and the specific transcript isoforms that are expressed are essential for normal cellular function. Deviations in gene expression or isoform proportion can be an indication or the cause of disease.

RNA sequencing (RNAseq) is a high-throughput next-generation sequencing technology that allows for the interrogation of gene expression on a massive scale. RNAseq generates short sequences that reflect pieces of mRNAs present in a sample. RNAseq can therefore be used to explore differences in gene expression, reveal transcript isoform identities and compare changes in isoform proportions. In this dissertation, I design and apply advanced analysis techniques to RNAseq, phenotypic and drug response data to investigate disease mechanisms and drug sensitivity.

Research Goals: The work described in this dissertation accomplishes 4 aims. Aim 1) Evaluate the gene expression signature of concussion in collegiate athletes and identify potential biomarkers for response and recovery. Aim 2) Implement a machine-learning algorithm to determine if splicing can predict drug response in cancer cell lines. Aim 3) Design a fast, scalable method to identify differentially spliced events related to cancer drug response. Aim 4) Construct a drug-splicing network and use a systems biology approach to search for similarities in underlying splicing events.

Yunlong Liu, PhD, Chair

Sarath Janga, PhD

Jun Wan, PhD

Huanmei Wu, PhD

Jingwen Yan, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADP	Adenosine Di-phosphate
AKT	Protein Kinase B
ALS	Amyotrophic Lateral Sclerosis
ATP	Adenosine Tri-phosphate
AUC	Area Under the Curve (or Concentration Response Curve)
BAM	Binary Sequence Alignment Map
BWA-MEM	Burrows-Wheeler Aligner Measure Memory
CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary Deoxyribonucleic Acid
CI	Confidence Interval
CIGAR	Concise Idiosyncratic Gapped Alignment Report
CNS	Central Nervous System
CNV	Copy Number Variants
CPM	Counts Per Million
CTRP	Cancer Therapeutic Response Portal
DGE	Differential Gene Expression
DMD	Duchenne Muscular Dystrophy
DNA	Deoxyribonucleic Acid
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EMT	Epithelial-mesenchymal Transition
ERK	Extracellular Signal-regulated Kinase
ESTs	Expressed Sequence Tags
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
FPKM	Fragments Per Kilobase Per Million
GDSC	Genomics of Drug Sensitivity in Cancer Database
GLM	Generalized Linear Model
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GTF	Gene Transfer Format
GWAS	Genome-wide Association Study
HGPS	Hutchinson-Gilford Progeria Syndrome
hnRNP	Heterogeneous Ribonucleoprotein
IL	Interleukin
IU	Indiana University
JAK	Janus Tyrosine Kinase
JM	Juxtamembrane Protein Domain
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAPQ	Mapping Quality
MDS	Multi-dimensional Scaling
ME	Mutually Exclusive Splicing Event
MEK	Mitogen Activated Kinase Protein

miRNA	Micro Ribonucleic Acid
MISO	Mixture of Isoforms Software
mRNA	Messenger RNA
NAMPT	Nicotinamide Phosphoribosyltransferase
NB	Negative Binomial
NCBI	National Center for Biotechnology Information
NES	Normalized Enrichment Score
NF-kB	Nuclear Factor Kappa Beta
NOD	Nucleotide-binding and Oligomerization Domain
PC	Principal Component
PCA	Principal Component Analysis
PI3K	Phosphoinositide 3-kinase
PKC	Protein Kinase C
Pre-mRNA	Precursor Messenger RNA
PSI	Percent Spliced In
QBGLM	Quasi-binomial Generalized Linear Model
RBP	RNA-binding Protein
RI	Retained Intron Splicing Event
rMATS	Replicate Multivariate Analysis of Transcript Splicing
RNA	Ribonucleic Acid
RNAseq	Shotgun RNA Sequencing
ROC	Receiving Operator Characteristic
RP	Retinitis Pigmentosa
RPKM	Reads Per Kilobase Per Million
SAM	Sequence Alignment Map
SE	Skipped Exon Splicing Event
SMA	Spinal Muscular Atrophy
SMART	Simple Modular Architecture Research Tool Database
SNP	Single Nucleotide Polymorphism
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonucleoprotein (or Complex)
SNV	Single Nucleotide Variant
sQTLs	Splicing Quantitative Trait Loci
SR	Serine-arginine Rich Proteins
STAR	Spliced Transcripts Alignment to a Reference Tool
STAT	Signal Transducer and Activator of Transcription Protein
TGF-beta	Transforming Growth Factor Beta
TMM	Trimmed Mean of M-values
TN	True Negative
TNFa	Tumor Necrosis Factor Alpha
TP	True Positive
TPM	Transcripts Per Million
UCSC	University of California Santa Cruise

Chapter 1 Introduction

1.1 Background

High-throughput experimental approaches like shotgun transcriptome sequencing provide a wealth of comprehensive data that can be used to answer challenging and complex biological questions. Transcriptomic sequencing data gives researchers insight into the regulation and production of gene transcripts; coded genetic elements that carry important cellular information, control cellular fate, and coordinate a cells response to stimuli. Transcriptomic sequencing data can be analyzed in numerous ways and many analysis strategies offer powerful approaches to studying disease.

Transcriptomic analysis is informative for understanding disease response and recovery. In physiological diseases, like concussion, the transcriptome can indicate the damage that has been sustained, the biological processes that are activated following injury, and give insight into the stages of healing. In hereditary or acquired genetic diseases, transcriptomic analysis can help researchers understand the origin of disease and plan the best course of treatment.

Possibly the most common analysis approach for transcriptomic data is differential gene expression analysis. Differential gene expression analysis compares the quantity of messenger-ribonucleic acid (mRNA) gene products in two or more groups of samples and is used to identify differences in gene expression across multiple conditions or time.

Transcriptomic data can also be used to study splicing of precursor messenger ribo nucleic acid (pre-mRNA), an important step in the production of gene transcripts that determines the combination of sequence elements included in mature mRNA. Differential splicing contributes to protein diversity in eukaryotic cells by generating many unique transcript isoforms from a single gene sequence [1]. As much as 95% of human multi-exon genes are impacted by differential splicing [2]. The splicing of specific transcript isoforms can determine cellular identity, alter cellular processes and contribute to disease. The splicing of transcript isoforms, regulation of splicing and the functional consequences from splicing are of major scientific interest.

Cancer, an uncontrolled cellular proliferative disease of the genome, has long been the second leading cause of death in the United States [3]. Splicing has been linked to cancer in a number of ways. Splicing helps direct many key cellular processes that are altered in cancers, such as epithelial-mesenchymal transition, and splicing has been shown to impact every major cancer hallmark leading to development and maintenance of tumor transformation [4]. The splicing of certain transcript isoforms has also been shown to influence cancer drug metabolism [5,6]. It is likely splicing influences drug response in cancer in a number of ways, but so far very few studies have explored associations between drug response and cancer. A comprehensive study showing that splicing influences drug response would increase our understanding of cancer-related processes and spur researchers to investigate new therapeutic strategies for treating cancer.

While the association between certain differentially spliced transcripts and disease has been established, relatively little is known about the numerous transcripts differential splicing enables. Historically, analysis of transcript isoforms was a painfully slow process accomplished by systematically manipulating a single isoform at a time. Now, high-throughput transcriptome sequencing can summarize isoforms present in a sample within one to two days. The cost of high-throughput techniques continues to decrease, and study sizes are increasing as a result [7]. However, existing isoform analysis strategies were designed with small sample sizes in mind and are often not compatible with high-volume data sets. No currently available analysis tool can analyze differential splicing across hundreds to thousands of samples, but it will be necessary to develop a method that can manage large datasets using low computational resources and perform analysis in minimal time to find new discoveries in large data sets.

While cancer treatment strategies differ by cancer tissue type, cancer treatment strategies typically involve the use of combination therapy [8]. Combining multiple drugs can boost the effectiveness of treatment, reduce negative side-effects and decrease rate of relapse [8]. Treatment regimens must be tailored in a way such that drugs have a complementary and additive effect against the cancer rather than enhanced toxicity to the patient. When a cancer does relapse, tumors have typically acquired immunity to previously used drugs and alternative therapeutics targeting the cancer must be selected [9]. It is common for patients with late-stage terminal illness to be considered for clinical trials were new therapeutics, off-indication therapeutics or non-standard drug combinations are tested [10]. Identifying the tissue-specific differentially spliced exons mediating drug activity

and investigating the network of influential genes across tissue types would inform a more comprehensive treatment strategy and potentially uncover new effective drug combinations.

In this dissertation my goals are to: 1. Evaluate the gene expression signature of concussion in collegiate athletes and identify potential biomarkers for response and recovery. 2. Implement a machine-learning algorithm to determine if splicing can predict drug response in cancer cell lines. 3. Design a fast, scalable method to identify differentially spliced events related to cancer drug response. 4. Construct a drug-splicing network and use a systems biology approach to search for similarities in response related splicing events.

1.2 Investigating Differences in Gene Expression Triggered by Concussion

1.2.1 Significance

Concussion is a common type of mild traumatic brain injury and while concussions in athletics gain a lot of attention, they also occur frequently in day-to-day life. Concussions occur in individuals of all ages from activities such as work, car accidents, recreation and household falls **[11]**. Concussion is diagnosed as a temporary impairment in normal brain function as a result of physical trauma and can go undiagnosed due to the inconsistency and range of symptoms **[11]**. Generally, individuals who have sustained a concussion recover within two weeks, but some experience symptoms lasting a month or longer **[12]**. To date, there is no explanation for the extended recovery period some individuals experience.

Recognizing the symptoms and intervening in hazardous activity is currently the recommended treatment strategy for concussions **[13]**. Sustaining a second impact can greatly exacerbate the injury, leading to swelling, delayed brain degeneration and even death **[14]**. Because of the importance of rapid diagnosis, researchers have attempted to develop fast diagnostic tests. Methods based on brain imaging and blood-based protein or molecular biomarkers have so far failed to yield a robust target for concussion diagnosis. Identification of an effective and reliable concussion biomarker would enhance the objectivity and speed of diagnosis, thereby improving patient outcome and reducing the chances of sustaining compounding injuries.

1.2.2 Critical Need

Despite years of research into concussions and mild traumatic brain injury, key biological factors related to response and recovery are poorly understood. Due to physiological and temporal limitations, biological details surrounding concussion are difficult to study. Primary research in concussion is typically limited to samples from animal models with induced mild traumatic brain injury. A few studies have been conducted with human samples, however studies have been small in size . Without a clearer understanding of the human biological response and factors underlying recovery, identification of concussion biomarkers and development of informed therapeutic approaches will not be possible.

1.2.3 Innovation

Whole blood ribonucleic acid (RNA) sequencing data from over five hundred collegiate athletes, spanning multiple time points before and after injury, was collected by the Concussion Assessment, Research and Education Consortium. Concussed athletes will be compared to control participants to identify differentially expressed genes related to concussion response and recovery. Analysis of this dataset will be the largest concussion study to date. Additionally, whole transcriptome analysis of RNA sequencing data has the potential to reveal new information that brain imaging and blood-based biomarkers have failed to detect. Finally, new techniques need to be developed to handle the complex analysis between multiple groups with paired pre-injury samples. These new methods will serve as an example for researchers to follow and aid in guiding analysis of large concussion datasets.

1.3 Predicting Drug Response with Differential Splicing Profiles

1.3.1 Significance

The primary goal of genomics-based cancer research is to identify genetic and molecular features that can improve clinical outcome. Drug response prediction models are valuable for identifying influential predictive features and exploring treatment options [15]. Because each individuals' genome, and therefore cancer, is unique the ability to tailor a treatment regimen to an individual is expected to improve efficacy and reduce side effects [15]. Predictive modeling is a key component of precision medicine will allow the most effective therapy to be selected by using the patients' own genetic information [15].

As discussed in the background, splicing has a major influence on the expression of gene products. Predictive models usually include gene expression as a primary feature type. Despite the frequent use of expression data models typically ignore the fact that genes are expressed through specific isoforms. Inclusion or exclusion of individual exons in gene transcripts can alter protein functionality. Studies evaluating predictive models have found that increasing the variety of data types used in predictive models is essential to increase performance [15–17]. Any predictive method that doesn't incorporate isoform-specific information is not representative of true gene complexity. Building and evaluating predictive models with splicing data will help researchers evaluate the relevance of splicing in drug response, identify new treatments expression-based models cannot and provide a method to build more complex predictive models that integrate multiple data types.

1.3.2 Critical Obstacles

Currently there are numerous tools available for quantifying splicing of whole isoforms and individual exons, but no standard method for drug response prediction with splicing data exists. New methods specific to splicing data need to be developed that can properly transform the data to be compatible with machine learning techniques. As splicing expands the number of transcripts in the transcriptome it also expands the complexity of data used in predictive modeling. When building predictive models, it is important to consider the number of features used and any correlation among features to prevent overfitting [18]. The model construction process must be able to select only the most informative and unique features while removing features that introduce noise. These goals will require a fast, efficient modeling algorithm optimized for splicing data.

Since the ultimate goal of drug response modeling is to identify biologically relevant targets, it is essential that any model designed to predict outcomes from biological data be interpretable. The identities of exons responsible for predictions must be traceable. Previous work has shown that not all differentially spliced exons change the structure or function of proteins, and while predictive models can be designed to report the major features responsible for predictions the selected exons may not be truly important for drug response [19]. Following extraction of relevant features, methods for annotation and evaluation must be applied.

1.3.3 Innovation

One major difficulty in investigating the role of splicing in drug response is the lack of a dataset large enough to train accurate predictive models. Additionally, previous work investigating splicing in drug response was done in only a few specific cancer types. In this study, two large data sets from public databases were integrated to maximize the number of samples and cancer types explored. The Cancer Cell Line Encyclopedia (CCLE) is a public resource supported by the Broad Institute and the Novartis Institute for Biomedical Research that provides access to detailed genomic data from over 1,100 cancer cell lines [20]. While CCLE hosts a large volume of genomic data, it only provides pharmacologic data for a small number of drugs. The Cancer Therapeutic Response Portal (CTRP) on the other hand, also a product of the Broad Institute, is a public resource that provides access to pharmacologic data spanning hundreds of compounds in close to 1,000 cell lines [21]. We overlapped data between the CCLE and CTRP, searching for cell lines common to both resources, to create a dataset with pharmacologic measurements in 501 compounds and pre-treatment RNAseq data for up to 850 cell lines per compound. To our knowledge, the dataset used in this study is the largest integration of splicing and drug response data to date.

Because very few studies have explored predictive models using splicing data, and the data in this study is from a unique integrated dataset, it is necessary to produce a new model design and training/testing pipeline. It is also necessary to compare the splicing-based model performance to a known standard so that the general impact of splicing and potential benefits from integrating splicing in existing models can be quantified.

Therefore, we construct three models using the same base algorithm: 1. A splicing-data only predictive model, 2. A gene expression only predictive model, and 3. A combined splicing- and expression-based model that integrates multi-omic data. To our knowledge, this will be the largest study of its design, spanning a considerable number of drugs, and the first major characterization of the influence of splicing in a variety of cancer types.

1.4 Differential Splicing Analysis of Large-scale Data Sets

1.4.1 Significance

Although splicing of certain gene isoforms is known to influence drug response, little is known about the multitude of additional isoforms genes produce and their potential impact. It is likely that a number of differentially spliced exons change the way cells respond to drugs, and as such contribute to drug resistance. A comprehensive study investigating differentially spliced exons between pre-treatment transcriptomic profiles of cell lines categorized by drug sensitivity will help identify baseline splicing differences that change the effectiveness of drugs following treatment. Finding exons related to drug response will lead to research in new potential drug targets, enable investigation of upstream splicing regulation through proximal sequence information and expand the cancer therapeutic knowledgebase.

As stated in the background, existing methods for differential splicing analysis are slow and designed for small sample sizes. The dataset in this study, like many datasets to be generated in the near future, is large and complex. A method capable of analyzing high-volume data in short periods of time with few computational resources will be an

extremely valuable tool for splicing research and will be necessary to keep pace with the volume of high-throughput data generated by more economic and comprehensive studies.

1.4.2 Critical Obstacles

A significant challenge that complicates a next-generation sequencing study utilizing large numbers of samples is processing the vast quantity of associated raw sequencing data. Tens of millions of reads are generated for each sample, all of which must be aligned to a reference genome. After obtaining genomic coordinates for the reads, the aligned data must be searched and counted. Splicing of transcripts is a combinatorial problem. A single gene sequence could potentially produce an isoform for each unique combination of exons. A dictionary of known reference isoforms can be used to reduce the complexity of splicing quantification. Even when ignoring unannotated isoforms, reads must be counted for over 40,000 unique combinations of genomic positions in each sample. The dataset in this study has over 800 samples that will need to be individually processed before differential splicing analysis can be performed. To accomplish the processing of the study dataset a combination of custom scripts and publicly available tools, as well as the vast computations resources of the Indiana University (IU) supercomputing system, will be required.

Since few large-scale studies have been done with splicing data, a new model considering the format and properties of splicing data is required for downstream analysis. As mentioned above, counting of supporting reads in exon-centric approaches involves searching tens of thousands of positions in each sample. After counting, existing methods

estimate the percentage of isoforms and confidence interval for each combination of upstream, target and downstream exon combination. Then, differential testing is performed with a separate statistical framework. The new differential comparison method needs to be more efficient. Generalized logistic modeling is capable of using read count data and modeling the data distribution directly, avoiding the intermediate estimation procedure. I will use generalized linear modeling with quasi-binomial distribution to dramatically reduce the time and computational resources needed for analysis while retaining the data-specific requirements for comparing splicing data.

1.4.3 Innovation

Existing differential splicing analysis techniques are relatively slow, resource hungry tools that are designed for studies with small sample sizes. The method proposed in this study is a fast, efficient and scalable method for differential splicing analysis built on generalized linear modeling. Introducing a powerful new technique for differential splicing analysis will enhance the speed and reduce the computational burden of splicing analysis, allowing for much larger datasets to be processed in future studies. Additionally, this study represents the first occasion differential splicing analysis will be done with a large number of samples in a variety of cancer types. New associations between exons and cancer drug response will be identified. I will also investigate the cis-regulatory elements and functional consequences from differentially spliced exons, which previous whole-isoform studies could not do. The findings from this study can be used to identify potential new drug targets, improve the efficacy of cancer treatment regimens and improve our biological understanding of differential splicing.

1.5 Network Analysis of Drug Splicing Data

1.5.1 Significance

Drug treatment strategies are extremely important in cancer care and are often employed with surgical intervention or independently as the sole course of treatment [22]. Although cancer therapeutics can substantially prolong patient survival, cancer frequently becomes resistant to drugs after prolonged exposure [9]. Cancer drug resistance then leads to cancer relapse, aggressive disease and patient death [9]. Partially to combat drug resistance, drug treatment strategies in cancer usually involve multiple compounds [8]. Drug combinations can have additive effects, enhancing either the therapeutic efficacy or toxicity the patient experiences [8]. Drug combinations can also be neutral, providing no real benefit to the patient. Identifying effective new drug combinations could enhance the therapeutic outcome of cancer treatment regimens, reduce toxicity and prolong or prevent relapse.

The therapeutic landscape of differentially spliced exons remains largely unexplored. It is likely that a trove of information related to the influence of differentially spliced exons on drug sensitivity is waiting to be discovered. A large-scale, comparative analysis solution dependent on splicing data will allow researchers to link the splicing-based resistance profiles of drugs. Identifying communities of drugs with similar resistance profiles, and revealing the hidden relationship between splicing and drug response, could improve existing treatment regimens and lead to new complementary drug combinations.

1.5.2 Critical Obstacles

As mentioned in previous sections, the influence of splicing on cancer drug response has been rarely studied. An analysis approach specific to the challenges of splicing data will need to be developed. Hundreds of individual compounds and the differentially spliced exons associated with drug response will need to be evaluated for similarities.

To overcome the combinatorial complexity of pairwise drug-drug comparisons, a network-based approach will be used. Analysis techniques specific to the network design and structure will need to be developed.

1.5.3 Innovation

To date, there has never been a study looking at the commonality between splicing signatures influencing drug response across multiple compounds, or the possible splicing interactions drugs may have when used in combination. The total number of differentially spliced exons related to drug sensitivity is not known. Whether or not spliced exons can influence response to multiple drugs is unclear. Additionally, the potential for drug interaction through intersecting differentially spliced exons has never been explored.

The analytical strategies and knowledge generated by a comprehensive network analysis with splicing data could be extremely valuable for both the research and cancer treatment communities. This study will lay out a process for extracting disease-specific drug-splicing information that can be used to plan complementary drug treatment regimens and will have an impact beyond cancer. It will also provide biological insight into

mechanisms behind drug resistance and the importance of pre-treatment splicing profiles in anti-cancer drug response.

1.6 Objectives

The primary objective of my dissertation is to investigate the applications of transcriptomic sequencing data to investigate disease. First, I use differential gene expression and deconvolution analysis to describe the biological response to concussion. Second, I develop a machine-learning model that demonstrates the relevance of differentially spliced exons in cancer drug response. Third, I describe the implementation of quasi-binomial generalized linear modeling to uncover differentially spliced exons related to drug response. Fourth, I develop a network modeling approach to identify key differentially spliced exons that influence drug response in multiple drugs.

Chapter 2 reviews the current knowledge of gene expression, splicing, cancer treatment strategies and computational analysis methodology used in splicing analysis. Topics covered include the biology underlying splicing, the impact of splicing, and an overview of cancer biology and cancer treatment. Additionally, techniques commonly used to analyze changes in splicing are reviewed such as high-throughput genomic experiments, machine learning, differential splicing analysis methods and network analysis.

Chapter 3 describes differential gene expression analysis in a cohort of collegiate athletes. I detail differentially expressed genes relevant to the pathophysiology of concussion, the activation of immune response pathways through gene expression and

changes in response-related pathways over the course of recovery. I also introduce a method for differential analysis of deconvolution data while considering a previously obtained baseline sample. I find that transcriptomic data mirrors the pathophysiological response in damaged tissue and that there are large differences in gene expression immediately after concussion, but the differences quickly dissipate following activation of the immune response. I also find trends in immune signaling pathways across timepoints and discover differences in immune cell type proportions between injured and control participants.

Chapter 4 describes the construction of a predictive model trained to predict drug response using splicing data. I explain a common approach to quantify splicing in high-throughput data, use exon-centric measures from MISO to explore trends in splicing related to drug response and apply machine learning techniques to build an elastic net logistic regression model. I find that splicing can be used to predict drug response in cancer cell lines for the vast majority of cancer compounds and repeated modeling shows strong predictive power similar to that of expression data.

Chapter 5 describes a differential splicing analysis technique developed to process large volumes of data quickly using limited resources. I explain the design and advantages of quasi-binomial generalized linear modeling, identify differentially spliced exons that contribute to changes in drug response in cancer cell lines and perform RNA-binding protein enrichment analysis to uncover cis-acting sequence elements and trans-acting regulatory splicing factors modulating splicing. I also annotate exons with protein domain

and structural importance to connect differentially spliced exons to consequences in protein function. I show that quasi-binomial generalized linear modeling is capable of processing large volumes of data rapidly, and accurately identifies differentially spliced exons associated with drug-resistance strategies and functional protein domains.

Chapter 6 describes a network-based analysis that intersects differentially spliced exons modulating drug sensitivity across multiple compounds. I design a strategy to build a drug-drug splicing network and describe an algorithm that can be used to find modules of drugs that share exons related to drug response and discuss the mechanistic implications of the exons. I find that communities of drugs cluster together based on differentially spliced exons related to drug response. I also discover that drug modules form in multiple tissue types and detail genes implicated in mediating drug response through differentially spliced isoforms.

Chapter 7 summarizes the major findings from the dissertation and explains how the research done will have a substantial contribution to understanding response to concussion and differential splicing in cancer treatment. Potential improvements and future research directions are also discussed.

Chapter 2 Literature Review

2.1 Gene Expression and Splicing in Eukaryotes

The central principle of gene expression in biology dictates that DNA is transcribed to RNA and translated into protein [23]. In eukaryotes, DNA sequences of genes are comprised of three types of information: 1) untranslated sequence, the signaling and regulatory regions flanking the beginning and end of genes; 2) exons, the primary coding sequences that define the protein structure; and 3) introns, long stretches of non-coding sequence that are interspersed among exons and must be removed prior to protein translation [24]. During gene expression, DNA is first transcribed into pre-mRNA [24]. Then, introns are removed through a process known as splicing [24]. Finally, mature messenger RNA (mRNA) is relocated to the cytoplasm and translated by ribosomes to generate protein products.

The RNA splicing process is extremely important and determines the final sequence of the protein an RNA transcript produces [1]. Multiple protein sequences can be produced from a single pre-mRNA by varying the exons included in the final mRNA [1]. This is known historically as alternative splicing, although differential splicing is also used and may be more appropriate as “alternative” implies only two options. The use of “alternative” stems from many genes employing a single predominant isoform however advancements in transcriptome analysis have shown that genes routinely produce many isoforms. Large-scale studies report that up to 95% of human multi-exon genes may be differentially spliced [2]. Splicing is one mechanism that contributes to genome complexity and allows eukaryotes with limited genome sizes to produce much larger

numbers of gene products when compared to other species [1]. It has been proposed that splicing is a tightly regulated process and many alternative isoforms are nonviable or degraded before translation, however the function and ultimate fate of the vast majority of gene transcripts has not yet been determined and until we achieve a better understanding of differentially spliced gene products the potential impact of all isoforms should be considered [25].

2.1.1 Impact of Splicing on Biological Functions

Splicing is essential for control and regulation of many different cellular processes. Coordinated splicing networks essential for development have been identified in brain, heart and liver tissue [26]. For example, crosstalk between polypyrmidine tract-binding proteins (*PTBP1* and *PTBP2*) and serine/arginine repetitive matrix protein 4 (*SRRM4*) directs neuronal cell differentiation through modulating the inclusion of *PTBP2* exon 10 [26]. Initially, *PTBP1* works to suppress the inclusion of *PTBP2* exon 10 in neuronal progenitor cells but as cells differentiate *SRRM4* promotes *PTBP2* exon 10 inclusion and in turn neuronal development and tissue maintenance [26]. Likewise in heart, *CELF1* and *MBNL1* compete to regulate splicing essential for heart development [26]. The tissue-specific expression of many transcript isoforms and switch-like activity of some exons indicates a strong regulatory influence of splicing on human tissue complexity [27].

Mis-regulation of splicing is known to contribute to many different diseases. It has been estimated that 35% of single nucleotide variants (SNVs) that cause disease disrupt splicing [24,28,29]. Rare disease-causing variants altering splicing have been found

associated with spinal muscular atrophy (SMA), amyotrophic lateral sclerosis (ALS), Duchenne muscular dystrophy (DMD), Hutchinson-Gilford progeria syndrome (HGPS), Retinitis pigmentosa (RP) and cancer among others [30]. Splicing is also thought to contribute to milder differences in humans. On average, around 30,000 SNV differences are found between individuals [24]. Many of these 30,000 SNVs are synonymous, meaning they do not change the protein sequence, or are located in non-coding regions and contribute to phenotypic changes by altering splicing efficiency of individual exons [24]. Such variants are referred to as splicing quantitative trait loci (sQTLs) [24].

2.1.2 Evolutionary Origin of Splicing

The common accepted theory regarding the origin of eukaryotes is that there was a symbiotic endocytosis event between a large archaeal bacterial in the Asgard lineage and a smaller pre-mitochondrial bacterium in the alphaproteobacterial lineage [31,32]. During the evolutionary period following eukaryogenesis, a dramatic expansion in the diversity and complexity of the genome of the last eukaryotic common ancestor occurred [33]. It is likely during this time that the mobile genetic elements known as group II introns began inserting into existing genes, and later evolved into non-mobile introns processed by de-coupled specialized splicing machinery now known as the spliceosome [33]. Direct support for this theory has been observed in plants where a group II intron found in the mitochondria transferred to the nucleus and evolved into a spliceosomal intron [33]. Additionally, the structure and splicing mechanism of group II introns is highly similar to the modern spliceosome, and researchers believe group II introns are the predecessor to core spliceosome machinery in eukaryotes [34,35].

Yeast contains a small set of highly conserved splicing proteins that are likely the minimum core machinery required for splicing [36]. Eukaryotic organisms like drosophila and humans contain homologs for all yeast splicing proteins as well as additional proteins necessary for differential splicing [36]. Although other eukaryotes perform splicing, splicing in simpler organisms like yeast is mostly constitutive and the generation of multiple isoforms through alternative inclusion of exons is rare if present at all. The structure of the spliceosome in humans is more dynamic and unstable than yeast, making it harder to study but providing the flexibility required for interaction with a larger set of differential splicing-related proteins with selective and regulatory activity [37].

2.1.3 Splicing Components, Process and Outcomes

The collection of structural RNAs that form the scaffold and the many proteins that catalyze the necessary reactions for splicing are referred to as the spliceosome [36]. The spliceosome is comprised of five primary structural small nuclear RNAs (snRNA): *U1*, *U2*, *U4*, *U5* and *U6* [36]. Unlike some other snRNA complexes, the snRNAs of the spliceosome do not self-assemble into an active state on pre-mRNA introns [36]. Instead, a large number of additional proteins facilitate the interactions between snRNAs and proteins as well as the recognition of target pre-mRNA binding sites [36]. Spliceosomal proteins help correctly form the structure of the spliceosome, position the active sites of the spliceosome and ensure splicing is happening at relevant locations on the pre-mRNA [36].

The canonical splicing reaction occurs though eight conformational changes of the spliceosome, although the process can be broken down into 5 critical steps: 1) *U1* is recruited to a short conserved sequence at the 5' splice site of the intron and factor *U2AF* recognizes a short conserved sequence at the 3' splice site, recruiting *SF1* which is subsequently displaced by *U2* allowing it to bind on the branch point forming complex A; 2) the *U4*/*U6*.*U5* tri-small nuclear ribonucleoprotein (snRNP) complex is recruited forming pre-catalytic complex B; 3) *U1* and *U4* are destabilized and leave the complex, thereby producing the activated B complex; 4) a conformational change to B* is catalyzed by spliceosome member proteins and the first transesterification reaction occurs, where the 2' OH group on the branch point adenosine residue performs a nucleophilic attack on the 5' splice site; 5) the now available 3' OH of the 5' exon performs a nucleophilic attack on the 3' splice site, ligating two exons together, releasing the spliceosome and intron in the form of a lariat loop [36]. In canonical splicing, which accounts for the vast majority of *U2* dependent splicing, the 5' end of the intron contains the sequence "GT" and the 3' "AG" [38]. In non-canonical splicing the sequences of the 5' and 3' intron ends differ.

There are five primary types of differentially spliced events: 1) Alternative 5' splice site (A5SS) where a small piece at the start of an exon may be removed; 2) Alternative 3' splice site (A3SS) where a small piece at the end of an exon may be removed; 3) Skipped exon (SE) where an exon may be completely removed; 4) Mutually exclusive (ME) where two exons are alternatively included in a transcript and ever seen together; and 5) Retained intron (RI) where an intron that is normally spliced out is included in the final

transcript sequence [39]. Complex splicing of transcripts can also occur, such as multi-exon events or trans-splicing where two different transcripts are spliced together although complex events are thought to be rare and many can be described as some combination of the five primary event types. Lastly, inclusion of alternative transcript start, stop and poly adenylation sites are typically analyzed separately and not considered categories of differential splicing.

2.1.4 Splicing Regulation

Splicing is regulated through trans-acting RNA-binding proteins (RBPs), which can be expressed in a tissue- or condition-specific manor, that modulate the efficiency of splicing across groups of exons [40]. These key RBPs, called splicing factors, help the splicing machinery recognize exon-intron boundaries though binding on cis-regulatory sequences nearby [40]. The successful binding of splicing factors is dependent on the strength of the cis-acting motif. Splicing factors in particular exhibit transient activity, weak or short-term binding site association, and splicing efficiency is determined by the relative strength of motifs around an exon-intron boundary [41]. Neighboring splicing factors compete, encouraging or discouraging splicing at specific sites. The identity of splicing factors in the regulatory environment is dictated by other cellular processes like expression of certain transcription factors and tissue-specific genes.

There are two primary classes of splicing factors, Heterogeneous ribonucleoprotein particles (hnRNPs) and serine-arginine rich (SR) proteins [41]. Other types of RBPs have been implicated in splicing regulation as well, and growing focus on RBP research will

likely reveal many more splicing candidates in the near future. Splicing factors can act as "enhancers", increasing the chance exons are included in a transcript, or "silencers" depending on their physical properties and the location they bind to [41]. For example, SR proteins are generally thought to enhance splicing; however, SR protein *SRSF1* can act as an enhancer, when binding near the 5' splice site but a silencer when binding near the 3' splice site [42]. The activity of splicing factors may also change depending on their binding partners or other nearby splicing factors [42]. Taken together, what others have discovered indicates that regulation of splicing is very complex, that studying splicing regulation is essential to understanding splicing mechanisms in a disease context and that there is a lot to learn about how inclusion of specific exons in transcript isoforms is regulated.

2.1.5 Splicing in Cancer

Splicing has been shown to influence every major hallmark of cancer development and progression [4]. In some cases, like the splicing factor *SRSF1* discussed above, alterations in splicing factors and other spliceosome components can have substantial and far-reaching effects touching on many cancer-related processes [42]. Increased expression of the splicing factor *SRSF1* has been found in tumors, and studies have shown overexpressing *SRSF1* is sufficient to promote cancer transformation of mammary epithelial cells [42,43]. In other cases, like in epidermal growth factor (EGF) receptor *ErbB4*, the effects of splicing are more focused. The estrogen receptor gene *ErbB4* produces multiple isoforms characterized by two factors: whether they include a 16 amino acid cytoplasmic domain with binding sites for phosphoinositide 3-kinase (PI3K),

"CYT-1" for domain included and "CYT-2" for not included; and if they are susceptible to proteolysis at extracellular juxtamembrane (JM) domains, "JM-a" is susceptible and "JM-b" is not [44]. Depending on which domains are included, isoforms of *ErbB4* can have opposite functional effects. The *ErbB4* JM-b CYT-2 isoform stimulates apoptosis inhibiting cancer, whereas the JM-a CYT-2 isoform promotes cellular proliferation, survival and anchorage-independent growth encouraging cancer [44]. Both isoforms are functionally activated by several EGF-like growth factors and inhibited by ErbB kinase inhibitors, indicating that isoform-specific treatment strategies for cancer may be extremely important [44].

Splicing influences cancer cell metastasis through epithelial-to-mesenchymal transition (EMT); the transformation of cells from an immobile state, having cohesive cell-cell junctions, to a motile and invasive state where cells dissociate from neighboring cells and become elongated allowing them to travel through tissues [45]. Splicing has a large influence on EMT and is responsible for orchestrating a series of necessary events leading to EMT [46,47]. The key steps in EMT are tight junction disassembly, loss of membrane-bound E-cadherin, loss of apical-basal polarity, re-localization of cytoskeletal components controlling front-rear polarity and expression of matrix metalloproteinases [46]. Splicing has been found to modulate the key steps in EMT through exons targeted by *ESRP1* & 2 splicing factors; affected genes include *FGFR2*, *p120*, *NUMB*, *SCRIB* and *ENAH* [46,47]. Other splicing factors that modulate EMT are *MNBL1*, *RBFOX2*, *SRSF1* and *SRSF3* [46]. Lastly, splicing could facilitate chemoresistance through activation of

EMT and either the EMT-associated cellular processes or the acquired stem-like features [46,48,49].

Splicing influences cancer cell survival through processes related to control of cell cycle. One strategy healthy cells use to prevent formation of cancer is locking abnormal cells in senescence, ending the cell cycle and stopping further replication [50]. Global changes in spliceosomal gene expression have been observed during senescence, knockdown of individual spliceosomal genes was shown to induce senescence, and splicing factors such as *SRSF1*, *SRSF3*, *HNRNPA1*, *HNRNPA2* and *SF2* are known to be pivotal in senescent transformation [51]. Splicing can prevent senescence through control of EMT by inducing EMT-associated transcription factors *ZEB1/2* which repress cyclin kinase inhibitors and prevent *EGFR*-dependent senescence in esophageal squamous cell carcinoma [46]. Senescence can also be beneficial to long-term survival of cancer cells, preventing rapidly dividing cancer cells from succumbing to chemotherapeutics and allowing cancer cells to survive cancer treatment [50]. Cancer cells can later exit senescence and resume uncontrolled proliferation [50]. Surviving cancer cells can also acquire new resistance conferring mutations, making subsequent treatment less effective [50]. More research is needed to understand the role of splicing in senescence and whether splicing-based therapeutics could be used to prevent cancer cell growth or senescence-based survival strategies.

Splicing also influences cancer cell survival through apoptosis, and immune system evasion. Apoptosis, another process healthy cells use to stop cancer propagation, is where

damaged or diseased cells receive signals that induce programmed cell death [52]. Splicing controls apoptosis through multiple pathways and proteins, such as the example of *Erbb4* above [44]. Another well-known example is alternative splicing of Bcl-x_L and Bcl-x_S isoforms [52,53]. When Bcl-x_L, the anti-apoptotic isoform, is targeted with an antisense oligonucleotide cancer cells produce more Bcl-x_S, the pro-apoptotic isoform, inhibiting colony formation and eliciting apoptosis [53]. Additionally EMT, strongly regulated through splicing, confers resistance to apoptosis through the mitogen-activated protein kinase/extracellular signal-regulated kinase (MEK/ERK) and phosphoinositide 3-kinase/protein kinase B (PI3K/AKT) pathways [54]. Furthermore, the splicing factor *SRSF1* promotes splicing of *BIM* and *BIN1* isoforms without pro-apoptotic functions [43]. Lastly, splicing helps cancer cells evade the immune system through EMT, but also by defining isoforms of HLA-G cell surface markers that inhibit immunocompetent cells and through addition of exon 7 in MHC-I which reduces the potential for cytotoxic T lymphocyte stimulation [4,55].

2.2 Cancer and Treatment

Cancer is a disease of the genome, where somatic mutations acquired during the course of an individual's life alter the regulation or function of essential genes and lead to uncontrolled cellular growth and proliferation [56]. Cancer and cancer treatment strategies are complex because just as no two individuals are the same, no two cancers are the same. However, there are similarities between cancers stemming from the same tissue and cancers originating from the same type of environmental exposure. Similarities between cancers can be exploited during cancer treatment, as can the general properties

of cancer cells such as high rate of proliferation, by matching cancers to treatment strategies that have proven to be effective in cancers with similar genetic profiles.

Cancers typically originate from the abnormal regulation or function of two major classes of genes, oncogenes and tumor suppressor genes [57]. Oncogenes drive the growth and proliferation of cells [57]. Tumor suppressor genes act to suppress the growth and proliferation of cells by reducing or silencing expression of oncogenes [57]. Mutations in both oncogenes and tumor suppressor genes are required for cancer transformation, and in many cases numerous causative "driver" and innocuous "passenger" mutations are observed in active cancer [56]. Differentiating between driver and passenger mutations is difficult and is one aspect of cancer that complicates treatment.

Cancer treatment strategies can differ based on many factors such as tissue of origin, severity or stage, if a cancer is newly diagnosed or a patient is in relapse, and the types of cancer treatments used to previously treat the patient [22]. The goals of cancer treatment can also vary and range from merely alleviating symptoms to attempting to cure an individual. Numerous challenges around cancer treatment exist as well. Because each individual and cancer are unique, selecting the best treatment for an individual is difficult. Patients respond differently to the same course of treatment, and it is important to apply a treatment that is effective at killing cancer cells while maintaining a level of toxicity that is tolerable for the patient. Once a cancer has been exposed to anti-cancer therapeutics the genetic profile of the tumors can change, conferring resistance to the tumors and altering the efficacy of future cancer treatment. Ongoing research into cancer

genetics with high-throughput genomics has improved understanding of cancer and helped to develop new targeted treatments that are highly effective and low in toxicity [58]. Cancer genomics has also opened the door to personalized medicine, and it is now common to consider a patients' own genetic profile when recommending the best course of treatment [59,60]. Cancer treatment regimens continue to improve, and as we learn more about our own genetics our abilities to treat genomic diseases improve [58]. Recent breakthroughs in immunotherapy have dramatically improved cancer treatment outcomes [58]. However, cancer is genetically complex and even though it is possible to cure some cancers it is likely that even with the most advance scientific tools select cancers will remain chronic managed diseases rather than curable illnesses.

2.2.1 Anticancer Therapeutics

Anticancer therapeutics play a key role when treating the majority of cancers [22]. Anticancer drug molecules are classified by the type of action they take against cellular biology, and there are many classes of cancer drugs. The main drug classes of non-targeted chemotherapeutics are alkylating agents, intercalating agents, radiomimetic agents, topoisomerase inhibitors, antimetabolites, antimitotic drugs, polyamine inhibitors and iron-mediating drugs [58]. Non-targeted chemotherapeutics have an indiscriminate cytotoxic effect on cells but are useful for treating cancer because they disproportionately impact highly-replicative cells or induce further genetic damage to already mutated cancer cells leading to apoptosis. Unfortunately chemotherapeutics are also toxic to normal healthy cells and special attention to dosing and spacing of drug administration must be paid.

The use, type and importance of specific cancer therapeutics depends primarily on the cancer and prognosis a patient receives. Chemotherapeutics can be applied as the sole course of treatment or in combination with other forms of cancer intervention. Referred to as neoadjuvant therapy, chemotherapeutic treatment before surgery is typically aimed at reducing the size of a tumor to increase the likelihood of success following surgery [61]. Adjuvant chemotherapy is the use of chemotherapeutics after surgery and is designed to eliminate any remaining hidden cancer cells [61]. As our understanding of cancer has improved, so too have the number and variety of chemotherapeutics as well as their structured treatment regimens. Nanoparticles and extracellular vehicles have improved the efficacy of chemotherapeutics while decreasing the toxicity [62]. Nanoparticle encapsulation of drug molecules can increase drug solubility, allow for environmental triggering of drug release or even provide for direct targeting of tumors through conjugation of antibodies that bind to specific cell surface proteins like HER2 [62].

2.2.2 Drug Treatment Combinations

First generation chemotherapeutics, such as the topoisomerase inhibitor doxorubicin, are now typically used in combination with other chemotherapeutic drugs such as the alkylating agent cisplatin or the mitotic inhibitor paclitaxel [63]. This is because the subsequent generations of chemotherapeutics have been shown to have reduced toxicity and combination treatment strategies have been proven to be more effective at eliminating tumors with heterogeneous cell types common in cancer [63]. Chemotherapeutics are also increasingly being used in combination with targeted

therapeutics. In such cases tumor genetic profiling has identified a mutation, such as that resulting in growth factor dependence, which can be exploited using a highly-specific small molecule inhibitor [61]. Chemotherapy will be applied in conjunction with the small molecule inhibitor to destroy tumor cells that are resistant to the targeted therapeutic and ensure greater success. However, medical oncologists must be careful to use drugs with complementary effects. Two complementary cancer therapeutics can have powerful additive effects in cancer treatment, but adverse combinations can have neutral or severe negative effects like enhanced toxicity.

Identifying beneficial drug combinations typically begins with computational bioinformatic analysis of known molecules having associated pharmacological and genomic data. First, gene targets of drugs are identified either through annotated mechanisms of action, through merging drug-response and genomic data to separate genomic features relevant to drug activity, or through computational modeling of drug-protein interactions. Genomic features identified from analyzing drug-response and genomic data can be used to narrow the number of drug-protein interactions explored during computational modeling. Two strategies are commonly used to determine possible drug targets with computational modeling, molecular docking and ligand shape matching. In molecular docking, the chemical structure of the small drug molecule is inserted into hydrophobic pockets of computerized protein structures [64]. The affinity of the pocket for the molecule is calculated based on free energy and an affinity scoring function [64]. In ligand shape matching a database of known ligand structures is searched using the structure of the small molecule and probable matches are identified [65]. In addition to

chemical structure, molecular docking and ligand shape matching algorithms also consider key chemical properties like charge, electrostatic and van der Walls energies [64,65].

Next, disease-specific gene targets are identified, again through annotation of known disease mechanisms or through the combination of disease severity and genomic data. One example of an analysis method that can be used to identify disease target gene relationships is genome-wide association studies (GWAS). Annotation of disease-related proteins is an ongoing process and frequently knowledge from new studies or databases can change the known disease targets.

Finally, drug gene targets and disease gene targets are overlapped, and the relationships are characterized. There are six relationship possibilities: overlapping exposure, where both drugs have targets in common as well as targets within the disease target list; complementary exposure, where both drugs have targets within the disease target list and no targets in common; indirect exposure, where two drugs share common targets but only one drug has targets within the disease target list; single exposure, where two drugs do not share any common targets and one drug has targets within the disease target list; non-exposure, where two drugs share targets and neither has targets within the disease target list; and independent action, where all drug and disease target lists are unique and no overlapping targets are found [66]. Only drug pairs that have a complementary exposure relationship with the disease have been shown to potentially improve treatment outcome [66]. After potential complementary drug pairs have been identified a long process of

testing and validation must follow to ensure the new treatment strategy is beneficial, safe and effective.

2.2.3 Development of Cancer Treatment Regimens and New Therapeutics

Development of new cancer treatment regimens, and especially new cancer therapeutics, is an extremely expensive and time-consuming process. The cost to develop a new cancer therapeutic is estimated to be between $1.3 and $2.7 billion [67]. Usually a minimum of 12 years is required to develop a drug but often development runs longer, even 18 years or more [68]. Basing new therapeutics on previously approved structurally-similar molecules saves on development time and costs, and allows a drug company to bring a similar product to market to compete with rivals. Because of this many cancer drugs molecules are closely related to each other, both in structure and activity, and drug design innovation stagnates. However, some molecules with similar structures can have slightly different activities and side-effects. Similar molecules can therefore be potentially suited to treat a different type of disease, used in patients that do not respond well to another drug or be used in a new combination with another drug. Another way to save time and cost in developing new treatment strategies is drug repositioning, where a drug that has been previously approved to treat one disease is applied to another [69]. Drug repositioning, along with modifying the specifics of cancer treatment regimens like order and duration of treatments, is common in development of cancer treatment strategies. Even so, newly positioned drugs and treatment regimens must be supported by experimental evidence in animal models and go through phase I, II and III trials that can take 7 years or more to complete [63,68].

2.3 High-throughput Genomics, Gene Expression and Splicing Quantification

Before high-throughput genomics, studying the transcriptome was a tedious and time-consuming process. A randomly primed complementary DNA (cDNA) library, prepared from mRNA using purified reverse transcriptase, was created for each sample and individual cDNA fragments would be cloned into vectors to facilitate sequencing [70]. From cDNA clones, expressed sequence tags (ESTs) which are sequences complementary to the original mRNA would be generated [70]. ESTs are therefore pieces of expressed genes, generally between 200-600bp in length, and could be mapped back to chromosomal locations with hybridization-based labeling technologies [70]. Because ESTs are incomplete sequences they can only be used to infer which genes are expressed and not the high-quality mRNA sequence, exon composition or mRNA quantity.

Advancements in molecular biology and instrument technology now allow for sequencing the entire transcriptome of even single cells. However, due to cost, it is common to sequence the transcriptome of bulk samples with mixed cell types or to sequence a small fraction of the transcriptome of many single cells. The information gained from transcriptome experiments is highly dependent on the experimental design. For example, if the goal of the experiment is to quantify micro-ribonucleic acid (miRNA) then a library preparation procedure with a size selection step is required [71]. If the experimental goal is to study long non-coding RNA then a total RNA library preparation procedure is required as poly-a selection from hybridization-based preparation procedure would fail to capture many non-coding RNAs [72]. The downstream analysis strategy is also highly dependent on the experimental design.

2.3.1 High-throughput Transcriptome Sequencing Technologies

Illumina sequencing technology is currently the market leader in high-throughput sequencing, primarily due to the massive output capabilities and low cost [73]. Illumina machines generate more sequence information in a single run (up to 6Tb on the NovaSeq 6000®) than any other existing technology, support a wide range of experiment types, yield very high-quality data with error rates below 1% [73]. Illumina also provides more competitive pricing than smaller companies [73]. However, Illumina technology only produces reads up to 250bp long, and library preparation requires amplification of nucleic acids which can introduce errors and create sequence duplicates. Another limitation of short-read sequencing is that reads are much smaller than the mRNA molecules they are generated from, meaning that the complete mRNA sequence and therefore isoform structure remain unknown.

Illumina library preparation for mRNA sequencing begins with clean, extracted mRNA. The template mRNA is usually either captured via polyA probe hybridization or enriched for by degrading ribosomal RNA [74]. Processed mRNA is then fragmented into smaller pieces, between 500-800bp long [74]. Fragmented mRNA is then randomly primed and cDNA is produced with reverse-transcriptase [74]. Second strand synthesis then replaces the original mRNA template creating double stranded cDNA [74]. Next, sample index sequences used to identify samples during multiplexing and Illumina adapter sequences are attached [74]. Prepared cDNA fragments are then hybridized to Illumina flowcells, the delivery mechanisms for sequencing reagents [74]. Clonal amplification of the bound fragments is then performed to boost the sequencing signal [73].

Sequencing is accomplished through sequencing by synthesis. Based on Sanger sequencing, fluorescently labeled terminator nucleotides are added to the flowcell and incorporated into the newly synthesized strand by DNA polymerase [73]. Each of the four DNA bases (A,C,G,T) is labeled with a unique fluorescent wavelength. A camera takes an image recording the color at each clonal DNA cluster [73]. The terminating 3' blocker and dye are then chemically removed and another cycle begins [73]. In the case of paired-end sequencing, the free ends of the DNA fragments are hybridized again to the flowcell and the previously attached ends are released, flipping the fragments and allowing sequencing from the other side. Base calling is done by decoding the colors and quality measures are dependent on the clarity and intensity of spectra collected during the run.

Another popular instrument is the ThermoFisher Ion Torrent® sequencing system [73]. This technology is similar to Illumina's in that it uses sequencing by synthesis detect the sequencing signal [73]. However, sequencing with the ThermoFisher Ion Torrent® system requires fragmented templates be attached to beads rather than a flowcell [73]. The size of sequencing reads is similar to Illumina and is between 200-400bp. A template is clonally amplified around the bead in a microscopic oil droplet and all beads are then deposited into microwells where sequencing takes place [73]. The sequencing signal is detected by a semiconductor which records the change in pH that occurs when a hydrogen atom is released following nucleotide incorporation [73]. Due to the chemistry, the platform is susceptible to errors when sequencing homopolymer stretches greater than

6 bases as the extended signal from multiple identical bases can be difficult to classify [73].

Both the Illumina and Thermofisher Ion Torrent® sequencing technologies are referred to as second generation (or next generation) sequencing technologies. Third generation sequencing technologies, which can produce a signal without amplification, are also currently used by laboratories on a limited basis while the sequencing chemistry is being perfected. Single-molecule real-time (SMRT®) sequencing provided by Pacific Biosciences (PacBio) is a popular third generation sequencing technology and is widely used primarily because of its read length [73]. The biggest advantages of PacBio's SMRT® technology are long average read lengths around 14kb (but can be as long as 60kb), real-time data collection and absence of clonal amplification [73]. Disadvantages are high cost, lower throughput and higher error rates (up to 11%) [73]. However, cDNA can be modified to allow sequencing the same fragment many times compensating for the high error rate and bringing the final accuracy above 99%. The extended read size allows for whole isoform sequencing, and complete mRNA isoform identification and quantification.

The SMRT® sequencing protocol for RNA, known as Iso-Seq®, uses polyA selection to capture mRNA. cDNA is synthesized with the Moloney murine leukemia virus reverse transcriptase that adds multiple cytosine residues to the end of the elongated strand [75]. A primer is added which hybridizes to the polyC residues providing a template to incorporate an amplification primer. Second strand synthesis then replaces the original

mRNA template. Barcodes for multiplexing can be added to cDNA using sample-specific primers. Double stranded cDNA is then ligated to hairpin primer (SMRT-bell®) loops that create one circular DNA product [73]. Circularizing the DNA product is what allows the same fragment to be sequenced many times over, thereby increasing the accuracy to over 99% [73]. Transcripts can also be selected by size before sequencing using a follow-up purification method. Sequencing is accomplished by hybridizing a sequencing primer and DNA polymerase to the DNA fragment, subsequently binding the DNA polymerase to the bottom of a well in a SMRT® cell and introducing phospholinked fluorescent nucleotides [73]. Rather than repeated cycling and chemical washing as in the Illumina process, SMRT® sequencing runs in continuous real-time [73]. A fluorescent signal specific to one of the four DNA bases is emitted as each nucleotide is incorporated. Instead of cycle-specific images as in the Illumina process, a video of the SMRT® cell is recorded and used for basecalling [73].

Another example of a third-generation technology is the Oxford Nanopore® technology [73]. The major advantages of the Oxford Nanopore® technology are long read lengths between 6kb and >60kb, single molecule sequencing with no amplification, minimal library preparation, fast sequence acquisition and low cost [73]. Some third generation machines, such as the MinION®, are highly portable and merely require a laptop and basic reagents to operate [73]. Additionally, nucleic acid modifications such as methylation can be detected during sequencing without any changes in workflow. Also, both DNA and RNA from a sample can be sequenced together in the same run. The sequencing trace from the run can differentiate between DNA and RNA so post-

translational RNA modifications can be studied as well [76]. The major disadvantage of Oxford Nanopore® is high error rates with accuracy ranging from 83% to 95% [73,76].

Library preparation for Oxford Nanopore® is very fast and simple compared to second generation technologies. Nucleic acids, either DNA or RNA, only require fragmentation and adapter ligation before sequencing. cDNA can also be sequenced if desired [77]. For direct RNA sequencing, the 5' methyl cap of the mRNA is removed and the sequencing adapter with bound motor protein is ligated to the 5' end. PolyA selection based enrichment of RNA can also be done, and in that case a PolyT primer with motor protein is hybridized to the 3' end of mRNA [77]. If greater accuracy is desired a 2D sequencing library can be prepared by ligating a hairpin primer to the end of the fragment and extending the complementary strand [73,78]. The hairpin primer connects the complementary strand to the template strand and allows the two to be sequenced together in one pass [73,78].

Nanopore® sequencing is accomplished by passing the prepared template through a voltage-sensitive synthetic membrane [73]. First, the primer-attached motor protein binds to a membrane-attached pore protein [73]. A current is applied to the membrane and the ionic differential pulls the template strand through. The pore registers changes in ionic current as the strand passes [73]. Disruptions in the current, such as those from various nucleotides and modifications, are measured and decoded in real-time [73]. The attached motor protein helps to reduce the speed of the template, thereby increasing the accuracy of base calling after the run [73].

2.3.2 Mapping of High-throughput Sequencing Data

After sequencing of samples is complete, sequencing data must be quality assessed and reformatted in a way that allows researchers to understand and manipulate the data. Quality of the individual base calls is produced by the base calling algorithm, the mathematical model that converts raw sequence output such as image files to sequence information, for every sequencing method discussed above. Although the specifics of the base quality score calculation can vary, all methods of calculation are aimed at estimating how likely it is that a designated base was incorrectly assigned and all platforms produce what is called a PHRED score [79]. The equation for PHRED score is:

$$Q_{PHRED} = -10 \times log_{10}(P_e)$$

The PHRED score translates to a number between 0 and 93 where 0 is an incorrect base call and 93 is an extremely accurate base call [79]. Experiments that do not meet a minimum average PHRED quality level are considered failed runs and need to be troubleshooted. Raw sequences are stored in standardized plain-test sequence files known as FASTQ files that include the PHRED score for each base, unique read ID, and read sequence [79]. These files are usually compressed to save storage space.

The majority of sequencing experiments for the purpose of gene expression and splicing quantification are done using tissue or samples from model organisms or humans (*Homo sapiens*). Genomes for *Homo sapiens* and model organisms such as mouse (*Mus musculus*) or rat (*Rattus Norvegicus*) have been well characterized though decades of

research, and standardized genome sequence references for these species are publicly available on the internet through reference sequence consortia such as the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute's Ensemble and the University of California Santa Cruise (UCSC) Genomics Institute [80–82]. Using a reference genome allows researchers to map the location of sequencing reads back to chromosomal positions, refer to an accurate annotation of genetic features so genetic elements such as genes can be quantified and produce comparable results across sequencing experiments.

Mapping of sequenced reads to reference genomes is done using a wide array of algorithms that fall into two categories; ungapped aligners and splice-aware aligners. Ungapped aligners, such as Burrows-Wheeler Aligner Measure Memory (*BWA-MEM*), were designed to align reads to whole-genome DNA sequences and are useful for DNA-based experiments or estimation of alignment rates to continuous known sequences [83]. Ungapped aligners are not suited for RNA sequencing analysis. Splice-aware aligners like *Bowtie2*, *STAR* and *HISAT2* allow reads to be split during alignment [84–86]. Splitting reads during alignment is important because, as discussed above, during mRNA production transcripts are spliced to remove intronic sequence stretches that divide exons. Reads from RNAseq data frequently cross exon boundaries and a portion of the read will include sequence from another exon thousands of bases up- or down-stream. Unless a splice-aware aligner is used many junction reads will either fail to align or be discarded for poor alignment quality.

During alignment, minimum quality parameters are applied to ensure accurate mapping of sequencing reads to genomic coordinates. Quality parameters are usually influenced by the type of experiment a researcher is analyzing, the overall quality of the data being analyzed and the algorithm being used to map the reads. Limits can be set on parameters like the minimum number of matched bases, the maximum number of mismatches, the minimum number of bases mapped to another exon in a split read and the maximum genomic distance between paired reads originating from the same fragment. Aligned reads are usually filtered based on their mapping quality (MAPQ) score [87]. The MAPQ score is more complicated than the PHRED score but the premise is the same, to calculate the probability a read is aligned to an incorrect position. The score spans from 0 to 255 where 0 is incorrectly mapped, increasing scores designating better mapping positions and with some tools (as is the case with *STAR*) 255 reserved to flag reads with unique positions in the genome. The basic form of the MAPQ score for an individual read is:

$$MAPQ = -log_{10}\left(1 - \max\left(\frac{10^{-\mathrm{MMS}_i}}{\sum_{j-\{i\}} 10^{-MMS_j}}\right)\right)$$

In MAPQ, *MMS* is the sum of the PHRED base scores for mismatched bases in an alignment, MMS_i denotes the current alignment and the sum of the mismatch scores for alignments of the same read excluding *i* are in the denominator [87].

Alignment results are stored in a sequence alignment map (SAM) format file [88]. SAM files include the unique read id, raw sequence, CIGAR string detailing the alignment compared to the reference, MAPQ score, PHRED scores and other flags to designate

known read information [88]. SAM files are typically converted to binary format (BAM) to save on storage space.

2.3.3 Quantification of Gene Expression

Following read alignment, gene expression is quantified by counting the number of reads that align to each gene. Genomic positions for genes are obtained from annotation files matched to the genome reference sequence version used during alignment. These files are maintained by the various genome reference consortia and detail many different types of genomic features that researchers may be interested in. In some cases, features that are still under investigation will be present. To avoid later inflation of hypothesis testing during statistical analysis it is important to select an annotation that includes only well-characterized reference genes when counting reads for gene expression purposes.

There are two basic strategies for counting reads assigned to genes; exon union and exon intersection [89]. The exon union method counts reads mapped to all annotated exons belonging to a gene [89]. The exon intersection method counts only reads mapped to exons that are present in all isoforms (constitutive exons) [89]. Each strategy has an inherent weakness; the exon union method tends to underestimate gene expression and the exon intersection method can cause reduced power during differential expression analysis [89]. More advanced isoform-based methods, that rely on statistical inference, exist as well [89]. However, isoform-based methods assume the identities of isoforms based on annotation and observed reads, and some of the isoforms selected for quantification my not actually be in the sample. Additionally, the exon union method has

been previously shown to perform similarly to more advanced isoform-based methods and better than the exon intersection method, so the exon union method is more commonly used than others [89].

If the ultimate goal of an experiment to compare the expression of a gene (or genes) across multiple samples, then as long as the same technique is used for all samples the specific quantification technique used is less important. However, if the goal is to compare expression of genes within the same sample, more attention should be paid toward selecting the best counting method. Lastly, it is important to note that raw read counts cannot be used directly for analysis or visualization of gene expression without normalization [89].

2.3.4 Quantification of Splicing

Quantification of splicing in high-throughput data falls into two categories; isoform-centric and exon centric quantification. Isoform-centric quantification techniques are equivalent to the isoform-based methods of gene expression quantification. Exon-centric quantification techniques on the other hand aim to calculate the proportion of all transcripts that contain an individual exon.

Isoform-centric techniques employ three basic steps: First, the identity of potential isoforms in the sample is inferred from a combination of the genome reference annotation and the alignment of observed reads in the dataset; Second, reads that can only originate from a single isoform and reads shared amongst fewer isoforms than the total are used to

compute a prior estimate of relative expression; Third, constitutive reads are assigned to isoforms using a statistical approach that maximizes the likelihood estimation and pseudo normalized expression values in the form of reads or fragments per kilobase per million (RPKM or FPKM) for each isoform are produced [90]. Depending on the tool, an isoform-centric algorithm may proceed ineratively until an approximate best possible solution is found and the algorithm converges.

Exon-centric approaches are simpler than isoform-centric approaches and do not attempt to infer the identities of isoforms in a sample. Instead, the proportion of transcripts containing an individual exon is calculated, primarily based on junction read counts [91]. A percent spliced in (PSI) value (also known as inclusion level) for each combination of upstream, cassette and downstream exons is produced [91]. A single combination of upstream, cassette and downstream exons is frequently referred to as a splicing event. Inclusion junction reads, reads mapped to a flanking exon and a target exon that support the presence of the target exon in question, may be twice as abundant as exclusion junction reads since two junctions are present when an exon is included (the upstream-target junction and target-downstream junction) but only one junction is present when an exon is skipped (the upstream-downstream junction) [91]. Therefore, the PSI value is approximately the ratio between the number of inclusion junction reads and the sum of the number of inclusion junction reads added to twice the number of skipping junction reads [91]. The effective length of the event (the total number of bases considering the upstream, target and downstream exons limited by read length) can be used to adjust the PSI value to increase the precision [92].

Non-junction reads from the upstream, target and downstream exons may also be used with exon-centric quantification to estimate a confidence interval [91]. When confidence interval estimation is done, reads are randomly sampled to build a distribution of PSI values and the final mean PSI estimate is reported with the PSI distribution range [91]. Narrow confidence intervals can indicate strong confidence in the PSI estimate, but also might be from a lack of reads during sampling. To avoid false-positives and ensure events are real, splicing events should always be filtered by the total number of reads supporting the event.

Isoform- and exon-centric approaches each have their own respective benefits and drawbacks. Isoform-centric approaches are good for accurately summarizing the expression of genes across multiple isoforms when fewer, well annotated isoforms in a gene exist. Isoform-centric approaches can also allow researchers to identify an imbalance in gene expression that can be missed through the common exon union quantification technique [89]. However, as mentioned in the gene expression quantification section, Isoform-centric techniques may infer the identity of isoforms that are not actually in the sample; this is especially true if the gene structure is complex. Exon-centric approaches are better when gene structure is complex, or when novel (unannotated) splice sites can be present. Annotation of events in exon-centric approaches relies on observed reads and gene annotation so no inference is required. Because exon-centric techniques target a smaller window of a gene they are well suited for downstream analysis, such as RNA-binding motif enrichment, around exons that might explain the regulation of the spliced event. Both isoform- and exon-centric

approaches can identify novel splice junctions and therefore never-before seen isoforms. However, in isoform-centric approaches a false-positive junction can propagate error by influencing the expression of other isoforms. Regardless of the method, care must be taken to ensure the novel splice junctions are reliable.

2.4 Differential Gene Expression Analysis Methods

Differential gene expression (DGE) analysis is the process of comparing gene expression between two groups of samples to identify alterations in gene expression due to a treatment or condition. DGE analysis is a long-researched field so the properties of the data have been well characterized and many analysis approaches have been developed. The process used for computational analysis of any data begins with characterizing the data and understanding its properties. The properties of the data then dictate the type of normalization required, the best approaches to processing the data and what the expected outcome may be.

2.4.1 Normalization of Gene Expression Data for Visualization

The purpose of data normalization is to remove the influence of technical effects that can systematically bias downstream analysis [93]. Several factors can bias RNAseq expression data such as the total number of reads sequenced for individual samples, the differing lengths of genes, mRNA expression dynamics, and changes in experimental factors or machine performance across multiple runs [93]. It is necessary to normalize RNAseq expression data so that values are comparable within and between samples. The

type of normalization used to correct for technical effects depends on how the data is being used.

Normalization for the purposes of visualization is usually simpler than the process required for DGE analysis; this is because the goal during visualization is to create an obvious scale or separation between data to illustrate change, and the outcome is more qualitative. The primary technical concerns when visualizing data are total number of reads sequenced in each sample and the length of the genes. Normalizing for the total number of reads accounts for differences in the amount of library for each sample loaded during sequencing. Normalizing for gene length allows for the expression of longer genes to be compared to shorter genes. Commonly used data transformations include counts per million (CPM), reads or fragments per million (RPKM or FPKM) and transcripts per million (TPM) [94]. In the following equations, *n* is the number of counts for a gene, *N* is the total counts for a sample, *l* is the length of a gene in bases, and both *i* and *j* are genes [94]. The equation for CPM is:

$$CPM_i = \frac{n_i}{N} \times 10^6$$

Notice that CPM does not account for gene length. The equation for RPKM and FPKM does account for gene length and is:

$$(R/F)PKM_i = \frac{n_i}{l_i N} \times 10^9$$

The difference between RPKM and FPKM is related to the type of RNAseq data used to produce the counts. When single-end RNAseq data is processed individual read counts are used and the measure is RPKM [94]. When paired-end data is processed fragments are usually counted instead to produce FPKM [94]. However, FPKM cannot be calculated from RPKM in a paired-end experiment by multiplying FPKM by two because counting fragments requires both reads from the same fragment to map during alignment [94]. The equation for TPM is:

$$TPM_i = \frac{n_i}{l_i} \times \left(\frac{1}{\sum_j \frac{n_j}{l_j}} \right) \times 10^6$$

Also, notice that in TPM, the per-base counts for a single gene are divided by per-base counts for all genes effectively making it a fraction of one million reads. Therefore, TPM is particularly useful for comparing genes within the same sample [94]. CPM, RPKM and FPKM are all measures commonly used for plots such as smear plots and heatmaps.

2.4.2 Normalization for Differential Gene Expression Analysis

Normalization for DGE analysis is complicated and can vary based on the analysis approach being used. The goal of DGE analysis is quantitative and therefore more advanced technical concerns need to be addressed to ensure accuracy and reproducibility. To properly normalize the data for DGE analysis, both sample- and gene-specific effects must be controlled. Rather than transforming the data values, as is the case with normalization for visualization approaches, normalization and scaling factors are applied

during generalized linear modeling (GLM) through parameters. Three corrections must be applied when performing DGE analysis: scaling the library size to account for differences in total reads, correcting for the variation introduced from sampling a pool of mRNA from the total RNA in each sample, and accounting for the inherent biological variation between replicates [93]. One example of a popular method that simultaneously scales the sample library sizes and accounts for mRNA sampling is trimmed mean of M-values (TMM) [93].

The TMM method assumes the majority of genes are not differentially expressed and attempts to estimate the ratio of total RNA production from the observed count data. The method calculates a scaling factor that is then used to adjust the library size for each sample, thereby shifting the model mean [93]. To approximate the relative difference in total RNA production using the TMM method, $\log_2$ fold changes for each gene with respect to the total counts are calculated between samples [93]. The method is 'trimmed' because a percentage of top and bottom genes, usually 30%, are ignored to avoid bias from highly and lowly expressed genes [93]. Then, the inverse of the asymptotic variance is used to weight the $\log_2$ fold changes, increasing the influence of genes with high read counts, and the mean (TMM) is taken [93]. The square root of the TMM value is multiplied against the non-reference sample library size, which produces a new 'effective' library size that is passed to a DGE analysis model as an offset [93].

Inherent biological variability is a major source of gene expression variation and complicates DGE analysis [95]. It is important to use replicates within groups in a

RNAseq experiment so that biological variability can be corrected for. While differences in library size and composition are normalized outside DGE modeling, normalization for biological variability is done my tuning DGE model parameters. Typically, the negative binomial (NB) distribution is used to model gene count data in RNAseq experiments and differentially expressed genes are identified by using a GLM to quantify the difference in NB distribution parameters between groups of samples [96]. The NB distribution is similar to the Poisson distribution, which assumes the majority of gene counts are collected around a mean and some counts will be abnormally high. However, the NB distribution includes an additional dispersion parameter that can help account for common and gene-specific deviations that the Poisson model cannot [96,97]. Therefore, due to biological and technical variance, the NB distribution is better suited to RNAseq analysis as it can account for deviations from the Poisson distribution [97].

Biological variability is corrected for by passing a dispersion parameter to the NB GLM model for each gene being analyzed [98]. Three types of increasingly powerful dispersion parameters can be calculated; common, trended and tag-wise (or gene-wise) dispersion. The final dispersion method used in a model depends on the number of genes and the number of samples in a DGE experiment, which dictate the power of the dispersion calculations. Given the number of samples in modern high-throughput RNAseq experiments, power is less of an issue and tag-wise dispersion is most frequently used. However, Tag-wise dispersion calculations depend on common and trended dispersion values [99,100].

Common dispersion is a single value representing the biological variability across all genes and is calculated using the effective library sizes, gene counts and maximum likelihood estimation [99]. The same common dispersion value is used for all genes during NB GLM modeling and is useful for situations where there are very few (or no) sample replicates [99]. However, common dispersion is an oversimplified interpretation of biological variability as it is well documented that genes with lower expression have higher dispersion and higher-expressed genes have lower dispersion [100].

Trended dispersion is where the common dispersion is fit to gene 'neighborhoods', gene groups comprising at least 25% of total genes with similar expression, based on the mean-dispersion trend of the observed data and a loess dispersion curve of degree 0 summarizing the local dispersion likelihood [100]. If trended dispersion is used groups of genes will share the same dispersion value [100]. Trended dispersion is also oversimplified because each gene is expected to have it's own dispersion, as each gene is subject to a unique genetic regulatory environment, but more reasonable than common dispersion [100].

Lastly, tag-wise dispersion is a method that calculates a unique dispersion estimate for each gene, while balancing the dispersion information gained from other genes [100]. Tag-wise dispersion calculation uses an empirical Bayes framework to find the weighted average between the dispersion calculated from individual gene information only and the trended dispersion [100,101]. Using tag-wise dispersion shrinks the gene-specific

dispersion towards the trended dispersion allowing for more precise dispersion estimates to be used while mitigating overestimation from low sample observations [100].

2.4.3 Analysis Tools and Presenting Results

As mentioned above, DGE analysis is primarily done using the negative binomial GLM framework. The vast majority of DGE analysis is done with one of two software packages; *edgeR* or *DeSeq*. The *edgeR* software package pioneered the simplification of DGE analysis in RNAseq by providing a series of functions and a standardized pipeline capable of translating complex analysis designs into basic code [97]. *DeSeq* followed by attempting to make improvements on the *edgeR* pipeline, adding more flexibility and bias control [102]. Both tools have undergone many improvements since the original release and have incorporated additional features that extended their functionality and made it possible to analyze almost any type of RNAseq experiment with either package. However, neither tool properly supports advanced modeling such as mixed-effect and longitudinal modeling since advanced modeling typically requires custom solutions.

Regardless of the tool used, results for DGE analysis are typically visualized using the same types of plots. Multi-dimensional scaling (MDS) or principal component (PC) plots are used to check the objective clustering of samples and are based on groups of genes with the greatest orthologous variance. MDS and PC plots help confirm that samples assigned to each analysis group have similar gene expression trends and can assist in identifying outliers or mislabeled samples. Smear plots show the $\log_2$ fold change against the average expression (usually CPM normalized) for all genes between two groups of

samples. Smear plots are useful for checking that the data fits the assumptions of DGE analysis; specifically, that the majority of genes are low to medianly expressed, that expression of genes is not substantially biased by the treatment of one group, and that the model properly weighted identification of significant differentially expressed genes away from genes with extremely low or high read counts. Volcano plots, with significance on the y-axis and $\log_2$ fold change on the x-axis, separate genes by degree of significance and are good for highlighting the genes with the largest differences between two groups.

Heatmaps show raw data for each gene across all samples in a dataset and illustrate contrasting expression values between sample groups. To emphasize differences heatmaps typically arrange data in obvious trends and can implement clustering algorithms to sort data by gene values, sample values or both. However, heatmaps can be biased through selection of specific genes, normalization of the data and the clustering process. Additionally, plotting a large number of genes usually results in an unreadable or superficial plot. Heatmaps are best reserved for displaying small sets of curated genes, or qualitative grouping of samples and any conclusions suggested by heatmaps should later be supported by additional quantitative analysis techniques.

2.5 Differential Splicing Analysis Methods

Special considerations for splicing analysis begin during mapping. Because splicing quantification methods often rely on mapped reads to identify unannotated transcript structure, and quantification of reads is heavily reliant on accurate mapping of reads across junctions, additional parameters need to be set to ensure the quality of mapped

data. Besides mapping quality, the minimum number of bases mapped to an exon and the maximum distance a read can span across an intron are set. A minimum of six bases inside an exon ensures a one in 4096 probability of a false-positive alignment and is a high bar when used in conjunction with a maximum intron size of 300,000. When mapping high-quality reads for gene expression, trimming adapters and low-quality bases is an optional step as the mapping algorithm will automatically soft-clip mismatched bases from the ends of an alignment. However, for splicing analysis it is common to trim reads beforehand and then force the aligner to map the entire read to maximize the number and quality of junction reads. The mapped insert size (distance between paired-end reads) is also useful to ensure quality as it can be used to selectively count reads that map to the same transcript within a reasonable distance and eliminate unrealistic mapping results [91].

Differential splicing analysis methodology depends on the quantification technique used. isoform-centric approaches compare the expression of whole isoforms using approaches similar to differential gene expression techniques [103,104]. Methods for differential isoform-centric splicing analysis can include procedures for grouping isoforms and summarizing differences in expression at multiple levels [103]. For example, expression at the coding sequence, isoform- and gene-levels can be compared [103]. This dissertation is focused on exon-centric analysis approaches. Therefore, details related to analysis of isoform-centric data will not be reviewed. Exon-centric approaches use a specialized statistical framework to compare the distributions of PSI values and separate significant differentially spliced events [92].

2.5.1 Normalization of Exon-centric Splicing Data

Normalizing exon-centric splicing data for visualization is very similar to normalization of gene expression data. If the goal is to understand read coverage across an event, per-base RPKM can be used. Because the PSI is a simple ratio, it negates the influence of read coverage bias. Therefore, when visualizing differences in splicing outcome, the PSI value (and optionally the confidence interval) can be used directly. Differential splicing analysis approaches usually begin with raw count data and normalize by converting the counts to PSI [91,92]. As with differential expression analysis, technical and biological variability are corrected for during modeling.

2.5.2 Analysis Tools and Presenting Results

The current gold standard tool for analysis of differential exon-centric splicing analysis is Replicate Multivariate Analysis of Transcript Splicing (*rMATS*) [92]. While other tools have been developed, *rMATS* is the only method that accounts for isoform uncertainty, biological variability and considers information about splicing patterns from junction reads [92]. First, *rMATS* models isoform uncertainty by classifying reads as either inclusion or exclusion reads (considering the structure of the five types of splicing events; A5SS, A3SS, SE, ME and RI) and fitting a binomial distribution for each sample PSI using the effective lengths of the inclusion and skipping event spaces [92]. Next, biological variability within groups is modeled through fitting a normal distribution with mean equal to the logit transformation of group mean PSI [92]. Finally, a likelihood ratio test is applied to determine the probability of group distributions differing more than a user-defined threshold [92]. Using a threshold for the likelihood ratio test restricts

significant events by biological relevance, but a threshold of zero can be used if the minimum effective change in PSI is not known. Filtering for biological relevance can also be done downstream of significance testing by removing events that to not meet a minimum difference in group mean inclusion levels.

Visualization of splicing data is an important confirmatory step in splicing analysis and begins by loading aligned BAM files, a reference genome and annotation into a genome browser. Reads supporting individual events can be checked by mapping back the exon positions from the spliced event coordinates. Inspecting reads that map to exons and junction sites is essential for checking the accuracy of count data, identifying poorly mapped reads, observing the influence of parameters used during mapping and checking the annotation used to infer the transcript structure. Sashimi plots can be produced with BAM files and show the per-base RPKM and junction read counts across a spliced event [105]. In sashimi plots, the field of view is restricted to only include exons relative to a single splicing event. To summarize results across multiple events, a volcano plot with PSI on the x-axis and p-value or false discovery rate (FDR) on the y-axis is frequently used. Standard box and violin plots are useful for showing the distribution of PSI values within sample groups.

2.6 Machine Learning and Predictive Modeling

Machine learning is the process of training predictive models with experimental data to identify patterns that are not otherwise observable [18]. Machine learning is particularly useful for analyzing large and complex datasets; where more basic analysis methods,

such as plotting, correlation, and cluster analysis, are not powerful enough to reveal hidden trends in the data. Because modern biological experiments often involve high-throughput techniques and can incorporate many different types of data, the importance and popularity of machine learning is growing [18].

Machine learning is used to meet one of three basic goals: 1) to produce a predictive model that can determine an outcome when given never-before seen data; 2) to divide, classify or cluster observations into groups; or 3) to identify underlying features that influence the outcome of a process, condition or event. Data used as input for modeling can be classified into two types: continuous data, such as a range of numbers like the temperature outside; or discreet data, like if an individual is wearing a black, yellow or red shirt [18]. In biological fields, machine learning models are typically used to identify underlying features that explain a disease, phenotype or trait; because of this, it is important that a machine learning model is robust, reproducible, and interpretable. The goal of the model, the data used to train it, and the desired properties influence the type of machine learning model used.

2.6.1 Types of Machine Learning Models

Machine learning models are divided into two primary groups; models that predict a continuous outcome and models that classify observations into a limited number of groups [18]. There are a wide variety of algorithms, each with their own unique mathematical basis, that accomplish modeling. Even within a single modeling strategy, there are many variations and parameters that can be applied to fine-tune the model

produced. Three popular modeling approaches are regression, random forest, and neural networks.

Regression is a technique that attempts to calculate the parameters for a line equation where one or multiple covariates are weighted, and the total sum of the products between the weights and covariates is the prediction [18]. This technique is very powerful because the output of the linear equation can be controlled using a link function and one of many standard probability distributions. When defining the link and data distribution, regression is instead referred to as generalized linear modeling. The link function transforms the sometimes-non-linear output of an equation to be linear, with the distribution of the predicted values dependent on the mean of values from the linear equation. Linear regression is used to generate continuous predictions; however, logistic regression, where the output of the linear equation is converted to a probability between zero and one, can be used to classify observations [18]. Regularized regression is an advanced from of regression that applies additional weights to covariates during model fitting and can control the influence or restrict the number of covariates in the final model [18]. Ridge regression is one technique for controlling the maximum contribution a covariate can have and Lasso is a technique for controlling the number of covariates [106]. The method known as “Elastic net” uses Ridge and Lasso together, balancing the contribution of each approach through a ratio [106]. Elastic net is a popular technique in biological data modeling because it can integrate many types of biological data and select the most influential features from the covariates thereby reducing model complexity.

The random forest approach to machine learning is a non-linear classification technique that relies on a simple algorithmic building strategy [18]. Random forests are comprised of many individual decision trees, each of which votes on the predicted outcome from the ensemble [107]. Bootstrap aggregation, or bagging, is used to sample data observations that will be used to build a tree. Bagging randomly selects a subset of observations from the total dataset with replacement, meaning that the same observation may be selected more than once [107]. The "bagged" observation set is the same size as the original dataset, but with duplicates, and represents approximately 63.2% of unique data (due to the effect of replacement). Trees are constructed by randomly selecting a subset of predictive features, calculating the effectiveness (using information gain or entropy) of the features ability to divide the data, and splitting the data by the best predictive feature [107]. The process of selecting features and adding decision nodes proceeds down the tree until all observations are classified [107]. The construction of random forests is primarily controlled through two parameters: *ntree*, which limits the number of trees in the forest; and *mtry*, which limits the number of predictive features randomly selected for each tree [108]. The parameter *mtry* usually depends on the total number of predictive features in the dataset, with the square root of predictive feature count used for classification models and the predictive feature count divided by three for regression models. The final outcome of random forest is decided by comparing the number of trees that select each outcome. The outcome with the largest number of trees supporting it wins [107]. Accuracy of a random forest can be estimated internally by testing observations not used in building specific trees, called out-of-bag observations, with only the trees the observations were excluded from [107]. The relative contribution, and therefore

interpretability of the model, can be assessed through calculating predictive feature importance. Importance is calculated by permuting the value of a feature within a tree, using the out-of-bag samples to find the loss of accuracy, and then averaging the loss across all trees with the selected feature [107].

Neural networks are named after their operational similarity to neurons in the brain and were originally designed to study brain function [18]. Like random forest, neural networks are composed of many individual learners. The basic decision unit of a neural network is a synthetic neuron, also called a perceptron [18]. Like real neurons, each synthetic neuron takes basic input values, sometimes aggregated from multiple upstream neurons, and applies a mathematical function converting the signal to output [18]. The mathematical equation within a synthetic neuron is comprised of a weight, which like regression is multiplied against the input, a bias term and an activation function [18]. The bias term is used as an offset to further adjust the weighted output and the activation function converts multiple inputs into a single output [18]. The sigmoid function, for example, is commonly used to represent the action potential of a real neuron and will return 1 if an input threshold is met or zero if not. There are many types of neural networks, each of which has many options for customizing the model input and training process [18]. While neural networks generally offer the best performance of all machine learning approaches they are not very reproducible, require an extremely large amount of data to train, and are typically not interpretable [18]. Because biologists are primarily interested in the underlying explanation behind predictions, neural networks are not frequently used [18].

Finally, each machine learning algorithm is subject to inductive bias, the predisposition for a model to solve a task a certain way, that is dependent on the process used to fit the model [18]. For example, the inductive bias of the random forest modeling approach is that it summarizes data as a series of binary decisions. Inductive bias can be a limitation of the modeling approach or an advantage. Random forest will excel at identifying sequential relationships in data that determine the outcome. If the data features are sequentially related then random forest will likely produce an accurate and efficient model, even if the data are complex. If the data is not related sequentially then random forest may not perform as well as other modeling techniques.

2.6.2 Feature Selection, Outliers and Missing Data

Regardless of the machine learning approach used, feature selection (also known as dimensionality reduction) must be done. Feature selection is the process where non-informative features are eliminated from the dataset before training to reduce the model complexity and prevent overfitting [109]. When selecting features it is important to use an objective approach to avoid biasing the model later. The most basic approach to feature selection is filtering based on simple data metrics like sparsity and variance. Features with many missing or zero values, and those with identical values across observations, will not contribute to the predictive power of the model and will only introduce noise. Advanced feature selection techniques can be used after basic filtering to remove even more features, thereby improving the speed of training and reducing the storage space required.

There are many advanced techniques for feature selection, and the method selected depends on the type of data, the model being trained, and the relationships within the data a researcher wants to retain [109,110]. Feature selection techniques can be supervised, where the outcome is considered during the calculations, or unsupervised. One of the more popular unsupervised techniques is principal component analysis (PCA) [111]. PCA uses eigenvectors to identify features responsible for the greatest orthologous variance, summarizing their feature sets into single numeric values [111]. Feature sets, or components, are sorted by the total variance across samples they account for and the features from top *N* components, *N* determined by a significant drop in variance for components ranked lower, are retained for modeling. Tests for feature independence or outcome correlation, such as the chi-squared test, Pearsons's correlation coefficient, and linear discriminant analysis are also used to reduce the number of features [109].

Observations may also be removed before modeling if they have extreme values for the predictive features or outcome; this prevents the model from being influenced by outliers. Another way to handle outliers is to transform the data with an equation. For example, if the majority of feature values are about the same order of magnitude, but some observations have a 10-fold or greater difference then log transforming the data will restrict the contribution of those values in the modeling process.

Some machine learning approaches, like random forest, require values for all observations in order for a feature to be used in modeling. In such cases, missing data must be supplemented with pseudo-values that are representative of the feature but do not

bias the learning process. One strategy for filling missing data is population-based statistics. For example, one can simply substitute using the mean or median of the feature values. Another approach is random or semi-random imputation [112]. In random imputation a value is drawn, usually with replacement, from a randomly selected feature and sample. In semi-random imputation, first the available values from filled samples of the missing feature are compared to the values in other features and the replacement value will be drawn from a subset of features with values that are similar to the missing value distribution. Finally, some models prefer that any features with missing values be removed from the dataset. While removing features due to missing data tends to remove an extensive amount of data from other samples, it also prevents training bias due to unfavorable imputation.

2.6.3 Model Training

Training of machine learning models begins by dividing the available data observations into training and testing datasets [18]. All operations associated with training models, including feature selection, happen within the training data set [18]. The testing dataset is reserved to assess the performance of a trained model on never-before-seen data [18]. Data can be divided in any ratio, but commonly 70% training and 30% testing or 90% training and 10% testing are used. Testing and training datasets are divided randomly. However, before any operations are conducted on the data it is recommended to set a random "seed"; a number that is used by a random number generator to initialize a randomized drawing routine. Setting a seed is important because many machine learning approaches require randomized sampling or parameter tuning and a seed allows someone

to reproduce the model building process later. If building a model for classifying data, it is important to balance the number of observations in each class [18]. Otherwise, the model may be biased to predicting one class over another. If the goal is to compare multiple machine learning approaches, identical training and testing data sets must be used for all models. A validation dataset may be used to compare the performance of a trained model on data from another source. If a validation set is used, it is frequently obtained from a separate experiment, external publication or public database.

There are three basic types of machine learning techniques; unsupervised, semi-supervised and supervised learning. In unsupervised learning, no labels or outcome are provided with the observations and the data is separated by inherent feature differences [18]. In semi-supervised learning, some of the data is labeled and learning is initialized or corrected using known information [18]. In supervised learning, the labels or outcome are known and the model learns how to optimize the predictions accordingly [18]. Every machine learning approach has parameters that can be adjusted to fine-tune the model building process. Basic parameters dictate the power or performance of the model and are usually subject to an optimization procedure to ensure the best outcome [18]. Hyperparameters control how quickly a model is trained and can have a smaller influence on the model outcome, but greatly influence the time required for training [18]. There are two popular methods of tuning model parameters during training: grid search, where a list of possible parameter values is supplied to the model and training proceeds through each one; and gradient descent, where a large approximate starting value is set to a parameter and then sequentially smaller values are tried in later training iterations, honing the

optimum solution. The improvement in model performance between repeat trainings is determined by a loss function [18].

In machine learning, the loss function calculates the difference between the ground truth and predictions of a trained model [18]. Popular choices for the loss function include R-squared, adjusted R-squared, root-mean-squared error, mean absolute error, accuracy, kappa and entropy. Because there are two basic types of machine learning approaches, continuous prediction and classification, methods to calculate loss fall into the same two families. R-squared, adjusted R-squared, root-mean-squared error and mean absolute error are all metrics used for calculating loss in continuous prediction. R-squared and adjusted R-squared are focused on assessing the variance explained by a machine learning model through comparing the variance in the predicted values to the variance in the ground truth values. Root-mean-squared and mean absolute error are focused on quantifying the deviation between the predicted values and ground truth values. Accuracy, kappa and entropy are metrics used in classification. Accuracy is simply the fraction of correct classifications out of total observations. Accuracy can be deceiving though as there may be an imbalance between the number of observations in each class. Because of this, methods like kappa and entropy consider class membership to better assess model balance [113].

When training machine learning models it is common to repeatedly sample a portion of the training data, fit a model and asses its effectiveness against the unselected training data via the loss function. This process is referred to as cross validation and is useful for

adjusting hyperparameters, estimating performance on external data, and limiting the influence of random effects introduced during the modeling process by allowing researchers to select the best trained model [18]. Cross validation can be combined with grid search or gradient descent to achieve a thorough but time-intensive training procedure.

2.6.4 Model Evaluation

While the loss function is used to assess performance and adjust parameters during learning, the overall performance of the best fit model is estimated after training using the testing data set. The metrics for evaluating continuous models are the same as the loss functions during training. On the other hand, performance for classification models is usually reported in terms of accuracy along with qualifying metrics like sensitivity, specificity, precision, recall and F1 score. Qualifying metrics are calculated using the four ground truth classification counts: true positives (TP), correctly assigned positive observations; false positives (FP), incorrectly assigned negative observation; true negatives (TN), correctly assigned negative observation; and false negatives (FN), incorrectly assigned positive observation. The classification metric equations are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2}{\frac{1}{Sensitivity} + \frac{1}{Precision}}$$

It is also common to visualize the performance of classification models using a receiver operating characteristic (ROC) curve. A ROC curve plots the sensitivity (y-axis) against one minus the specificity (x-axis). The larger the area under the ROC curve, the better the overall performance of a classification model is.

Comparing the performance metrics from training and testing data in machine learning models is an important check to understand the real-world value. Underfitting is when a model has poor performance on training data and indicates that training did not fully capture the relationship between the outcome and predictive features [18]. Overfitting is when the training performance is excellent, but the model fails to predict testing or validation data sets correctly [18]. Overfitting is especially undesirable in biological data as it can cause researchers to report incorrect associations that are later disproven or detrimental to future research. Revisiting the selected machine learning approach, training procedure, and underlying dataset are all steps that can be taken to improve performance of machine learning models.

2.7 Network Analysis Methods

Network or systems biology is the study of biological interactions through computational modeling of synthetic networks[114]. Networks exceed at representing large volume, high complexity interaction data in a digestible format, and network mathematics is a well-developed field that offers a wide array of techniques for characterizing interactions. Analyzing data using a network-based approach allows for discovery of complex interactions that are otherwise invisible [114]. Network representations of biological data are extremely flexible and can support a wide array of data types. Network-based approaches are also well suited for integration of multiple data types [114]. High-throughput genomics has enabled the development of network systems biology by generating large volumes of multi-omics data that requires integration [114]. High-throughput genomic data also has extraordinary potential to inform researchers on the underlying rules of cellular biology.

2.7.1 Network Components, Design and Structure

The components of a network are very simple. Network entities are represented through nodes, sometimes called vertices, that can be any biological gene, protein, molecule or feature in general [114]. Nodes are connected to each other pairwise through links, sometimes called edges, that represent the relationship between the two entities [114]. Links can be binary or weighted, and directed or undirected [114]. Using a network framework, biological phenomena like protein-protein interactions, transcription factor regulation, enzymatic pathways and small-molecule signaling can be represented with ease. Networks are also defined by a mathematical structure, based on the number and

probability of connections between nodes [114]. The number of connections a node has is called the degree [115]. The two most common network structures in biology are random structure and scale-free structure.

In a random network, each node has approximately the same probability of connecting to another, and the degree of each node follows a Poisson distribution [114]. Random networks show no connection preference to individual nodes on average, and if the network is disrupted by removal of a node approximately the same number of edges will be destroyed. The distribution of node degree in a scale-free network follows a power law, where the probability of new node connections is dependent on the number of existing connections [114].

Scale-free networks mimic the evolutionary nature of biological interactions in that certain genes have substantially more interacting partners than others [114]. One reason biological networks have evolved to be scale-free is gene duplication [114]. Gene duplication creates an identical copy of a gene and therefore its translated protein retains the interacting partners [114]. Additionally, if an interacting partner is duplicated it is more likely to have a connection to the previously duplicated gene pairs [114]. Evidence supporting gene duplications role in network structure has been found in that evolutionarily older genes tend to have more interacting partners [114]. Another reason biological networks have evolved to be scale-free is that a scale-free network provides greater systemic protection in the off chance a node is removed. For example, if a loss-of-function mutation occurs randomly across all genes, then in a scale-free network it is

much more likely that a mutation will occur in a gene with few connections since most connections occur in a limited number of nodes [114].

2.7.2 Network Metrics

Network metrics are useful for describing structure within a network and can also be essential in algorithms that search for network communities. One example of a node-level metric, node degree, has already been introduced. In a directed network a node can have both an in-degree, the number of connections directed to the node; and an out degree, the number of connections directed from the node [115]. In a weighted network, the node strength is the sum of all weighted edges connected to a node [115]. Strength can also be broken down to in- and out-strength in a directed network [115]. Betweenness centrality is the count of the shortest paths between neighboring nodes that flow through a target node and indicates the importance of a node in transferring information across the network [115]. Additional metrics have been developed that are more specific to the type of information a network represents.

Network-level metrics are also useful for describing the overall shape and content of networks. Network-level metrics can also help researchers decide what analysis algorithms to use. Density is the total number of edges in a network divided by the number of all possible edges and indicates how saturated a network is [115]. Assortativity is the preference for nodes with similar characteristics to connect with each other and positive assortativity can indicate rapid spread of information [115]. Transitivity

measures the number of node trios with 3 edges compared to trios with only 2 edges and signifies the clustering potential of a network [115].

2.7.3 Network Module Discovery

Network module discovery, also called community identification or clustering, is an extremely important field of research that is essential to understanding network structure. A network module can be loosely defined as a set of nodes with more connections between members that would be expected at random [116]. This definition also implies that nodes in a module are connected more frequently to each other than they are to other nodes in the network [116]. Many different module discovery approaches exist, all of which are tailored to the type of data within the network or the network structure and metrics. Although efforts have been made to compare module discovery algorithms, often there is no single best solution for biological data and it is common to combine multiple techniques or results from a number of algorithms to achieve a better outcome.

There are two components of a module identification algorithm; the method the algorithm uses to find linked nodes in the network and the scoring function. One method for identifying linked nodes in a network is random-walk [117]. Random-walk based algorithms begin at a starting node and randomly select a connected node to add to the module. The edge between the starting node and new node will be considered in the module scoring function. The random-walk is then repeated from the new node and the process continues until a condition is met, such as a specific number of iterations or a marked decrease in the combined module score. One example of a measure that is used to

score modules is called modularity [116]. To calculate modularity, the squared probability of both intra-cluster and inter-cluster edges is subtracted from the probability of intra-cluster only edges [116]. The overall probability of a module occurring at random in a specific network can also be calculated [118]. The module probability can then be used to filter weaker modules and select the most significant communities for downstream analysis.

Chapter 3 Differential Gene Expression Analysis of Concussed Athletes Reveals Differences in Immune Signaling Pathways and Immune Cell Types

3.1 Acknowledgement of Contributions

The analysis performed in chapter 3 was made possible through collaboration with members of the Concussion Assessment, Research and Education (CARE) Consortium. Participant blood specimens were collected through a partnership between academic, medical and sport professionals at numerous collegiate institutions across the United States. Specimens were transported to the Clinical and Translational Sciences Institute Indiana Biobank for storage and were processed and sequenced by the Indiana University Center for Medical Genomics. I performed quality control, computational analysis and visualization of the sequencing data.

3.2 Introduction

Concussion is a type of mild traumatic brain injury resulting in brief loss of normal brain function due to a head injury [119]. Despite being a coded diagnosis in ICD10, concussion is a non-specific term and symptoms can range in type and severity [119]. Concussion poses a major public health threat. A 2017 study found 15.1% of high school students (2.5 million) playing a sport, received at least 1 concussion in the last 12 months [120]. TBI in service members has been more prevalent in recent conflicts like Iraq and Afghanistan [121]. Between 2000 and 2015 over 300,000 service members sustained a TBI and of those more than 80% were concussions [121]. Furthermore, concussions from automobile accidents and accidental falls are the most common mechanisms of concussion, and account for almost 75% of all traumatic brain injury hospitalizations

[122]. Quickly identifying the symptoms of concussions and intervening to prevent further brain injury is essential to recovery. Currently, athletic programs and other institutions are advised that immediate removal from activity when a concussion is suspected is the best course of action [13]. Unfortunately, concussion symptoms can be hard to detect. Studies show athletes underreport concussions not only because the symptoms can be subtle, but also because of their commitment to the activity [13,123]. Development of a rapid diagnostic test to assess concussions would greatly improve the diagnosis of and intervention after concussions.

Although many individuals recover from concussions within two to three weeks, some individuals experience extended lag times in recovery [11]. Persistent symptoms of concussion, also known as post-concussion syndrome, can last months [11]. Research into concussion diagnostics has largely focused on blood-based biomarkers. Yet, no biomarkers or underlying genetic factors have been identified that help explain post-concussion syndrome [11]. Little is known about the gene expression signature following concussion, or potential expression biomarkers that could aid in diagnosis and recovery prediction. Given that blood-based assays have identified potential biomarkers, we hypothesize that gene expression-based diagnostic or prognostic biomarkers may also exist [124]. We anticipate investigating post-injury gene expression signatures will reveal differentially expressed genes that are relevant to concussion response and recovery. We intend to identify processes coordinated through changes in gene expression and uncover trends in gene expression that are informative for long-term recovery prognosis.

This study describes the initial findings from whole transcriptome RNAseq analysis on concussed individuals, spanning preseason baseline and multiple post-injury timepoints. This study also introduces a comprehensive dataset that will be a valuable resource for researchers investigating the consequences of head impacts, traumatic brain injuries and gene expression biomarkers.

3.3 Materials and Methods

3.3.1 Study Participants and Sample Collection

The Concussion Assessment, Research and Education (CARE) Consortium is a partnership between the Department of Defense and the National Collegiate Athletic Association that was formed to further the study of concussion neurobiology and consequences of exposure to repetitive head impacts [125]. To date, CARE has enrolled over 50,000 volunteer collegiate varsity athletes and military academy recruits participating in competitive sports [126]. For this study, whole blood was collected from a cohort of 552 varsity athletes and military cadets participating in various sports between 2015 and 2019. Samples were drawn into PAXgene tubes (BD Biosciences, Cat. No. 762165) at six timepoints: baseline (Base), before injury; post-injury (PostInj), taken within six hours of injury; 24-hour (24hr) taken between 24 to 48 hours after injury; asymptomatic (Asymp), when an athlete begins return-to-play progression; seven days post unrestricted (7PostUR), when an athlete has been cleared for return-to-play; and six months (6Mo) from the date of injury [125]. Athletes were divided into three groups based on injury status: non-contact controls (NCC), athletes who did not participate in contact sports; contact controls (CCT), athletes who participated in contact sports but did

not sustain a concussion; and injured athletes (INJ), athletes who sustained a concussion during an athletic event. [ansote] The full study protocol is available from (Broglio et al. 2017) [125]. Regardless of group status, only participants with baseline samples were retained for analysis.

3.3.2 Sequencing Library Preparation

Total RNA was extracted from blood cells using the PAXgene Blood RNA kit (Qiagen) followed by DNase I treatment to remove contaminating genomic DNA. Dual-indexed strand-specific cDNA libraries were prepared from eluted total RNA using the Kapa mRNA HyperPrep kit (KapaBiosystems) along with QIAseq FastSelect Human Globin removal kit (Qiagen). Libraries were prepared in a 96-well plate using a Biomek FxP Laboratory Automation Workstation. Each plate was pooled using the QIAgility Automation System. Pooled libraries were loaded onto a flowcell that was sequenced with 2×150 bp paired-end configuration on a NovaSeq 6000 instrument (Illumina, Inc.).

3.3.3 Gene Expression Quantification and Differential Expression Analysis

Sequence reads from RNAseq experiments were aligned to the human genome (hg38) using *STAR* v2.5.2b [85]. Gene expression levels were quantified by counting the number of RNA fragments aligned to exonic regions of genes using the program *featureCounts* [127]. The data were analyzed as individual timepoints compared to baseline to avoid eliminating participants with missing time point data. Differential expression analysis was performed with *edgeR* using negative binomial generalized log-linear modeling (GLM) and likelihood ratio tests [97]. When calculating distribution parameters with the

estimateDisp function, the robust option was used to nullify extreme outliers. Genes with very low read counts were removed before differential expression analysis to reduce the number of individual statistical tests performed and to avoid inflated significance values. Genes were filtered if they had less than 1 count per million mapped fragments (CPM) in a minimum number of samples at each timepoint. Given the large number of samples in each group, we defined the minimum number of samples as 25% of the smallest group in the comparison (i.e., injured vs. contact controls). At each time point, the minimum sample thresholds (N) were: PostInj (48); 24h (58); Asymp (61); 7PostUR (57); and 6Mo (42).

Differential expression analysis was performed at each timepoint compared to baseline and a contrast between the injured and contact control groups was calculated. Comparison of contact control and non-contact control participants at the PostInj timepoint was used to define the gene expression background levels for concussed athletes on rest following injury. In all comparisons, genes with Benjamini-Hochberg false discovery rate (FDR) ≤ 0.05 were considered significant [128].

3.3.4 Gene Ontology Analysis

Gene ontology (GO) analysis is a type of enrichment analysis where the top differentially expressed genes in an experiment, defined by a significance cutoff, are matched against reference gene lists that have been annotated to biological terms or functions. Enrichment tests determine if a term is significant by comparing the number of matched genes in a list to a random background. GO analysis was performed in R using the *clusterProfiler*

package [129]. Differentially expressed genes with FDR ≤ 0.05 were converted to Entrez gene IDs using Biomart [130,131]. Biological process, cellular component, molecular function, and Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic reference lists were searched with *clusterProfiler* to determine enrichment for terms at the PostInj timepoint. Only terms with FDR ≤ 0.05 were considered significant. Terms with enrichment lists containing ≥ 50% common genes were merged.

3.3.5 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is a type of enrichment analysis that does not depend on a significance cutoff, making it ideal for situations where few or no differentially expressed genes are identified. Instead, genes are ordered by significance and a running score is obtained as matching proceeds down the full list. GSEA was performed on differential expression results at all timepoints, with genes ranked by the sign of the fold change multiplied by the -log(p-value) from differential expression analysis, using the GSEA v4.1.0 app and MSigDB v7.3 [132,133].

3.3.6 Deconvolution Analysis

Deconvolution analysis is the process where cell type proportions can be estimated from bulk RNAseq data based on marker gene expression. Deconvolution analysis was performed with *CIBERSORTx* using raw read counts and default software normalization [134]. A GLM was used to test the difference between estimated percentages of cell types output by *CIBERSORTx*. The cell type percentage of the Base sample and the group were used as covariates to predict the cell type percentage of the timepoint sample. All

timepoints were tested, but only a few cell types from the PostInj timepoint were significant at FDR ≤ 0.05.

3.4 Results

3.4.1 Participant Demographics and Dataset

A total of 2,489 blood samples were collected from 552 athletes amongst all groups. Participants without baseline samples in the contact controls and injured groups were filtered, leaving 129 contact control, 230 injured, and 102 non-contact control individuals with a combined total of 2,125 blood samples. A breakdown of participant demographics is in **Table 1**. Because some sample collections were missed, unaccounted for, improperly recorded, or failed quality control, the number of samples at each timepoint differed between groups. During the study, nine contact control participants sustained a concussion and were subsequently reclassified as injured participants; as a result, samples from these athletes were present in both contact control and injured groups. Therefore, the baseline samples for these nine participants were duplicated for the reclassified sample sets, but the other blood draws for these participants (40 in contact controls and 33 in injured) remained unique in the dataset. In addition, one contact control athlete served as a control in two different seasons, and while the same baseline sample was used, unique blood draws separated by season distinguished the sample sets. The first contact control sample set for this participant was composed of five unique blood draws and the second set was composed of two unique blood draws.

We constructed 130 contact control and 230 injured sample sets, each having a baseline blood draw and at least one sample from a later timepoint. A summary of sample group numbers at each timepoint is provided in **Table 2**. The distribution of participant sample sets is shown in **Figure 1**. Non-contact control samples were used to represent time-based gene expression variance. For each of the 102 non-contact control participants, the first sample drawn was considered the Base sample and each subsequent sample for that participant was individually paired with the Base sample as a separate sample set. As a result, there were 428 non-contact control sample sets.

Factor	NCC	CCT	INJ
Total	102	130	230
Sex			
Male	82(80.4%)	99(76.2%)	182(79.1%)
Female	20(19.6%)	31(23.8%)	48(20.9%)
Age (SD[-])	19.3(1.2)	19(1.2)	18.9(1.2)
Military Status			
Military	0	37(28.5%)	126(54.8)
Non-military	102(100%)	93(71.5%)	104(45.2%)
Race			
African American	13(12.7%)	30(23.1%)	45(19.6%)
Asian	0	1(0.7%)	3(1.3%)
Hawaiian/Pac. Isl.[+]	1(1%)	1(0.8%)	4(1.7%)
Indian/Alaskan	3(2.9%)	0	0
MSU*	7(6.9%)	9(6.9%)	21(9.1%)
White	78(76.5%)	89(68.5)	157(68.3)
Ethnicity			
Hispanic	8(7.8%)	9(6.9%)	13(5.7%)
MSU[*]	1(1%)	12(9.2%)	29(12.6%)
Non-Hispanic	93(91.2%)	109(83.8%)	188(81.7%)
Injury Sustained			
Competition			88(38.3%)
Practice/Training			132(57.4%)
Outside Sport			10(4.3%)
Sport			
Football		61	101
Ice Hockey		10	20
Soccer		29	47
Lacrosse		8	16
Rugby		9	27
Wrestling		2	7
Cross Country/Track	37	1	2
Intramurals		6	9
Softball	7		1
Baseball	37		
Basketball	13		
Field	8		
Other		3	
Unknown		1	

Table 1 Cohort demographics of CARE participants. "Other" includes skiing, boxing and handball. - = Standard Deviation; +Pac. Isl. = Pacific Islander; *MSU = Multiple, skipped, or unknown.

	CCT	INJ
PostInj	125	96
24h	117	173
Asymp	123	194
7PostUR	115	171
6Mo	85	138

Table 2 CARE sample group balance. Sample group numbers at each timepoint for contact control and injured participants.

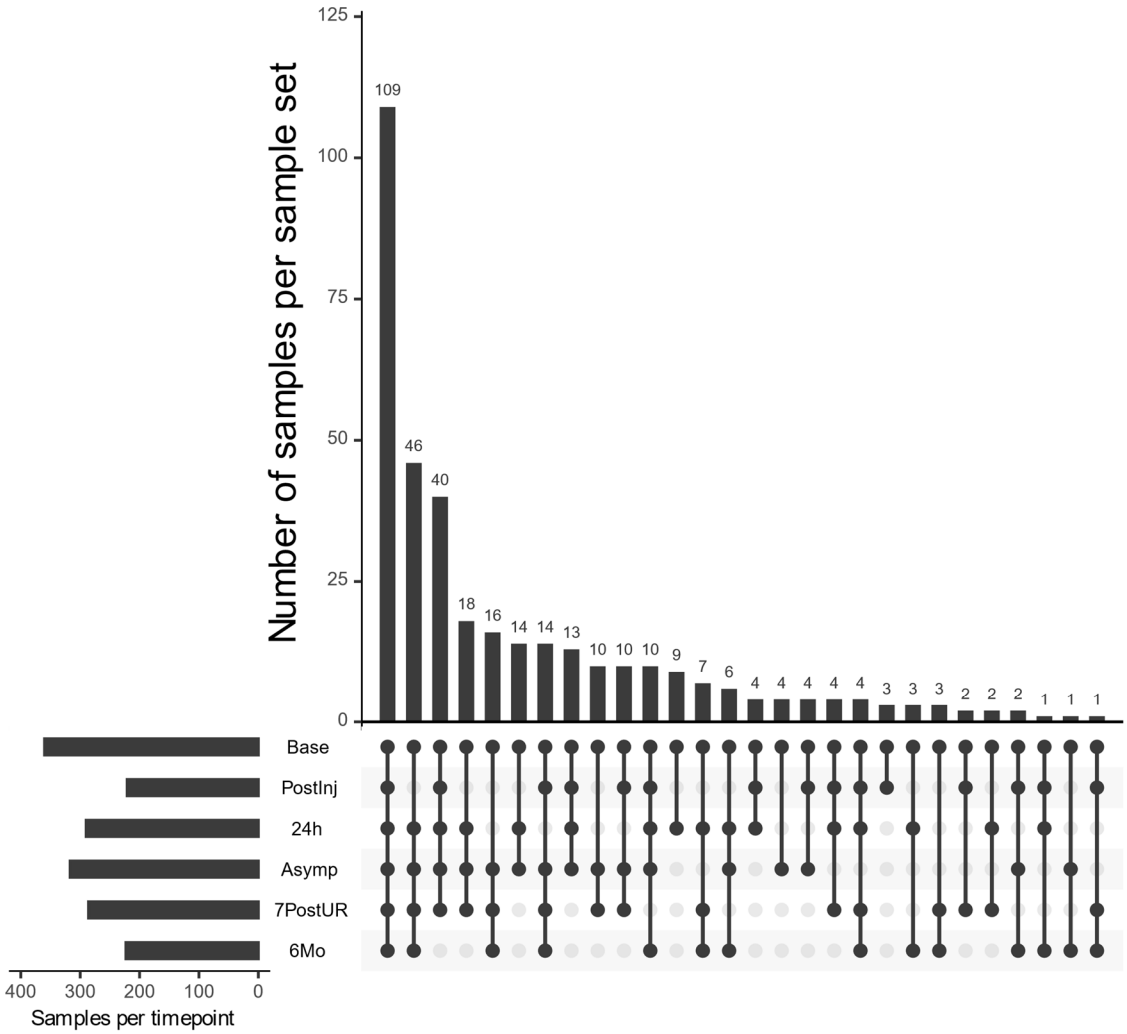

Figure 1 CARE sample set distribution. Upset plot showing set membership for contact control and injured blood samples. The main panel shows the total number of observations with multiple time points. The time point coverage is annotated below. A black dot indicates a blood sample was drawn at a given time point. The left barplot indicates the total number of samples at each timepoint.

3.4.2 Differential Gene Expression Analysis Reveals Many Altered Genes Immediately Following Concussion

To investigate the timeframe whereby sports-related concussion altered gene expression patterns in peripheral blood, we performed differential gene expression analysis for each timepoint. We fit a model comparing a selected timepoint to the paired Base sample within contact controls and injured participants separately. Then, the injured and contact control models were contrasted to identify changes specific to injured participants. By first modeling within the contact control and injured groups, and then comparing between them, we controlled for gene expression changes resulting from contact sport participation and not associated with injury. We found that the highest number of significant differentially expressed genes occurred at the PostInj timepoint (N = 860) and that the number of differentially expressed genes was dramatically reduced by the 24h timepoint (N = 8). Volcano plots of differentially expressed genes at all follow-up timepoints are in **Figure 2**.

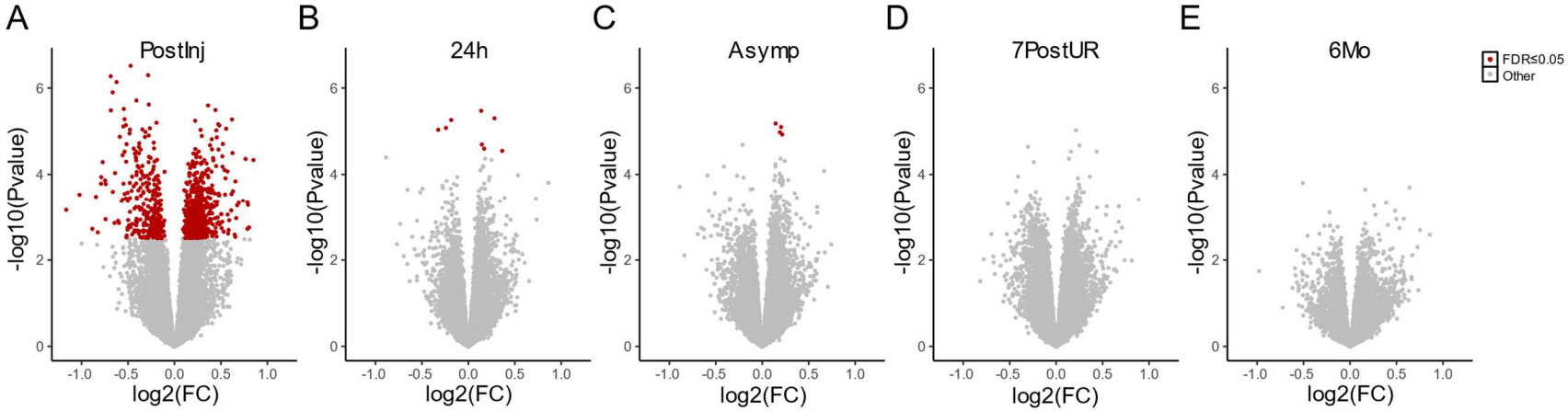

Figure 2 Volcano plots for CARE differential gene expression analysis. Genes with FDR ≤0.05 are in red.

3.4.3 Gene Expression Changes After Concussion Mirror Pathophysiology

Dysregulation of calcium metabolism and calcium dependent signaling is a known consequence of brain injury [135]. After concussion, the disruption of membranes in neuronal cells triggers ionic flux [135]. Cells attempt to restore homeostasis by activating ion pumps, including calcium pumps, which in turn consume ATP and starve the brain of energy [135]. Multiple genes related to calcium metabolism were altered in expression at the PostInj timepoint, including *CAMK2G*, *CAMKK2*, and *CAMKK1* which were all positively regulated in concussed participants. At least 15 components of solute transporters were altered in expression including four members of the SLC22 family (*SLC22A15*, *SLC22A16*, *SLC22A1*, and *SLC22A4*), each of which were increased in expression and can transport carnitine. Carnitine is synthesized from the amino acids lysine and methionine, and used in cells to transport long-chain fatty acids into mitochondria for energy production [136]. Together, the upregulated genes we observed related to calcium and energy metabolism suggested a compensatory effect following injury, and matched the pathophysiology reported for concussions.

3.4.4 Changes in Blood-based Protein Biomarkers Are Not Observed in Gene Expression Data

We also investigated known potential biomarkers for traumatic brain injury diagnosis [124,137]. Two FDA approved biomarkers used in the i-STAT TBI plasma test (Abbott), *GFAP* or *UCH-L1*, did not meet the minimum expression threshold for analysis at any timepoint. The gene encoding Tau, *MAPT*, also did not meet the minimum expression threshold for analysis at any timepoint. Neurofilament Light Chain, *NEFL*, was expressed

at all timepoints but no significant differences were observed between injured and control participants.

3.4.5 Gene Ontology Enrichment Analysis Identifies Activation of Immune Signaling Processes

To explore the biological function of differentially expressed genes after concussion, we performed GO term and KEGG pathway enrichment analysis on the differentially expressed genes at the PostInj timepoint. The resulting lists were ranked from smallest to largest FDR and the top two biological processes were neutrophil activation and neutrophil mediated immunity. Other significant biological processes were also related to immune response, such as "cytokine production involved in immune response", "positive regulation of cytokine secretion", "interleukin-1-mediated signaling pathway", and "regulation of CD4-positive, alpha-beta T cell activation". Inflammation following primary injury is thought to be one of the mechanisms of neuronal tissue damage in concussions [138,139]. Acute inflammatory response and upregulated cytokine production has also previously been observed in smaller studies [140,141]. In our study, we observed significant gene expression differences in multiple interleukin receptor genes at the PostInj timepoint, including *IL1R1*, *IL1R2*, *IL1RAP*, and *IL2RB*. Several other biological processes related to signal transduction pathways were found, such as regulation of small GTPase mediated signal transduction and protein autophosphorylation. Likewise, enriched KEGG pathways included natural killer cell mediated cytotoxicity, MAPK signaling pathway, and NOD-like receptor signaling activity (**Figure 3**).

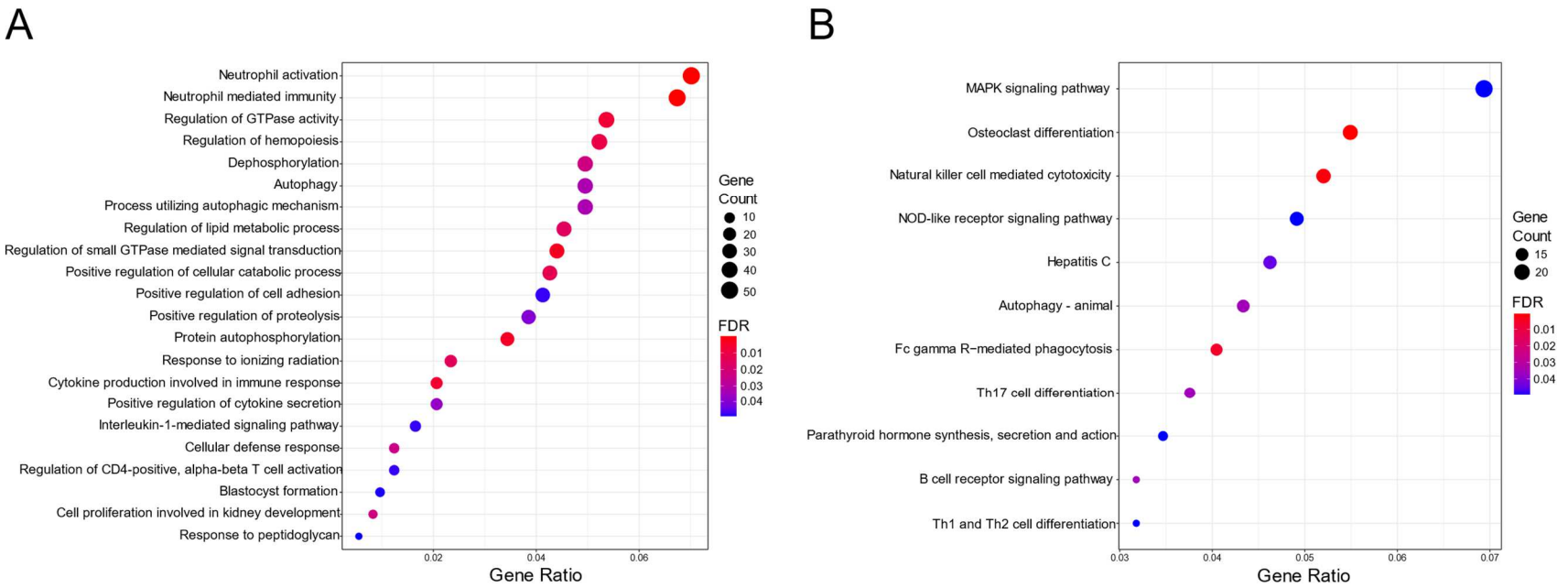

Figure 3 Enriched GO and KEGG terms in differentially expressed genes. Significantly enriched GO (A) and KEGG (B) terms at the PostInjury timepoint. The diameter of the point indicates the number of differentially expressed genes (FDR ≤ 0.05) matched to a term. Color is scaled from blue to red by increasing significance. "Gene Ratio" refers to the number of genes matched to a term compared to the total number of differentially expressed genes at the PostInjury time point.

3.4.6 Gene Set Enrichment Analysis Confirms Activation of Immune Signaling and Suggests Reversal in Immune Signaling During Recovery

Because the small number of differentially expressed genes at the later timepoints prohibited GO analysis, we used GSEA to compare enriched cellular processes and pathways at all timepoints. Select enrichment results using hallmark gene sets are in **Figure 4A**. Similar to the findings from GO analysis and KEGG pathways, GSEA also showed that the top-ranked pathways immediately following concussion were related to upregulation of immune-related signaling. For example, "*TNFa* signaling via *NF-kB*", "inflammatory response", and "*IL6 JAK STAT3* signaling" were all significantly positively enriched (FDR ≤ 0.05); each of which has also been strongly associated with response to concussion in multiple studies [142–144]. We then returned to our differentially expressed genes at the PostInj timepoint and identified multiple genes downstream of Janus tyrosine kinase (*JAK*) that were altered in expression including members of PI3K/AKT, MAPK and STAT signaling pathways. These genes include *JAKMIP1*, *JAKMIP2*, *PRR5L*, *MAPK13*, *STAT6*, and *BCL6*. Additionally, two PP2 regulatory subunits, *PPP2RB2* and *PPP2R5E*, were differentially expressed; PP2 being protein phosphatase 2, a serine/threonine phosphatase that targets MEK and AKT signaling cascade pathways [145]. Notably, significant enrichment for "*TNFa* signaling via *NF-kB*", "inflammatory response", and "*IL6 JAK STAT3* signaling" were also found at later timepoints, but with negative instead of positive enrichment scores.

We then asked if reversed changes in enriched pathways at later timepoints could be due to injured athletes being removed from play for recovery. We therefore explored enriched

processes at the PostInj timepoint between contact control athletes and non-contact control participants (**Figure 4B**). Interestingly, we observed that the immune signaling processes "*TNFa* signaling via *NF-kB*", inflammatory response, and "*IL6 JAK STAT3* signaling", were positively enriched in contact controls. Therefore, it appears that compared to athletes participating in non-contact sports, athletes participating in contact sports exhibit higher activation of certain immune signaling processes. These immune signaling processes become further elevated immediately following concussion but are then downregulated below contact control levels during recovery. GSEA results also indicate that some altered immune signaling pathways appear to remain repressed, compared to the contact control group, up to 6 months following a concussion.

We sought to confirm observations from the Hallmark gene lists and expand upon the results by performing GSEA with gene lists from two other reliable and popular databases, WikiPathways and Biocarta [146,147]. Pathways results using WikiPathways are in **Figure 5A** and Biocarta are in **Figure 5B**. In both WikiPathways and Biocarta, results previously observed in Hallmark gene sets were positively enriched (FDR ≤ 0.05) at the PostInj timepoint and negatively enriched at a later timepoint. Observed pathways included those associated with cytokine production and inflammatory response.

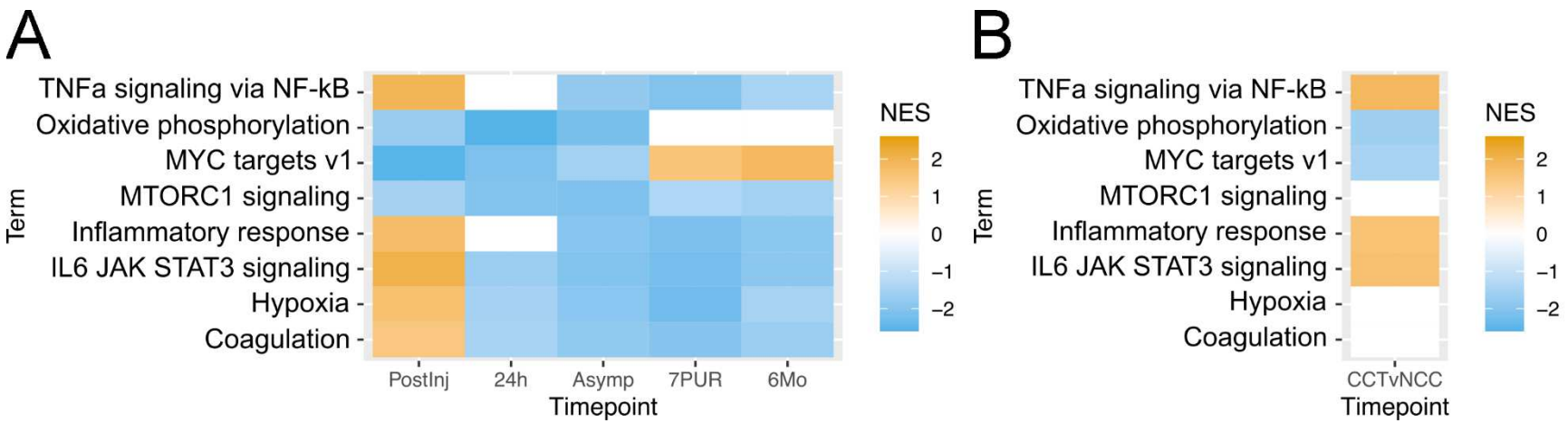

Figure 4 Hallmark GSEA of differentially expressed genes across timepoints. A. Normalized enrichment scores (NES) for Hallmark gene sets across all timepoints. B. NES for Hallmark gene sets using differential expression results from comparing the contact control group versus the non-contact control group at the PostInjury time point. Plotted terms in both figures were filtered from total results by selecting terms that were significant with FDR ≤ 0.05 at the PostInj timepoint and also significant with FDR ≤ 1e-5 at a later timepoint.

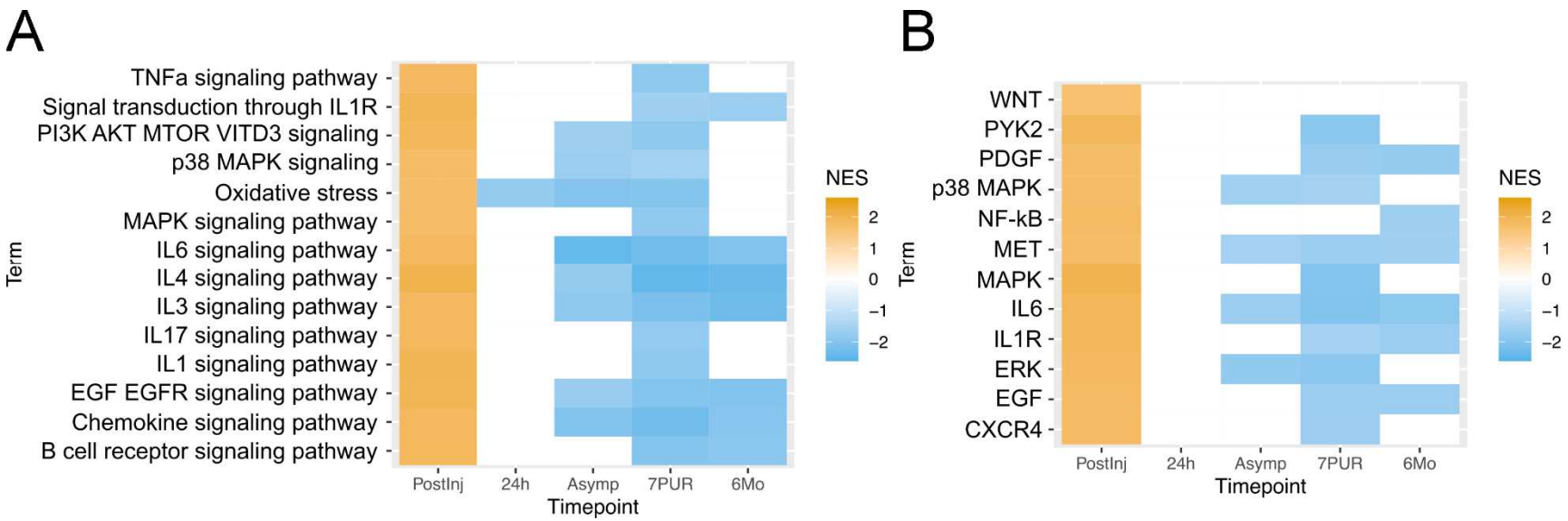

Figure 5 GSEA with terms from signaling pathway databases. NES for WikiPathways (A) and Biocarta (B) gene sets from GSEA across all timepoints. Plotted terms in both figures were filtered from total results by selecting terms that were significant with FDR ≤ 0.05 at the PostInj timepoint and related to pathways found in GO analysis, Hallmark GSEA or were otherwise immune-associated in literature searches.

3.4.7 Deconvolution Analysis Shows Increased Proportion of Neutrophils in Concussed Athletes

To understand possible changes in circulating cell type populations in response to concussion we performed deconvolution analysis. RNAseq counts were analyzed with *CIBERSORTx* to identify immune cell proportions. A GLM was used to test differences between injured and contact control groups at the PostInj timepoint. Two cell types were determined to be differentially proportioned; Neutrophils (FDR 2.3e-2, **Figure 6A**) were more prevalent in the injured group and resting natural killer cells (FDR 2.49e-5, **Figure 6B**) were less prevalent in the injured group. Our deconvolution results help support findings in another study where an increase in neutrophils at the site of injury was observed [148].

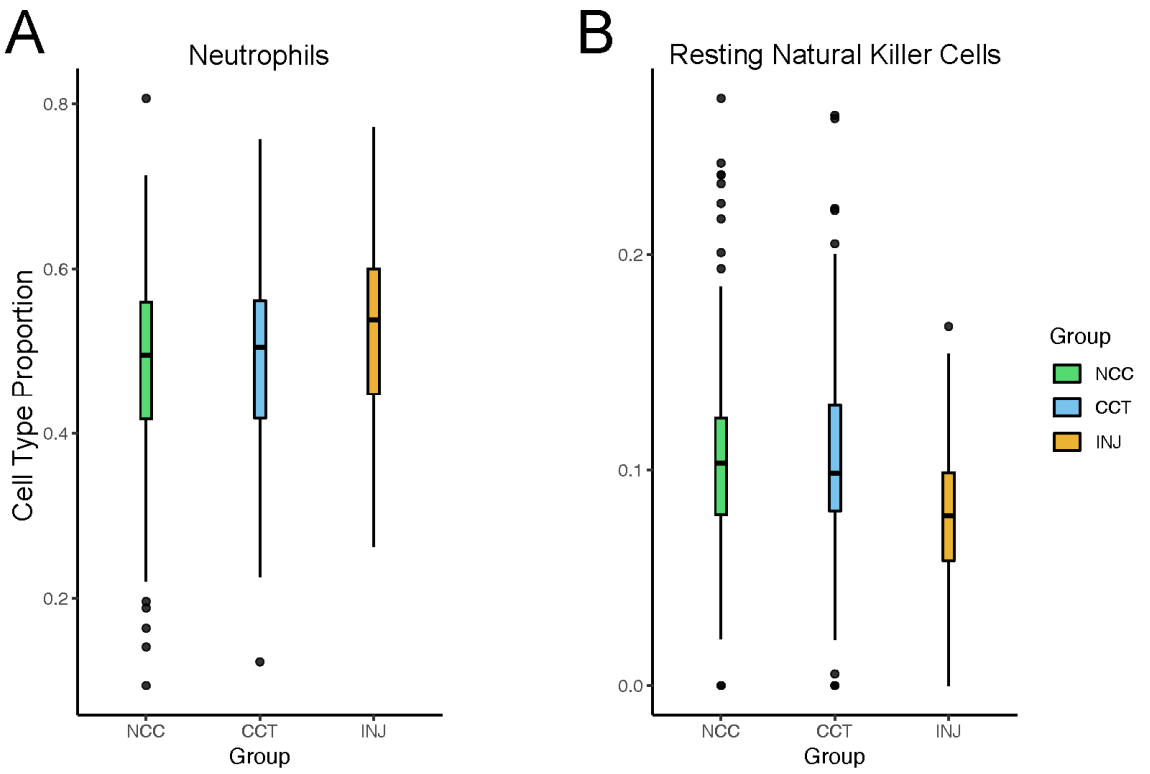

Figure 6 Cell type proportions at the PostInj timepoint. Boxplots of cell type proportions for neutrophils (A) and resting natural killer cells (B) estimated with *CIBERSORTx*. The box denotes the first-to-third quartile and the inner-line represents the mean. Whiskers extend to 1.5× the interquartile range and outliers are marked as points. The NCC group proportions are represented in green, CCT in blue and INJ in orange.

3.5 Discussion

To our knowledge, this study is the largest concussion transcriptome study to date. Our findings confirm results from numerous smaller studies and expand on the existing knowledgebase by showing trends in concussion-related pathways during recovery. We observed that maximal gene expression changes in peripheral blood cells were found immediately following concussion, and that these gene expression changes were consistent with a major immune signaling response. We identified compensatory changes in genes associated with calcium and energy metabolism which matched the pathophysiology of concussion. We also identified enhanced expression in genes that mediate immune signal transduction. GO and GSEA confirmed activation of immune signaling after injury. Furthermore, we observed that immune signaling processes were later suppressed, compared to contact controls, during recovery. Lastly, deconvolution analysis revealed the proportion of neutrophils was higher in injured participants compared to contact control athletes.

We did not observe changes in expression of genes coding for known blood biomarkers; Therefore, we speculate that protein and molecular biomarkers are present in blood not because related gene expression is enhanced, but due to changes in post-translational modifications or protein metabolism. Because of additional downstream factors, RNAseq of peripheral blood samples may not yield the same diagnostic targets as blood-based protein biomarker testing.

One difference between our studies and smaller studies is that others have noted differentially expressed genes sometimes days after injury, where in our study almost no differentially expressed genes were observed after 24 hours. We postulate that we did not observe highly significant differences in gene expression at later timepoints because of an increase in biological variability, due to our complex data set. While there may be low-level differences in gene expression later, as other studies and our GSEA analysis suggest, we suspect that the typical response to a concussion is a short-term surge in gene expression aimed at triggering cytokine and immune signaling processes that help to coordinate the immune response. Immune response then transitions from innate activation to adaptive activation and long-term recovery begins. This view is mirrored by conclusions from at least one other study in human traumatic brain injury [149].

One strength of our study is that using RNAseq technology, rather than microarray-based chips, allowed us to quantify significantly more genes than smaller studies previously done in concussion. One limitation of our study is that peripheral blood samples are not able to reflect the physiological environment of a concussion as well as brain tissue from the site of injury can. Despite this, blood is the most probable sample type for diagnostic testing in concussions and so reflects the diagnostic testing environment. Another limitation is that injury during a concussion is sustained by non-localized anatomy (other than merely the brain). We expect, however, that the variance introduced by the wide array of injuries in this study likely reduced the potential for identification of non-concussion related gene expression changes. Additionally, we did not investigate the influence of sex on gene expression following a concussion. Sex is known to be an

influential factor in concussion recovery; however, our goal was to identify wide-spread and common gene expression changes. Because individuals were first compared to their own paired baseline, the changes we identified are independent to sex. Finally, our study could benefit from additional samples and more complete time courses. Although longitudinal analysis is a primary goal of the CARE initiative, we feel that additional samples may be necessary to fully explore the longitudinal effects of concussion on gene expression. Increasing the sample size also increases the statistical power, which is required to separate low-level differences in gene expression and could improve findings at individual timepoints as well.

Chapter 4 Differentially Spliced Exons Predict Cancer Drug Sensitivity

4.1 Introduction

Alternative splicing is also known to contribute to the development and progression of cancer, and has been linked to every major signature of cancer transformation [150]. Furthermore, splicing variants can help cells evade cancer therapies and investigators have already started to explore splicing-focused therapeutic options [151–156]. Additionally, certain gene isoforms have been found to alter cancer drug response through altered kinase signaling [157,158]. Therefore, it is likely that alternatively-spliced isoforms play large roles in drug response, and that additional research in this area could have a major impact on development of targeted therapeutics and drug response modeling.

Precision medicine, or tailoring treatment strategies to the patient, is dependent on clinical and molecular profiling [159]. Currently, precision medicine primarily relies on limited genetic screening of well-characterized high-impact genes, such as HER2 and KRAS [160]. However, complex predictive models built with machine-learning techniques are expected to revolutionize precision medicine in the years to come [161,162]. Nevertheless, the use of complex predictive algorithms has yet to be widely accepted in clinical settings [161]. While early models lacked sufficient study sizes or could not be validated, a major concern of current models is the failure to account for the complexity of tumor transcriptomes [163]. Many predictive models have been trained solely on gene expression data or a combination of expression data and limited sequence variant information, such as single nucleotide polymorphisms (SNPs), copy number

variants (CNV) and small nucleotide insertions or deletions (indels) [15]. Previous studies, however, have concluded that algorithms capable of integrating knowledge from various experimental techniques need to be developed in order for predictive modeling to progress [15,16]. As such, a variety of experimental data, including mRNA-splicing data, must be considered in order to build more realistic and comprehensive models.

To date, few studies have incorporated splicing information into predictive modeling techniques. One such study produced the *SURVIV* pipeline, a system for discovering mRNA isoforms associated with patient survival [164]. These authors used exon-centric quantification and a binomial GLM with length normalization function on invasive ductal carcinoma data. They found that splicing information not only predicted patient survival, but it also consistently outperformed expression-based models [164]. Additionally, the authors found that combining clinical, expression, and splicing profiles produced the best performance [164]. In another study, isoform-centric biomarker expression and drug response in cancer cell lines was investigated using a linear model to select an isoform for each response-mediating gene that showed the strongest correlation with drug sensitivity [165]. A small number of these biomarkers were validated in breast cancer cell lines and found to be significantly associated with four anti-cancer therapeutics [165]. Together, these two studies established a connection between mRNA splicing and drug response; thus demonstrating the potential utility for splicing data in tumor biology. However, a drug-response classification model has not yet been established and the relationship between individual exons and cancer drug response is still largely unexplored.

4.2 Materials and Methods

4.2.1 RNAseq and Drug Response Datasets

975 RNA-seq files corresponding to pre-treatment cancer cell lines were downloaded from the Cancer Cell Line Encyclopedia (CCLE) database and matched to post-quality control area under the concentration-response curve (AUC) values for 860 cancer cell lines from the Cancer Therapeutic Response Portal (CTRP) v2 database using the cell line name [20,21]. While integrating data from two separate sources is not ideal, this approach was chosen because it provided the largest available overlap between RNA-seq and drug profiling data. Intersecting these data sets for cell lines profiled with doxorubicin yielded 755 cell lines with drug response and RNA-seq data. Cell lines were split into three groups using the tertiles of the AUC distribution. The low AUC group was labeled “sensitive” (N = 253), the high group was “resistant” (N = 258), and the middle group was omitted from analysis.

4.2.2 MISO Splicing Analysis

Splicing analysis for predictive modeling was done with the Mixture of Isoforms software (MISO) [91]. RNA-seq files belonging to sensitive and resistant groups were analyzed using exon-centric version 2 annotations for hg19 and the standard pipeline from the MISO documentation website, http://miso.readthedocs.io/en/fastmiso/. Data corresponding to 40,178 skipped-exon events were obtained.

4.2.3 Gene Expression Quantification and Differential Expression Analysis

Read counts for predictive modeling with expression data and for differential expression analysis were calculated with *featureCounts* [127]. A genomic feature was defined as any record with a valid gene id and was counted at the meta-feature level. RNA-seq files were processed for 57,095 genomic features that were annotated in the reference GRCH37v87 gene transfer format file (GTF) file downloaded from ftp.ensembl.org, using a minimum read length overlap of 2. Differential expression analysis was performed on count data from the training dataset using *edgeR* [97]. Only features with ≥ 10 reads in ≥ 35% of training cell lines were evaluated, leaving 22,201 features before differential expression and downstream filtering. Log_{10} of counts-per-million were used as feature values. The same annotation set of quantified genomic features as those used for predictive modeling (57,095) were again used for assessing differential expression of RBPs. In this case, filtering to include features with at least 10 reads in ≥ 20% of cell lines reduced the number to 28,110 before differential expression analysis. In *edgeR*, a negative binomial generalized log-linear model with quasi-likelihood F-test (*glmQLFit*) was used. Differentially expressed features with an FDR ≤ 0.05 and a log(fold-change) of ≥ 1.5 were considered significant, producing a final number of 2,943 differentially expressed gene features.

4.2.4 Predictive Modeling

Elastic net logistic regression, using the *glmnet* and *caret* packages in the R programming language, was used to fit all predictive models [166,167]. Following splicing and expression analysis, feature selection was performed to restrict the parameters of the

models. A splicing feature was defined as a skipped-exon event identified by MISO and was required to have PSI values with confidence intervals (CI) between 0.01 and 0.2 for a minimum of 35% of cell lines. This requirement reduced the number of potential splicing features from 40,178 to 15,007. We also filtered events having a PSI standard deviation < 0.14, based on the top 5% of the remaining skipped-exon events, which reduced the number of splicing features to 805. Any missing values were then imputed randomly from all samples with data for a particular event. Gene expression features were filtered by requiring a minimum of 10 reads in ≥ 35% cell lines. This lowered the number of potential features from 57,905 to 22,201. Cell lines were divided into 70% training and 30% testing sets. This produced a total of 354 training cell lines (177 sensitive, 177 resistant) and 157 testing cell lines (76 sensitive, 81 resistant). Individual and combined models were trained on the same training cell lines. Expression features were further restricted after training and testing set separation by conducting differential expression analysis on the training set and applying the cutoffs: FDR ≤ 0.05 and a log(fold change) of 1.74 (top 5%). Expression- and splicing-only models were then trained using their respective filtered feature sets, while the combined model was trained by merging the two filtered feature sets and allowing elastic net to choose freely between the whole.

To train the models, a 10-fold cross validation approach with grid search (to scan for the highest performing alpha and lambda values) was used. The models were then assessed with the testing cell line data. Sensitivity, specificity, accuracy and precision were calculated. The area under the receiver operating characteristic curve (ROC AUC), F1 score and p-value (corresponding to accuracy against the no-information rate) were also

produced. Lastly, when building models to assess the generalizability with the remaining 500 drugs in CTRP, all event types including skipped-exon, mutually-exclusive exons, retained intron, alternative 5' splice site and alternative 3' splice site were used for modeling.

4.3 Results

4.3.1 Dataset and Drug Selection

When we intersected cell lines from CCLE and CTRP, we observed the number of cell lines with both data types differed by drug. Per-drug area under the concentration-response curve (AUC) values from the CTRP were plotted. A higher AUC value, which is a surrogate for cell growth under increasing concentrations of a designated drug, corresponds to superior drug resistance. We chose doxorubicin to further investigate because it is a widely active chemotherapeutic used to treat a variety of malignancies and it affects cells through multiple mechanisms, including DNA damage by intercalation and inhibition of topoisomerase II [168,169]. Additionally, we reasoned that doxorubicin would be a good drug for proof-of-principal testing because the alternatively-spliced exons we identified would likely be relevant to a variety of cancer types, whereas spliced exons associated with targeted therapeutics might be relevant only to cancers containing specific genomic alterations. Furthermore, doxorubicin has been used in many drug modeling studies and therefore, our results would be expected to have greater context and build upon an existing body of knowledge.

4.3.2 Classification of Cell Lines Prior To Training

Following drug selection, we labeled cell lines according to their AUC values: cell lines at or below the 33rd percentile of the AUC distribution were considered doxorubicin sensitive and cell lines at or above the 66th percentile as resistant (**Figure 7**). This provided a total of 755 cell lines with intersected RNA-seq and doxorubicin response data; 253 were classified as sensitive and 258 as resistant.

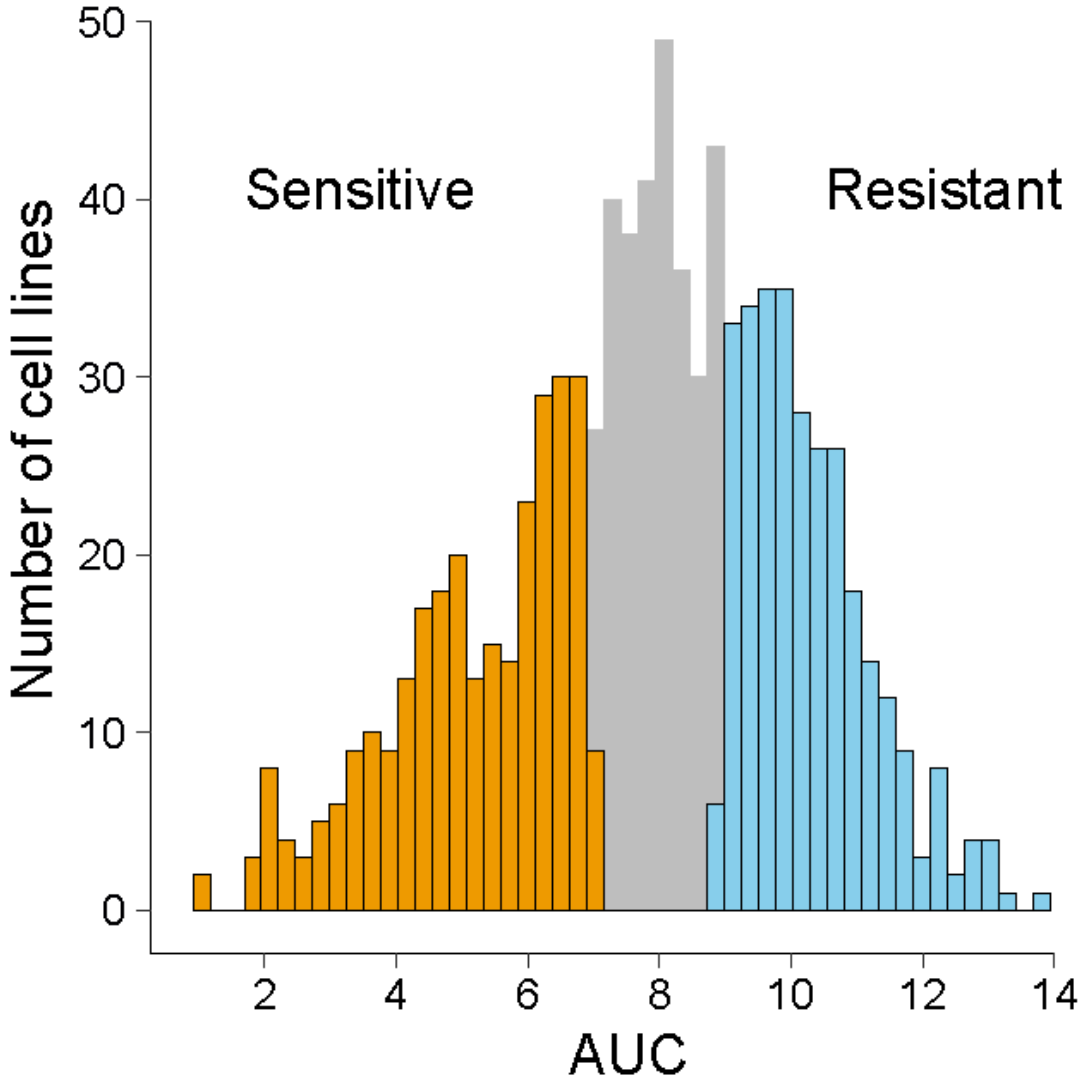

Figure 7 CTRP cell line response to doxorubicin. Distribution of the AUC, area under the concentration-response curve, values for doxorubicin in the CTRP cell lines. Lower and upper tertiles were labeled as sensitive (orange) or resistant (blue), respectively.

4.3.3 Splicing and Expression Data Individually Predict Drug Sensitivity Class

We postulated that alternative splicing profiles from untreated cancer cell lines would hold predictive power for doxorubicin drug-response. To test this hypothesis, we built a machine learning model with elastic net logistic regression and exon-centric splicing data. Skipped-exon event annotation, percent-spliced-in (PSI) calculation and uncertainty estimation were done with MISO [91]. For the splicing-based model, we required skipped-exon events (model features) to be present in a minimum of 35% of cell lines and to exhibit PSI values with CI between 0.2 and 0.01. We observed that PSI values with CI outside of this range tended to be either calculated on low read counts, or exhibited unrealistically precise distributions; these PSI values were filtered because small non-consequential changes in PSI would have been incorrectly considered highly significant. Skipped-exon events were then limited to only those with the highest (top 5%) PSI standard deviation, thereby targeting events with higher variance and selecting for greater model impact. From a total of 40,178 pre-filter skipped-exon events, 805 remained. Cell line data was then randomly split 70:30 into training (N = 354) and testing (N = 157) sets; each set consisted of approximately 50% sensitive and resistant cell lines. The predictive model was fit using elastic net logistic regression. The final splicing model contained a total of 42 non-zero weight events. Model performance was assessed on the testing data and performance metrics are provided in **Table 3**.

To assess whether splicing information would provide additional predictive power compared to an expression-based approach, we constructed an expression-only model. We first used *featureCounts* to quantify reads mapped to gene expression features [127].

To reduce the number of sparse genes, we filtered gene features with less than 10 reads in ≥ 35% of RNA-seq data. Using the same training set samples as the splicing-based model, we conducted differential expression analysis with *edgeR* to further reduce the number of features [97]. We retained genes meeting a Benjamini-Hochberg false discovery rate (FDR) threshold of 0.05 and ≥ 1.74 (top 5%) log(fold change) value [128]. Read counts were then transformed to $\log_{10}$ counts per million. Out of 57,905 pre-filter gene expression features, only 1103 remained. After running elastic net, we obtained an expression-only model comprised of 67 non-zero weight features. The performance of the expression-based approach was also strong (**Table 3**). In comparison with the splicing model, the sensitivity was lower (0.68 vs. 0.75), but specificity (0.96 vs. 0.88) and ROC AUC (0.90 vs. 0.85) were both higher. These metrics indicated that while splicing predicted more doxorubicin-sensitive cell lines correctly, it also predicted more false positives; on the other hand, expression-only modeling was more specific.

Model	Sens.	Spec.	Acc.	Prec.	AUC	F1	P-value
Splicing (S)	**0.75**	0.88	0.82	0.85	0.85	0.80	5.3E–15
Expression (E)	0.68	**0.96**	0.83	**0.95**	0.90	0.79	2.8 E–16
S + E	0.71	0.95	**0.83**	0.93	**0.92**	**0.81**	**6.2 E–17**

Table 3 Elastic net performance metrics. The strongest value for each column is in bold; Sens., sensitivity; Spec., specificity; Acc., accuracy; Prec., precision; AUC, area under the receiver operating characteristic curve.

4.3.4 An Integrated Modeling Approach Outperforms Stand-alone Models

Based on our findings that splicing- and expression-based models showed strengths in sensitivity and specificity, respectively, we asked whether integrating the information from both models would lead to increased model performance. An integrated model was fit by merging the 805 events obtained after applying the splicing filter with the 1103 gene expression features remaining after applying the differential expression filter. From this combined feature set, elastic net selected 95 splicing and 216 gene expression features. ROC plots for all 3 models are in **Figure 8**. The integrated model showed the highest accuracy and ROC AUC (**Table 3**). From this outcome we concluded that splicing information enhanced the expression-based model and that splicing and expression data contributed improvements to sensitivity and specificity, respectively, to build a more balanced model. Bootstrapping the model building process revealed that although the combined model consistently showed a slight increase in specificity, the overall performance of the combined and expression-based models was largely the same (**Figure 9**).

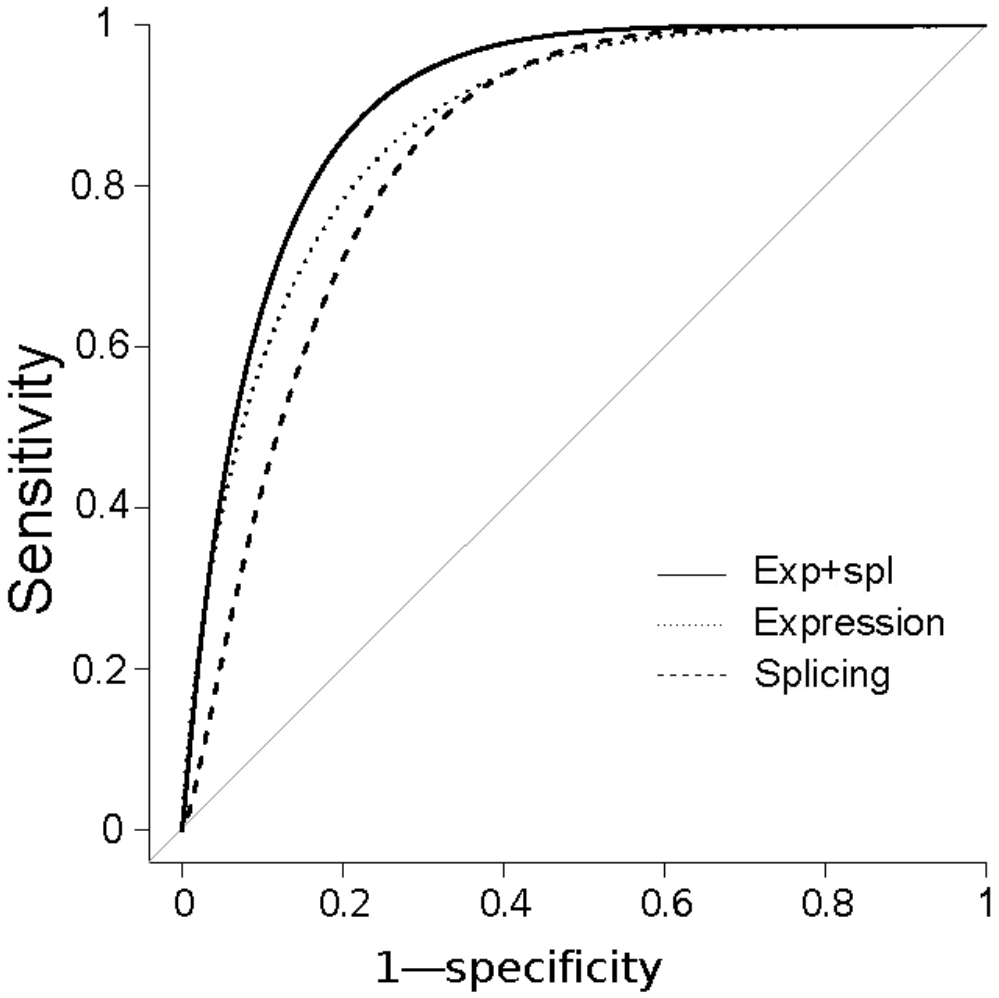

Figure 8 Comparison of model prediction of cell line response to doxorubicin. ROC, receiver operating characteristic, curve for prediction of cell line response to doxorubicin on the testing data set. Expression-only (dotted line), splicing-only (dashed line) and combined expression and splicing (solid line) models are shown.

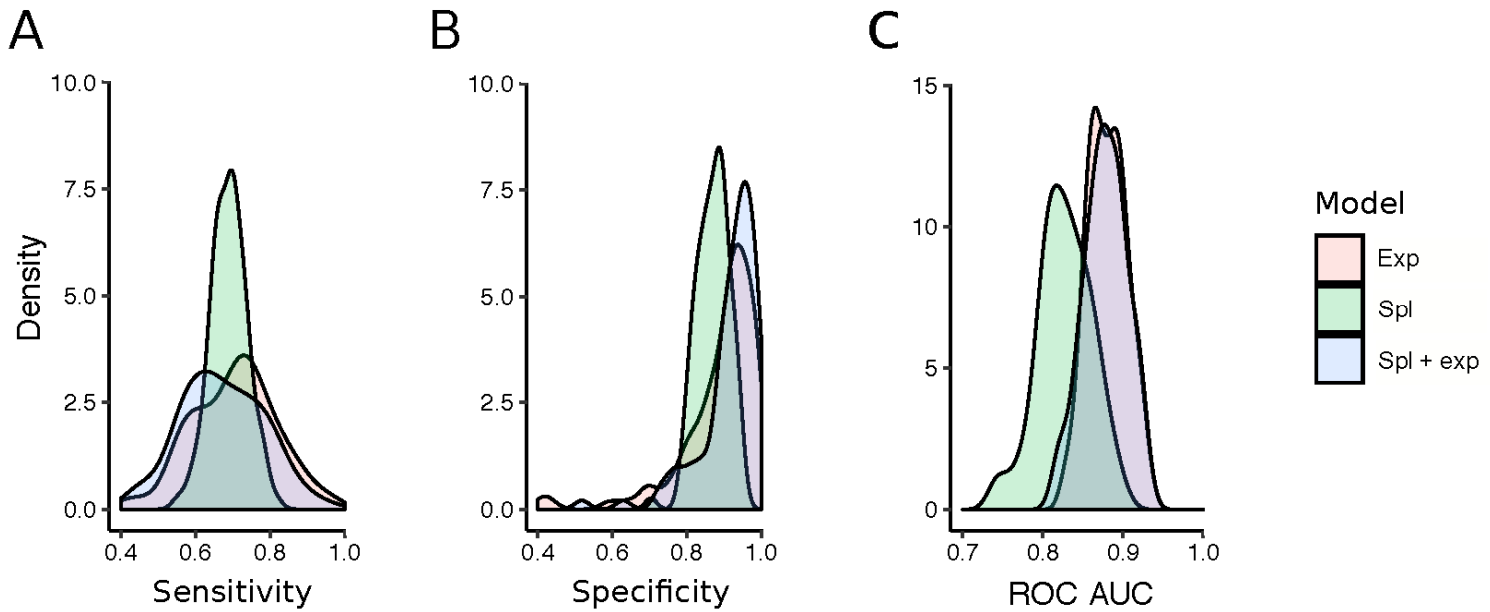

Figure 9 Reproducibility of predictive modeling. Sensitivity (A), Specificity (B), and (C) ROC AUC, receiver operating characteristic area under the curve, for 100 bootstrapped splicing models on doxorubicin.

4.3.5 Splicing Contributes Unique Predictive Features to Modeling

We next asked whether splicing contributed unique information to the final model, or if the skipped-exons selected by elastic net were also reflected by the gene expression features. We found that skipped-exon features in the splicing-only model were not located in genes in the expression-only model. Similarly, no overlapping expression and skipped-exon features were observed in the combined model. These findings indicate that the information contributed by splicing data to our models was unique.

4.3.6 Splicing Features Have Predictive Value in Many Drugs

We further applied our modeling approach to the other 500 drugs in CTRP and observed strong performance for the vast majority (**Figure 10**); This suggested that splicing could be important for predicting drug resistance across many compounds. To identify compounds where splicing was better at predicting response class, we then directly compared the ROC AUC from the splicing-only models to the expression-only models (**Figure 11A**). We found that splicing-only models for four compounds outperformed their paired expression-only model by >0.10; These compounds were BRD-K44224150, darinaparsin, tigecycline, and VAF-347. An ROC AUC advantage of ≥ 0.05 was observed in a total of 51 splicing-only models and 179 expression-only models, suggesting that generally expression-only models performed better than splicing-only models. This finding is also supported by a negative shift in the difference-distribution and mean-difference between splicing-only and expression-only ROC AUC values (**Figure 11B**).

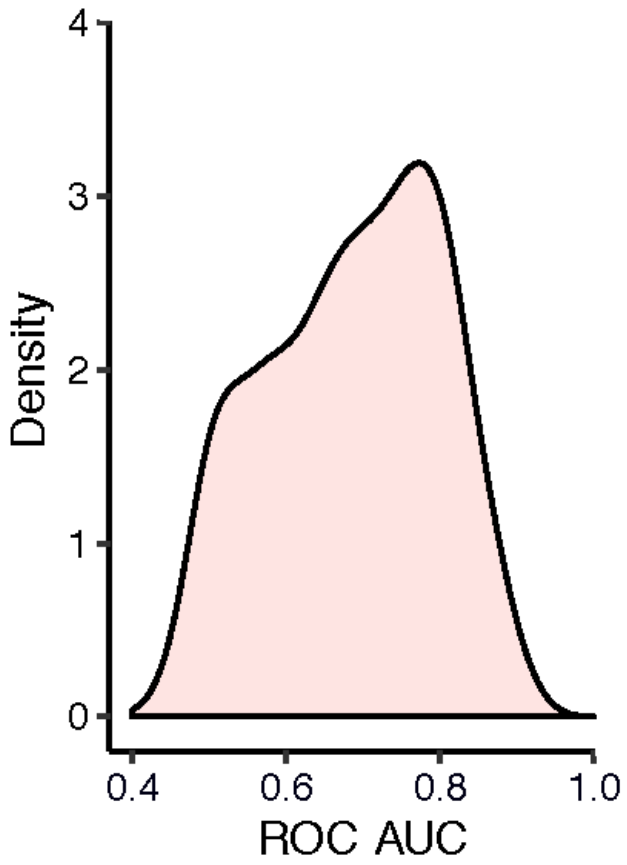

Figure 10 Generalized CTRP model performance. ROC AUC, receiver operating characteristic area under the curve, for splicing models on all 501 drugs in CTRP.

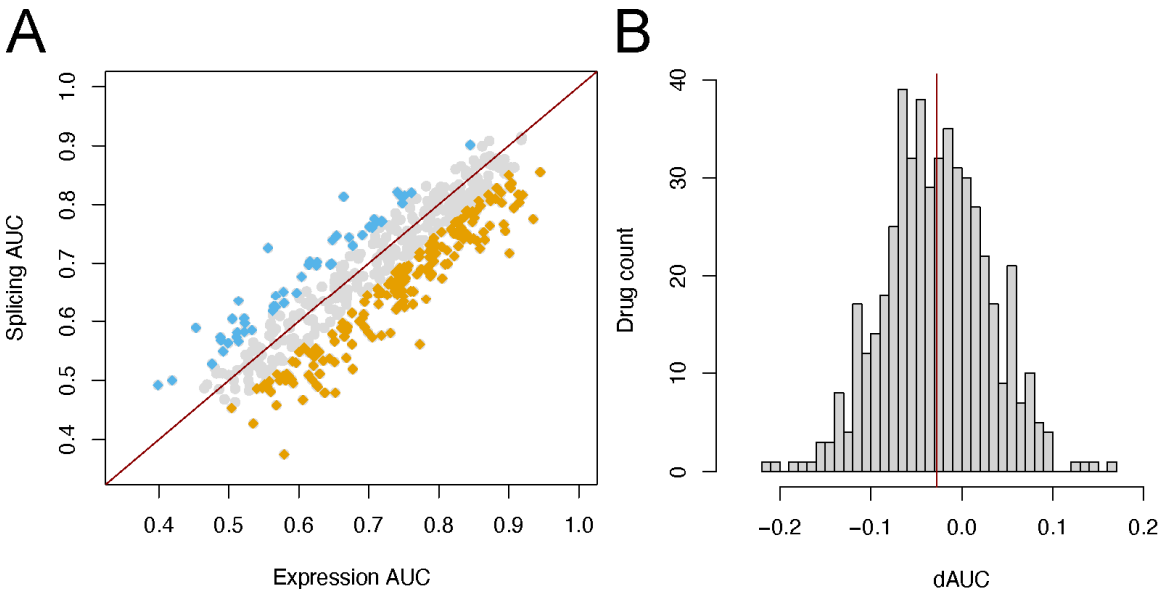

Figure 11 Direct comparison of splicing and expression model performance for all drugs. A) ROC AUC for splicing-only and expression-only models in all 501 drugs plotted against each other. The red line has an intercept of zero and slope of 1, dividing drugs between those where the splicing model performed better (top) and those where the expression model performed better (bottom). Drugs in blue and orange exhibited a ≥ 0.05 increase in AUC for the splicing or expression model, respectively. B) Histogram of the difference between splicing and expression ROC AUC for all 501 drugs. The distribution mean is indicated by the red line.

4.4 Discussion

The major conclusions of chapter 4 are that skipped-exon splicing data independently predicts drug response and, when integrated with gene expression data, can increase the power of predictive drug response algorithms. These conclusions are supported by the following experimental evidence. First, we demonstrated strong performance of the splicing-only elastic net GLM model and determined that the most balanced model was obtained by combining splicing and expression data. Second, we showed that splicing and expression models had no genes in common, which indicates that each data type contributed unique information. Third, we found that splicing holds strong predictive power in a large number of anti-cancer compounds.

After our analysis, we questioned whether our model could have selected predictive features based on differences in underlying cell-type proportions among sensitive and resistant groups. We considered this possibility because doxorubicin is known to be more effective against highly-proliferative cell types and it was possible that allocating cell lines to groups by response could have assorted proliferative cell lines unevenly [170]. We further investigate the assortment of cell types in chapter 5.

We did not choose to investigate the details or mechanisms behind the predictive splicing features because, while machine learning techniques excel at identifying relevant predictive features, the final selection of features included in the model can be arbitrary. Elastic net, to simplify the final model and prevent overfitting, restricts orthologous

predictors; Therefore, features extracted from an elastic net model may not be the best biologically relevant targets.

To accurately assess the contribution of differential splicing to predictive drug modeling, we sought to identify a comprehensive and well-characterized data set with drug response measurements and paired RNA-seq data. Although the widespread availability of high-throughput data sets offered a number of options for computational modeling, the majority of large-scale studies were done before RNA-seq became the predominant expression quantification method and most of the pharmacological profiling experiments were paired with array-based expression data, making splicing analysis impossible. We searched for data sets with large numbers of samples to increase the power of our machine learning based approach and to avoid overfitting. We also targeted diverse data sets to allow for investigation of predictive features with broad applicability, corresponding to multiple drugs and cell types. These criteria led us to integrate two large independent data sets, rather than use a single resource with limited transcriptomic or pharmacological data.

While some investigators have challenged the integration of drug response datasets, integration of these resources by others has shown reasonable consistency[171,172]. Additionally, other investigators have argued that isolated testing of individual cancer cell lines is an incomplete representation of tumors and that databases containing large collections of cells better represent the heterogeneity and tissue-level characteristics of cancer [173]. Because our goal was to specifically target global splicing patterns, we

sought to use large data sets to reduce the impact of individual differences across databases. Therefore, we feel that our approach accurately reflected the transcriptomic and drug response measures of various cancer types, that the number and composition of cell lines in it reduced the possible influence of lineage inconsistency, and that our dataset is a reliable source of information for investigating global trends in transcript splicing or expression.

When performing machine learning, we elected to build a classification model rather than a continuous model as the CCLE and Genomics of Drug Sensitivity in Cancer consortia (GDSC) recommended dividing cell lines into sensitive and resistant groups when analyzing drug response data across datasets [174]. This recommendation was based on the observation that using all cell lines in a database tended to introduce noise due to the increased variance in drug response from cell lines that did not have influential genetic differences [174]. We analyzed performance consistency by bootstrapping the model building procedure and found combined and expression-based models were almost equivalent. While we did not find splicing-based modeling to outperform expression-based modeling as previous researchers have [164,165], our approach differed from these earlier models as it was designed to determine the importance of alternative splicing in doxorubicin drug response using a minimalistic procedure rather than generating the best possible classifier. Nevertheless, while our work provides evidence that adding splicing information to expression-based models in a more controlled manner produces a better classifier, there remains room for improvement in the model building process.

Chapter 5 Quasi-binomial Generalized Linear Modeling: A Method for Differential Splicing Analysis

5.1 Introduction

Although long-read isoform sequencing technologies exist, they are often prohibitively expensive for large-scale studies. As a consequence, short-read data is commonly used to infer isoform-specific information; the drawback being that the true identities of mRNA isoforms remain unknown. This uncertainty must be accounted for in quantitative techniques [91]. Currently, there are two main approaches to quantify isoform outcomes in short-read RNA-sequencing data: isoform-centric and exon-centric quantification [175]. Isoform-centric techniques measure the expression of whole isoforms by integrating read data across multiple exons, whereas exon-centric techniques measure relative expression of individual exons. While both isoform- and exon-centric techniques are susceptible to the limitations of short-read sequencing, gene complexity and the heavy reliance on mathematical modeling to address combinatorial possibilities across exons often make isoform-centric approaches less attractive [176].

Following predictive modeling in chapter 4, we decided to comprehensively investigate alternatively spliced events related to doxorubicin sensitivity. To capture a more complete set of events, that were not subject to arbitrary selection by machine learning feature reduction techniques, we set out to conduct differential splicing analysis between sensitive and resistant cell line groups. As discussed previously (chapter 2 section 5), *rMATS* is the current gold standard for differential splicing analysis of short-read RNAseq data [92]. The statistical approach *rMATS* applies is excellent, however the

process is very slow, resource intensive and better suited to smaller sample sizes. In order to process the integrated CCLE and CTRP dataset we were required to develop a fast, scalable framework for differential splicing analysis.

5.2 Materials and Methods

5.2.1 Dataset and Cell Line Classification

The integrated CCLE and CTRP dataset from chapter 4 was once again used for work in chapter 5 [20,21]. Alignment and cell line classification was performed as in chapter 4 sections 4.2.1. The same procedure was followed for all drugs in the dataset. The total number of cell lines tested, and as such the number of sensitive and resistant cell lines, differed per compound.

5.2.2 Processing and Quantification of Spliced Exons

Identification of candidate skipped-exon events was performed with code from *rMATS* that had been modified for speed and to catalogue only events found in a reference GTF [92]. A total of 38,108 skipped-exon events were extracted from isoforms annotated in the GRCH37v87 GTF file downloaded from ftp.ensembl.org. Uniquely mapped and properly paired junction reads with a minimum exon overlap of 1 bp supporting the inclusion or exclusion of skipped exons were counted for each skipped-exon event. Events were filtered after counting, retaining only those with at least 1 inclusion and 1 exclusion read in 35% of classified cell lines. A total of 18,409 events passed the filter.

5.2.3 Differential Splicing Analysis by QBGLM

Splicing analysis by quasi-binomial generalized linear model (QBGLM) was done using raw read counts. A QBGLM was fit using the *glm* package in R [177]. The inclusion read percentage for a given event was modeled as the probability of success. Cell line label (sensitive or resistant) was set as the dependent variable.

Results were filtered for significance by requiring a Benjamini-Hochberg FDR-adjusted p-value ≤ 0.01 on the group weight (Beta1) [128]. A total of 4,309 events passed the filter. Events were further separated for relevance using the difference (Δ) in mean inclusion-to-total read counts (inclusion / inclusion + exclusion) in each group. A minimum 0.1 difference in mean inclusion-to-total read counts between sensitive and resistant groups was required to maximize biological relevance; only 277 events met this threshold.

5.2.4 GO Enrichment Analysis

Significant skipped-exon events identified from QBGLM were annotated for gene symbol by genomic position of the skipped exon using the Bioconductor *biomaRt* package [130,131]. Gene symbols for the entire set of significant events were then analyzed with R and the *clusterProfiler* package [129]. Results from biological process enrichment were then exported and assessed for relevance.

5.2.5 Motif Enrichment

Significant skipped-exon events were analyzed for enrichment of RNA binding protein motifs in three stages: motif matches were counted for significant events in a region of interest, the total count for the set of significant events was compared to a background of randomly drawn events and significantly enriched motifs found were then filtered and sorted based on their associated splicing outcome. Seven regions surrounding each exon of interest were extracted from hg19 (GRCh37). These regions were: 150 bp maximum or full length of the 5' upstream exon, 300 bp of its 3' flanking intron, 300 bp in the 5' upstream intron flanking the skipped exon, the entire length of the skipped exon, 300 bp in the 3' downstream flanking intron, 300 bp in the 5' intron flanking the 3' downstream exon and 150 bp maximum or full length of the 3' downstream exon. Sequences for each region were scanned using the Find Individual Motif Occurrence (FIMO) tool and the CISBP-RNAv0.6 RNA binding motif database [178,179]. Using a p-value threshold for motif matches of 6.7e-4, as compared to the default 1e-4, was necessary to find small splicing factor motifs in short extracted sequence lengths. Counts across significant events for a given motif were then compared to the genomic background in context by bootstrapping the same number of skipped-exon events (without replacement) from all annotated events in the genome, repeating the procedure 10,000 times. P-values for significant event motif counts were then calculated using this random distribution and FDR-adjusted using the Benjamini-Hochberg procedure [128]. This was referred to as the enrichment p-value. Fisher's exact test was then used on enriched motifs to identify those associated with preferential increased or decreased exon inclusion. P-values from

Fisher's exact test were referred to as inclusion p-value. In both enrichment and preferential inclusion analysis a minimum p-value of 0.05 was required.

5.3 Results

5.3.1 QBGLM Identifies Differentially Spliced Events

Genome-wide annotation for skipped-exons resulted in a total of 38,108 events that were then filtered to retain only those with reads supporting inclusion and exclusion for a minimum of 35% of cell lines in sensitive and resistant groups. This filter significantly decreased the event space and left 18,409 events for analysis. GLM with quasi-binomial family was then performed in R [177]. In the quasi-binomial distribution, the dispersion parameter provides for fitting of increased variance; this property is especially useful for biological data, where variability between samples is expected. Additionally, fitting variance by QBGLM helped account for the uncertainty introduced when using short-read data in splicing analysis and situations where a low number of reads inaccurately represents the probability of inclusion in some samples. Our procedure was also unique for splicing-data normalization in that no consideration was made for exon or read lengths. As such, QBGLM modeled uncertainty without the assumption that there was an equal probability of reads aligning to every position in the event.

Wald p-values, corresponding to the weight on the class of the cell line, were FDR-adjusted using the Benjamini-Hochberg procedure and filtered for significance ≤ 0.01 [128]. Events were again filtered after QBGLM by requiring a difference in mean inclusion-to-total read counts of 0.1 between sensitive and resistant groups. This filter

reduced false-positive identifications by selecting events that were more likely to exert meaningful biological consequences. In total, 277 significant alternatively-spliced events were identified; 180 with higher and 97 with lower exon frequency in resistant cells. A volcano plot of the results and examples of raw data for two significant events are presented in **Figure 12**.

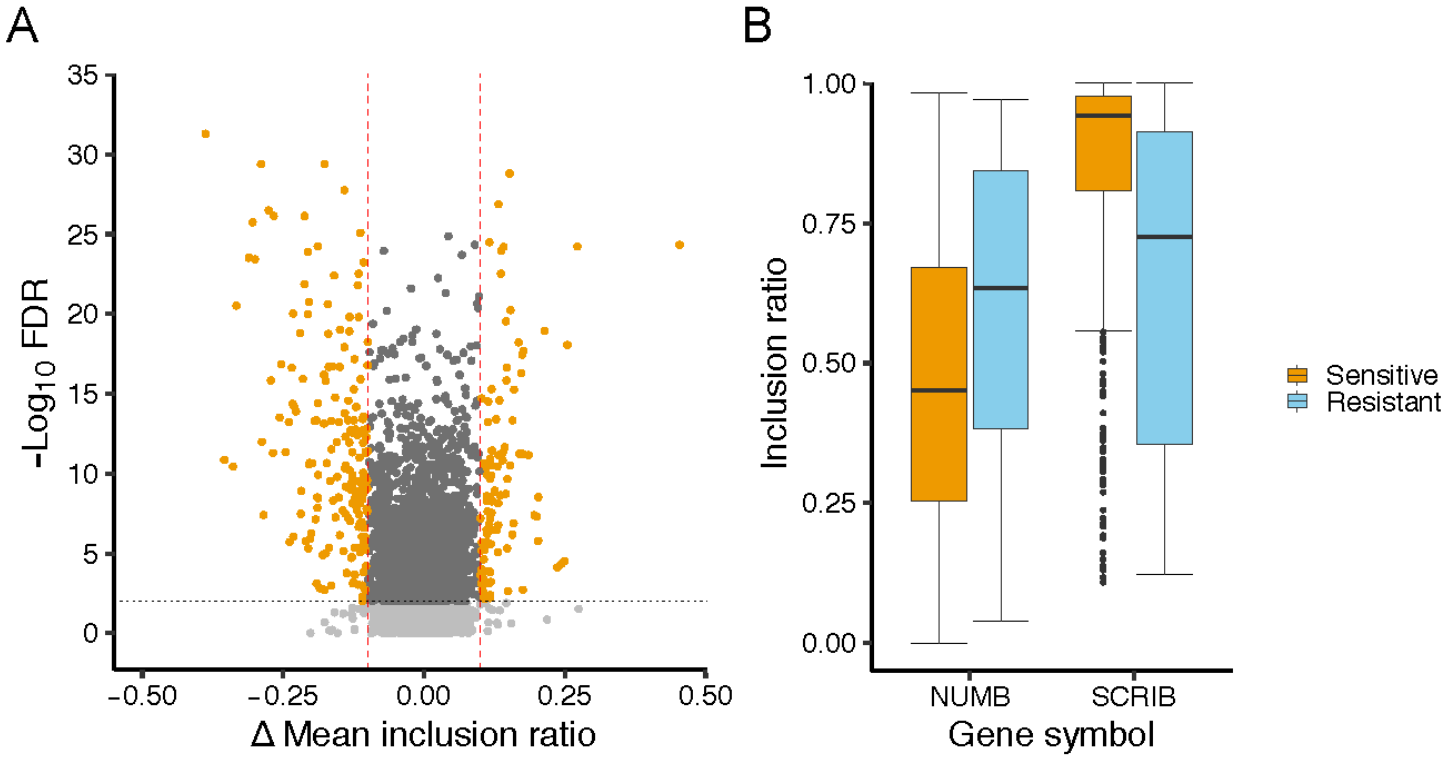

Figure 12 Differentially spliced events analyzed by QBGLM. A. Volcano plot of events analyzed by QBGLM. The horizontal dotted line marks the 0.01 FDR cutoff for significance. Vertical dashed lines mark the minimum 0.1 mean inclusion ratio difference between sensitive and resistant groups. B. Boxplots for inclusion ratios show overall change between sensitive and resistant groups for two genes with significant spliced events. The box denotes the first-to-third quartile and the inner-line represents the mean. Whiskers extend to 1.5× the interquartile range and outliers are marked as points.

5.3.2 Enrichment Analysis Reveals Connection to Epithelial-Mesenchymal Transition

Gene ontology (GO) term enrichment was performed on gene symbols from significant alternatively-spliced events to assess their relevance. Over-representation analysis with *clusterProfiler* revealed enrichment for several biological processes including cell junction organization (FDR 1.3e-04), cytoskeleton organization (FDR 1.7e-05), and positive regulation of GTPase activity (FDR 3.2e-05) (**Figure 13**) [129]. Alterations in these processes have been previously implicated in uncontrolled cellular proliferation, epithelial-mesenchymal transition (EMT), and drug resistance [180–184]. Noteworthy genes affected by splicing alterations included *SCRIB*, *ADAM15*, *MACF1*, *NUMB*, *VEGFA*, and *FOXM1*. While the majority of splicing consequences were in-frame alternatively included or excluded exons with no known significance, an exon identified in *NUMB* (exon 11, chr14:73745989-73746132) contained an alternative translational start site and another in *SCRIB* (exon 16, chr8:144889722-144889784) included a portion of a Protein Kinase C (PKC) phosphorylation site. *NUMB* is a key protein in cell fate determination and increased expression has been found to inhibit propagation of chronic myelogenous leukemia cells [185,186]. Additionally, *NUMB* mRNA processing is regulated by a variety of splicing factors, including *RBM6*, and alternative *NUMB* isoforms are consistently found in cancer [187,188]. *SCRIB* exon 16 has been reported to be associated with mis-regulation of EMT in specific cell types [45].

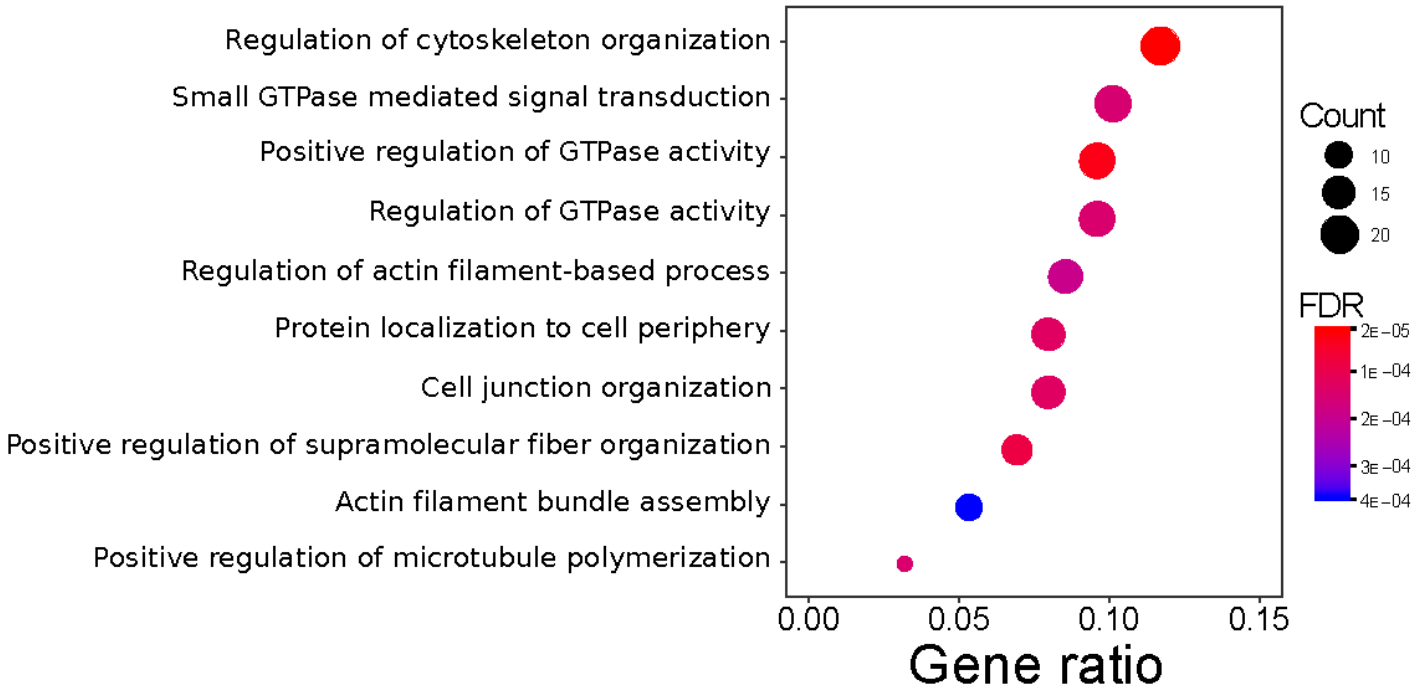

Figure 13 Enrichment of biological processes identified in differentially spliced events. GO biological processes identified with overrepresentation analysis were sorted by gene count ratio from top to bottom, with the highest ratio (found to total significant) of genes for a specific process on top. Point diameters are scaled by total number of genes in that process and warmer colors indicate stronger significance.

5.3.3 RBP Motif Enrichment and Regulatory Splicing Factors

To elucidate a regulatory mechanism for the splicing differences between response groups, we searched for RBP motifs corresponding to potential splicing factors. Motif analysis was conducted on seven sequence regions for each skipped-exon. These regions consisted of the entire skipped-exon sequence, 300-bp from the 5'- and 3'-ends of both flanking introns, and 150-bp in the upstream and downstream exons (**Figure 14**). Sequences from these regions were extracted from the hg19 reference genome and scanned for motifs using FIMO [178]. To determine enrichment for identified motifs, all annotated skipped-exons across the genome were scanned and null distributions of counts for each motif were made from bootstrapped events. RBP motifs identified were *SNRPA*, *PPRC1*, *RBM6*, *PCBP3*, *RBFOX1*, *EIF2S1* and *ELAVL1*. Locations and enrichment p-values of the identified RBP motifs are in **Figure 14** and **Table 4**. Fisher's exact test was used to determine association with higher or lower exon frequency. Splicing outcome, Fisher's p-values and descriptions of the identified RBPs are in **Table 4**.

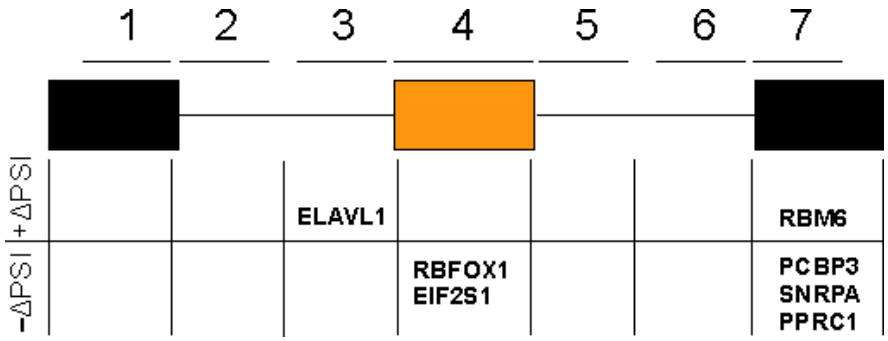

Figure 14 RNA-binding protein motifs identified in differentially spliced events. Significantly enriched RBP, RNA-binding protein, motifs by skipped-exon event region with respect to resistant cell lines. The above schematic shows two constitutive exons (black boxes), one skipped-exon (orange box) and two introns (connecting lines) as observed in skipped-exon splicing. Regions of interest are shown as horizontal lines numbered 1 to 7. These regions consisted of: (1) up to 150bp of the upstream exon; (2) 5′ 300bp of the upstream intron; (3) 3′ 300bp of the upstream intron; (4) the entire length of the skipped exon; (5) 5′ 300bp of the downstream intron; (6) 3′ 300bp of the downstream intron; and (7) up to 150bp of the 3′ downstream exon. +Δ and -Δ PSI indicate a higher and lower exon frequency in resistant cells, respectively.

RBP	Position	Enrichment P-value	Exon inclusion	Inclusion P-value	Description
RBFOX1	Skipped exon	8.7E–08	-	0.015	RNA-binding protein Fox-1 homolog 1 (RBFOX1). This protein and its family members (RBFOX2 & 3) bind to (U)GCAUG stretches. They are generally found to enhance splicing when bound downstream and suppress splicing when bound upstream [27,189].
EIF2S1	Skipped exon	1.6 E–07	-	0.011	EIF2S1 (or EIF2alpha) is one of three key members of the Eukaryotic Translation Initiation Factor 2 complex and is responsible for delivering Met-tRNA for initiation of translation [189].
RBM6	3' exon	1.2 E–04	+	0.021	RBM6 is an RNA binding protein first identified by cloning a tumor suppressor locus and has been linked to lung as well as other cancers [190].
PPRC1	3' exon	1.6 E–04	-	0.021	PPRC1 (or PGC-1) is a coactivational transcription factor commonly associated with metabolic stress and little is known about its potential role in splicing, however an important paralog of this gene (PGC- 1alpha) has been connected to altered splicing of VEGF [189,191].
ELAVL1	5' intron	1.9 E–03	+	0.005	ELAVL-like RNA Binding Protein 1 (ELAVL1) family members traditionally bind AU-rich elements in 3' untranslated regions of mRNA [189,192].
PCBP3	3' exon	0.015	-	0.002	PCBP3 is a member of the poly(rC)-binding protein family and is paralogous to PCBP1,2 & 4. Members of this family have strong motif

					homology and share a wide variety of functions, however PCBP3 lacks the nuclear localization signals that other members have [189,193].
SNRPA	3' exon	0.026	-	0.015	Small Nuclear Ribonucleoprotein Polypeptide A (SNRPA) is an essential component of the U1 splicing complex and is required for recognition of the pre-mRNA 5' end. The U1 complex binds to the 5' splice site of an exon-intron boundary [192].

Table 4 Enriched RBPs identified by motifs in significant events from QBGLM. The enrichment p-value is the FDR-adjusted p-value against randomly bootstrapped events from the genome. Exon inclusion is with respect to resistant cells; *i.e.*, "+" is higher exon frequency in resistant cell lines. The inclusion p-value was calculated using Fisher's exact test.

5.3.4 Enriched RBP Family Members Are Differentially Expressed

Finally, we asked whether any of these enriched RBPs were differentially expressed between sensitive and resistant cell lines. Differential expression analysis was conducted using *edgeR* on *featureCount* data from the two groups [97,127]. Significantly different expression patterns were observed for RBFOX and ELAVL family proteins (**Figure 15**). This finding was particularly interesting as RBFOX and ELAVL family members have previously been linked to EMT and other cancer-related processes [45,194,195]. Notably, Blencowe et al. previously found that *PPRC1* increased splicing of *NUMB* exon 5 in CGR8 mouse embryonic stem cells compared to differentiated N2A neuroblastoma cells [196]. In our work, *NUMB* exon 10 was differentially spliced. However, we did not see differential expression of *PPRC1* as it was filtered before *edgeR* analysis due to low read count. In contrast, RBFOX, ELAVL and *PPRC1* RBPs were not selected as predictive features in the expression-based predictive model. Based on these findings, we conclude that the additional biological information gained from splicing analysis could not be found using expression based analysis alone.

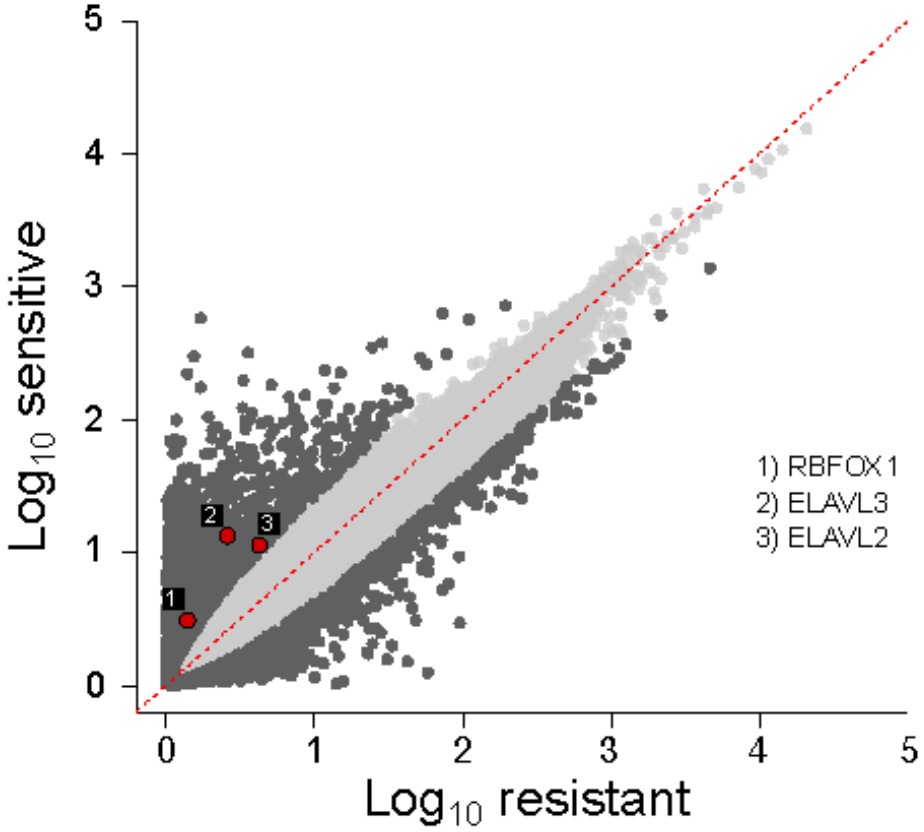

Figure 15 Differential expression of RBPs between sensitive and resistant cell lines. Mean expression of genes in sensitive and resistant groups using $\log_{10}$(read counts). Three differentially expressed RBPs belonging to the RBFOX or ELAVL families are numbered and shown in red.

5.4 Discussion

In this chapter we show that a quasi-binomial generalized model can accurately identify differentially spliced events between two groups of samples. Additionally, we demonstrated that an exon-centric approach positively impacts downstream analysis by identifying cis-acting RBP regulatory motifs and allowing researchers to find associations between regulatory elements of differentially spliced exons with essential biological processes. When employing RBP-motif enrichment, we identified several candidate splicing factors, including RBFOX and ELAVL family members, which were differentially expressed between drug-response groups. Moreover, we identified signatures of EMT, which affect cellular plasticity and stemness in tumor sub-populations and are thought to contribute to mechanisms behind cancer drug resistance [46,197–200]. Furthermore, our results indicate that splicing information provides new biological insights.

Following our analysis, we assessed the cell line origin of the classified data set to investigate if differences in the proportions of cell lineages could help explain the enriched biological processes we observed (**Figure 16**). The distribution of cell lineages, specifically the proportion of hematopoietic & lymphoid cells, differed greatly across sensitive and resistant groups: hematopoietic & lymphoid cells made up 44% of sensitive compared to only 1% of resistant cell lines. Hematopoietic cell types exhibit enhanced cytotoxicity to doxorubicin treatment, a consequence of treating highly proliferative cells with a topoisomerase-inhibitor [170]. These cells, being more stem-like in nature compared to solid tumor tissue, are also expected to display signatures of EMT as

stemness and EMT are related [201]. Our overrepresentation analysis of differentially spliced events resulted in a number of biological processes with relationships to EMT, proliferation, and drug resistance. Furthermore, we identified an exon in *SCRIB* previously described by Shapiro et al. to be alternatively spliced and associated with an EMT signature [45].

To determine the potential influence that cell type distribution might have had on the QBGLM results, we performed QBGLM in a tissue-specific manner on hematopoietic & lymphoid and lung cell types. We then overlapped the differentially spliced events from all tissues and the tissue-specific analyses (**Figure 17**), which showed that hematopoietic & lymphoid and lung cell types had many events in common. The vast majority of events found to overlap between the tissue-specific and all tissue-type analyses were present in both hematopoietic & lymphoid and lung cell types. Additionally, we did not observe an imbalance in the number of events from the overlaps between all tissue and tissue-specific analyses. These findings support the conclusion that QBGLM also identified events from other tissue types besides hematopoietic & lymphoid, and that many events found in hematopoietic & lymphoid cells are recapitulated by other cell types.

In conclusion, differential splicing analysis with QBGLM can be achieved in minutes. Even for groups containing hundreds of samples, analysis time is negligible if inclusion and exclusion reads are counted beforehand as part of a standard pipeline. While our analysis works well for large groups of samples, it struggles with smaller sets; however, we expect the model's ability to handle large groups of samples to be a key strength as

the volume of sequencing data and the number of samples included in studies continues to rise. In the next chapter we discuss a strategy that implements quasi-binomial differential splicing analysis in a tissue-specific manor to find splicing trends across drugs based on cell line drug sensitivity.

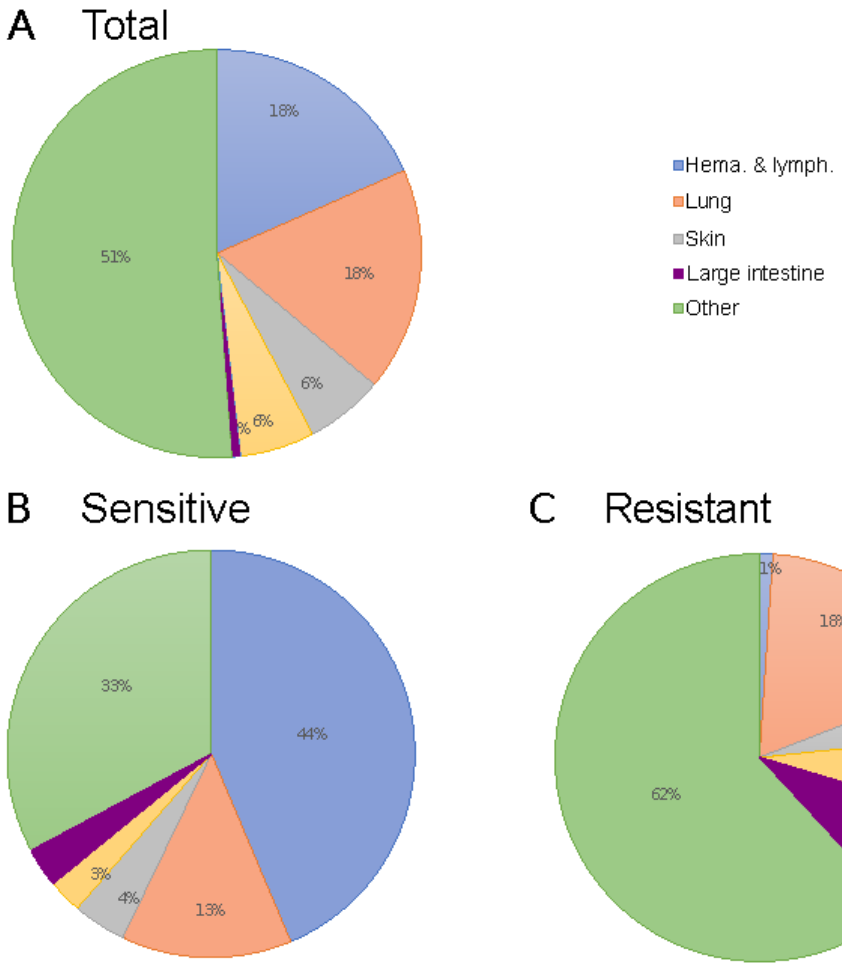

Figure 16 Tumor cell lineage for cancer cell lines tested with doxorubicin. A. The overall chart represents cell lineage distribution for all cell lines in the paired (CCLE and CTRP) dataset. B. Distribution for cell lines sensitive to doxorubicin. C. Distribution for cell lines resistant to doxorubicin.

Figure 17 Overlapping events between all tissue and cell-type specific differentially spliced events. Venn diagram of common differentially spliced events in all, hematopoietic & lymphoid and lung tissue types.

Chapter 6 Tissue-specific Network Analysis of Splicing Data

6.1 Introduction

In chapters 4 and 5 we identified differentially spliced exons in pre-treatment transcriptional profiles from cancer cell lines, their relationship with drug response and the regulatory elements that play a role in their splicing. We found that alternatively spliced skipped-exons were highly predictive of doxorubicin drug response and extending the same modeling approach to other drugs yielded similar results. We then hypothesized that drugs from the same class or with similar activities would share predictive splicing features.

As discussed in chapter 2 section 1, splicing has been linked to cancer drug resistance [151,152,202]. Certain spliced isoforms can manipulate kinase signaling and alter cellular drug response [151,152]. Despite this, few studies have explored connections between drug response and splicing. Additionally, it is well known that many drugs exploit similar targets or pathways as development of structurally homologous compounds is cheaper and faster than development of novel therapeutics. Yet to our knowledge, no one has investigated commonality between splicing signatures modulating cellular response to various compounds.

Here, we expand our work to incorporate all classes of alternatively spliced events and construct tissue-specific drug networks utilizing predictive splicing features. We describe the drug network characteristics and explore exons connecting individual drug modules.

6.2 Materials and Methods

6.2.1 Dataset and Splicing Quantification

The integrated CCLE and CTRP dataset from chapter 4 was once again used for work in chapter 6 [20,21]. Alignment and cell line classification was performed as in chapter 4 section 4.2.1. The same procedure was followed for all drugs in the dataset. The total number of cell lines tested, and as such the number of sensitive and resistant cell lines, differed per compound. Splicing quantification was performed as in chapter 5 section 5.2.2.

6.2.2 Filtering of Differentially Spliced Events

Differentially spliced events across sensitive and resistant cell lines, divided by tissue, were identified for each drug using the quasi-binomial generalized linear modeling framework from chapter 5. A FDR cutoff ≤ 0.05 was applied to filter low-significance events. To select for accuracy and biological significance in each event, we required at least 35% of the cell lines that were tested for drug response have splice-junction reads and a minimum difference in the fraction of included reads between sensitive and resistant cell lines ≥ 0.05.

6.2.3 Construction of Tissue-specific Drug Splicing Networks

To ensure adequate data for analysis, tissue types were selected based on the number of cell lines that were available from the integrated data set. We required a minimum of 60 cell lines per tissue (**Figure 18**). After filtering, networks were constructed for six tissue types; breast, central nervous system, hematopoietic & lymphoid, large intestine, lung,

and skin. Drug-drug networks for each tissue were constructed using drugs as nodes and a Jaccard index modified for splicing data to calculate pairwise edge weights:

$$w_{ij} = \frac{|i \cap j| - D_{ij}}{|i \cup j|}$$

In the modified Jaccard index, i is all events from the first drug in a pair, j is all events from the second drug, D_{ij} is divergent events, (i.e. the number of exons observed to have a higher inclusion level in sensitive cells of the first drug but lower inclusion level in sensitive cells of the second or *vice versa*), and w_{ij} is edge weight. A network adjacency matrix, where nodes are in rows and columns and edge weights the pairwise values, was used to represent each network. For bipartite network inspection, an incidence matrix was built with all significant events (from all drugs) in a network used as rows and drugs as columns.

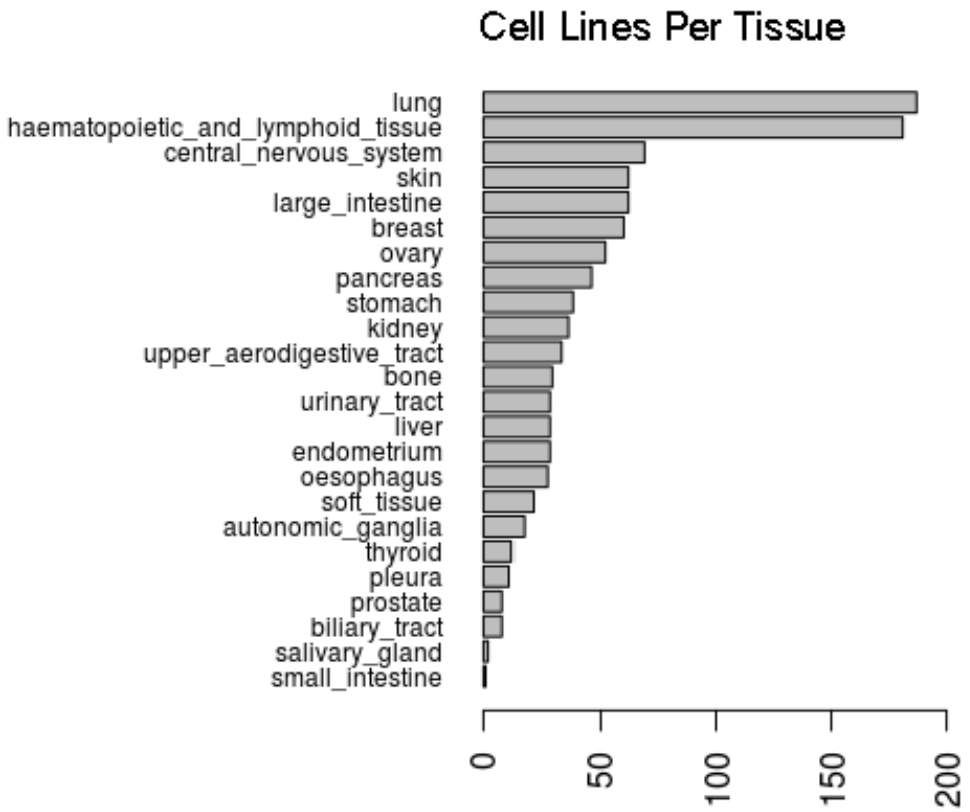

Figure 18 Cell line counts in the integrated CCLE and CTRP data set. Cell types are sorted by total cell lines with paired data in descending order.

6.2.4 Drug Module Identification

Module identification was accomplished using the network adjacency matrix in three steps. First, hierarchical clustering was performed on the network matrix in R using the *hclust* function with average distance [177]. Next, all clusters with between 3 and 15 members were extracted and clusters were merged if one contained all members of another. Significance of each module was then determined using the *cb-signi* software package [118]. Briefly, significance of network modules in random networks is determined by the total number of nodes in a network, the total number of edges in a network, the number of nodes in a module, and the number of edges in a module [118]. The probability of a community occurring at random, calculated through the cumulative probability of the other nodes in the community combined with permuting the worst node (most likely to not be a community member), is referred to as "community score", or *c-score* [118]. A low *c-score* (under 0.05) indicates the module is significant [118]. The *c-score* can be extended to consider multiple worst nodes and redefined as "border score", or *b-score* [118]. Clusters in tissue-specific networks were filtered for those having b- and *c-scores* (significance) of ≤ 0.05.

6.2.5 Drug Module and Event Annotation

Modules were annotated with drug activity from CTRP [21]. Gene symbols for events were annotated using the same GTF file used to identify potential spliced events. All protein structure, domain and functional information was annotated manually with the Simple Modular Architecture Research Tool (SMART) and GeneCards databases [203–205].

6.3 Results

6.3.1 Drug-splicing Networks Show Cluster Structure

Complex networks were built from adjacency matrices for six tissue types. Select network metrics are in (**Table 5**). Networks exhibited random network structure with a large number of edges, suggesting many spliced events are related to drug response in multiple compounds. Visualization of adjacency matrices after clustering showed obvious community structures throughout the networks (**Figure 19**). Blocks of higher edge weights about the zero-line of each heatmap indicated groups of drugs that were tightly connected (**Figure 20**), and similar patterns in connectivity to other nodes trailing to the axes indicated drugs in communities tend to have comparable edge weights with drugs outside of the community. The number of communities appeared directly related to the number of cell lines underlying each network.

Tissue	Cell Line Count	Nodes	Edges	Mean Degree	Unattached Nodes
Breast	60	500	21717	87	100
CNS	69	498	20580	83	106
H and L	181	501	241455	964	7
Large Intestine	62	494	22464	91	73
Lung	187	501	242268	965	17
Skin	62	501	36672	146	63

Table 5 Tissue-specific drug splicing network metrics. Cell line count is the number of cell lines from the tissue of origin available in the overlapped CCLE and CTRP dataset. CNS, central nervous system; H and L, hematopoietic and lymphoid.

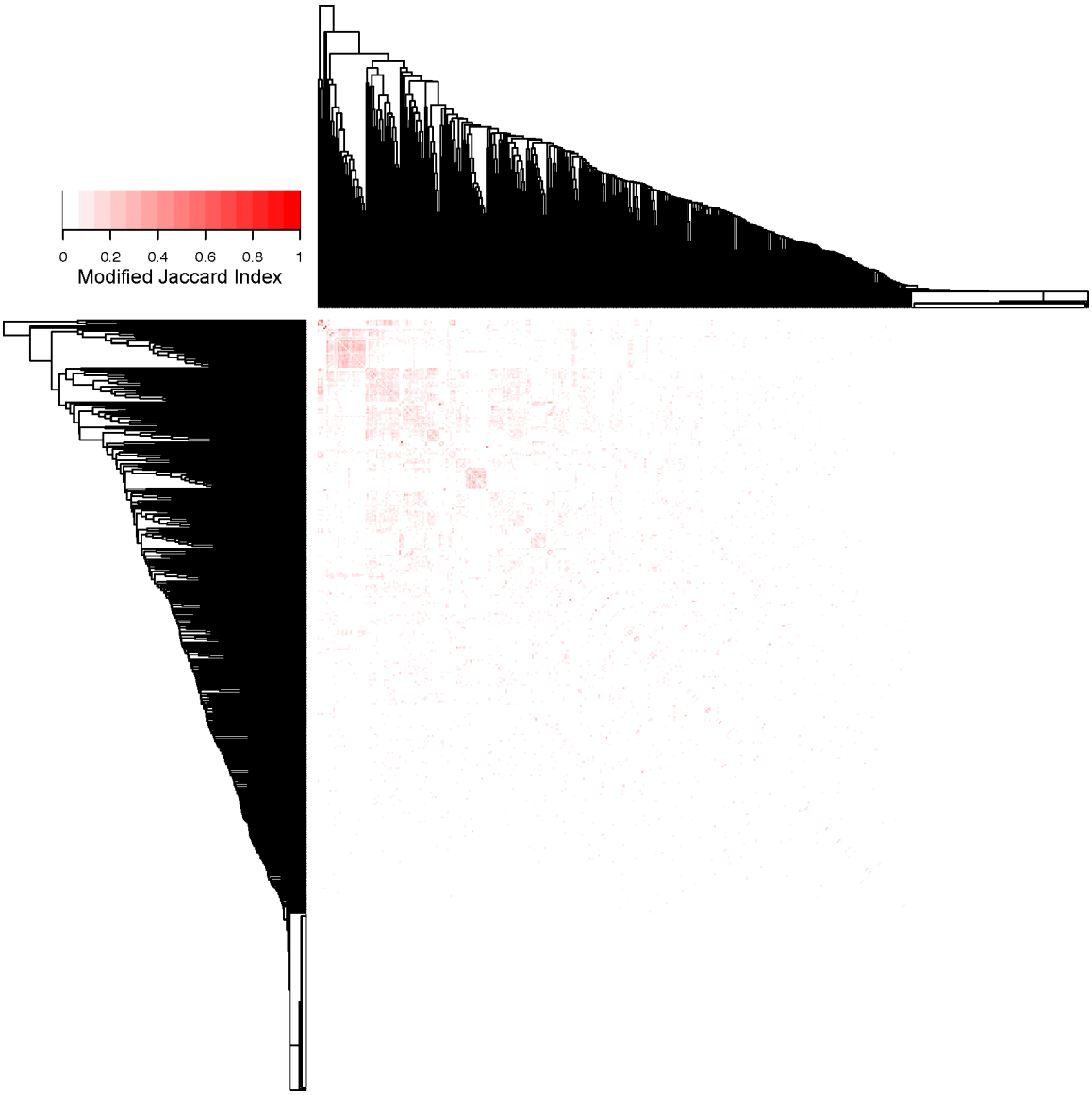

Figure 19 Heatmap of edge weights in the breast tissue drug network. Modified jaccard index for pairwise drug combinations scaled from white to red. Side dendrogram shows tree from clustering by average distance. Heatmap was imaged symmetrically by row and column to show patterns in edge weights.

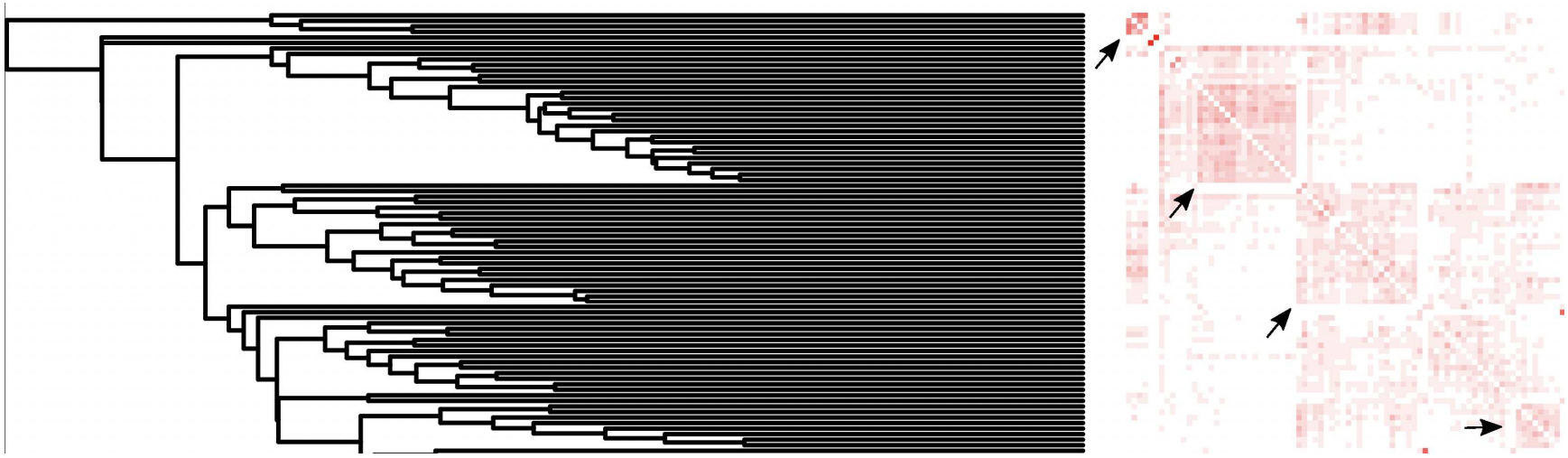

Figure 20 Clusters of high edge weights in the breast tissue drug network. Close up of breast tissue heatmap. Black arrows point to block-like organization of tightly connected drug communities.

6.3.2 Tissue-specific Module Identification Depends on Cell Line Coverage

To better understand community structure, we performed bipartite inspection on drugs and events comprising the edge weights. We discovered a small number of drugs with many differentially spliced events could drive wide-scale connectivity. These drugs created "hairballs" in the network by facilitating weak connections with other nodes and made module identification more difficult (**Figure 21**). Therefore, we applied a cluster- and significance-based module identification strategy that would take advantage of the global connectivity in the network to group community members, while considering the probability of community pairwise interactions to minimize the size of extracted modules.

The total number of modules we identified in each network, pre- and post- filtering, are in **Table 6**. Total module counts for each network paralleled the number of edges in the networks and like community structures were directly dependent on the number of underlying cell lines used for construction. However, filtered module counts were similar across networks. All tissue networks returned a substantial number of modules; this suggested that certain spliced events have a broad influence on pre-treatment sensitivity in various compounds, and the effects of splicing can be observed in numerous tissues. Modules identified for each network, along with significance scores, are in **Tables 7-12**.

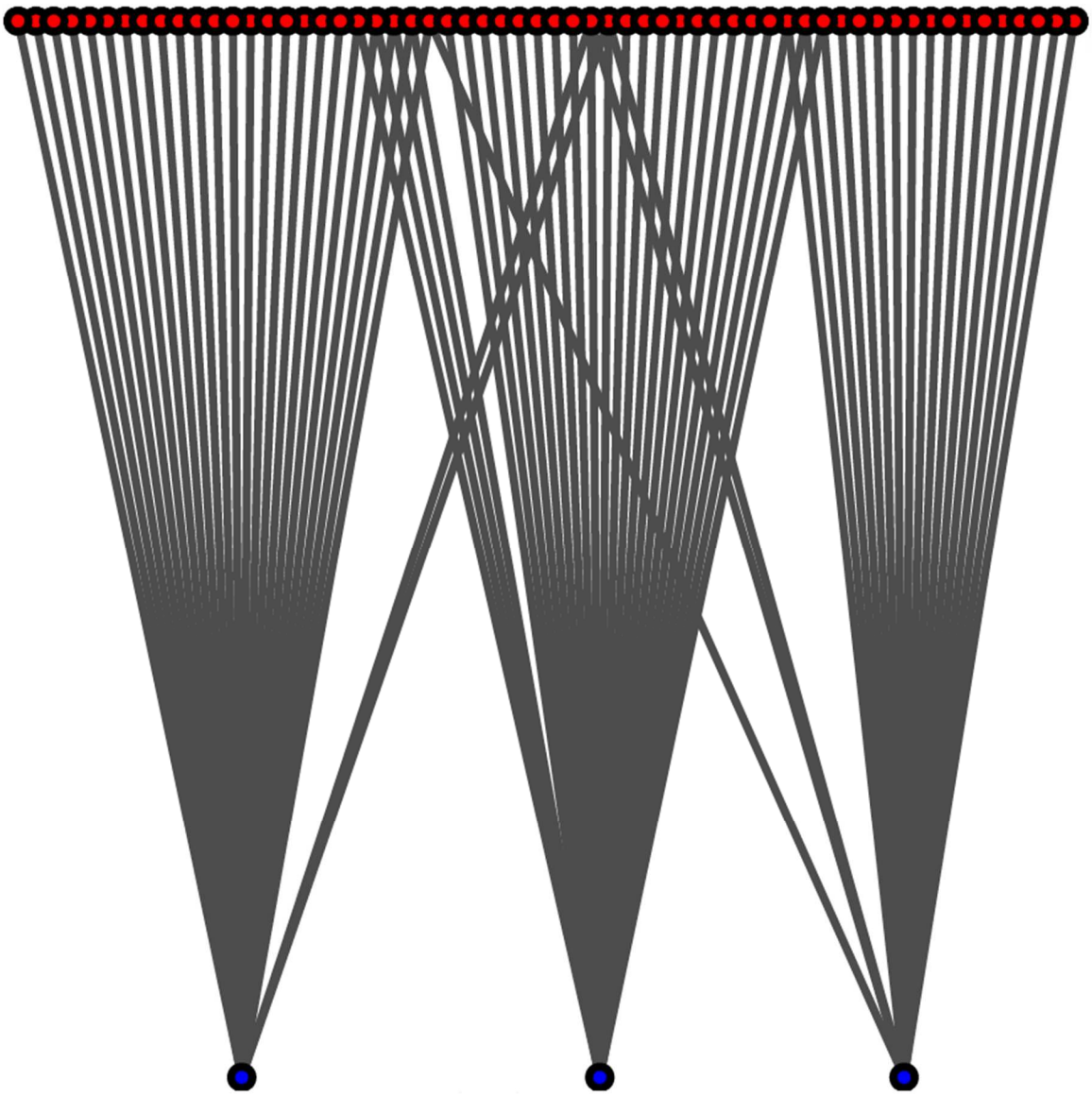

Figure 21 Bipartite plot of module 27 from the skin tissue drug network. This module is an example of three drugs with weak connections due to many differentially spliced events. Between three drugs a total of 10 out of 60 events were observed in at least two community members. Two events were present in 3 drugs. Despite these drugs clustering together, this module was eliminated during filtering because it was not significant by *c*- or *b-score*.

Tissue	Modules	Significant Modules
Breast	26	24
CNS	24	20
H and L	54	24
Large Intestine	24	22
Lung	54	24
Skin	29	26

Table 6 Tissue-specific module counts. CNS, central nervous system; H and L, hematopoietic and lymphoid.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	3	0.11	0	0
2	4	0.16	0	0
3	4	0.11	0	0
4	4	0.17	0	0
5	13	0.16	0	0
6	3	0.14	0	2.99e-24
7	15	0.2	0	0
8	10	0.12	0	0
9	8	0.16	0	0
10	5	0.07	0.0198996	6.88e-05
11	3	0.15	0	1.47e-31
12	12	0.06	0.84707	0.389374
13	15	0.12	7.04e-08	3.03e-10
14	3	0.09	0.329597	0.247971
15	5	0.1	4.80e-11	8.55e-15
16	3	0.23	0	0
17	3	0.15	0	0
18	3	0.15	1.69e-12	1.30e-12
19	3	0.17	1.06e-13	6.05e-14
20	3	0.12	5.31e-10	4.28e-10
21	5	0.09	2.00e-05	1.36e-05
22	13	0.12	0.000219527	2.01e-11
23	3	0.19	0	0
24	3	0.26	0	0
25	5	0.19	0.00601248	0.0058375
26	4	0.44	0	0

Table 7 Breast drug network module statistics. Each row represents a clustered module in the drug network.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	3	0.02	1	1
2	4	0.06	1	0.977287
3	10	0.19	0	0
4	7	0.09	0	0
5	13	0.1	0	0
6	7	0.18	0	0
7	6	0.08	0	0
8	3	0.17	0	0
9	8	0.11	0.0033545	0.00264043
10	10	0.11	6.88e-05	1.12e-08
11	3	0.07	1	1
12	11	0.4	0	0
13	5	0.11	0	0
14	7	0.15	0	0
15	6	0.12	9.11e-07	1.45e-08
16	3	0.11	1.55e-05	1.63e-05
17	5	0.17	2.45e-09	0
18	5	0.27	4.21e-12	4.14e-12
19	3	0.17	0.342685	0.327845
20	3	0.26	0	0
21	3	0.25	0.00010813	0.00011405
22	6	0.27	0	0
23	5	0.15	0.00010667	4.93e-07
24	3	0.33	0	0

Table 8 Central nervous system drug network module statistics. Each row represents a clustered module in the drug network.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	15	0	0.999343	1
2	6	0.03	0.261985	0.0744
3	11	0.03	8.93e-06	3.39e-12
4	14	0.01	0.999638	1
5	4	0.06	0	0
6	8	0.02	0.184204	8.21e-07
7	8	0.02	0.0389182	0.024968
8	15	0.02	0.0153037	0.0110133
9	7	0.06	2.23e-13	0
10	8	0.02	1	1
11	5	0.04	0.00057775	0.00066691
12	10	0.08	0	0
13	3	0.05	0.00014015	7.62e-05
14	6	0.08	7.99e-14	0
15	14	0.03	0.880236	0.272504
16	9	0.02	0.752034	0.779035
17	5	0.05	0.00047449	1.85e-06
18	4	0.03	0.0280366	0.00828057
19	8	0.04	0.946273	0.893683
20	6	0.04	0.009987	0.00562952
21	9	0.04	0.722979	0.691368
22	9	0.05	0.99999	0.999994
23	3	0.09	0.00272625	0.00290391
24	4	0.05	0.999904	0.958036
25	14	0.07	0.796684	0.389119
26	9	0.07	1.12e-09	0
27	13	0.07	0.111263	6.72e-05
28	3	0.07	0.0054361	0.00254429
29	4	0.05	0.796467	0.771549
30	7	0.19	0	0
31	6	0.12	0	0
32	4	0.07	1	1
33	11	0.1	1	1
34	3	0.05	1	1
35	7	0.13	1	1
36	8	0.19	2.28e-13	0
37	10	0.13	1	1
38	8	0.17	0.998039	0.988809

39	9	0.14	1	1
40	11	0.21	0.295204	8.12e-07
41	7	0.14	0.999993	0.999321
42	13	0.16	1	1
43	5	0.1	0.364562	0.354799
44	15	0.21	0.96689	0.806223
45	12	0.21	0.248356	0.191767
46	4	0.24	0.977163	0.979537
47	12	0.29	0.00160677	0.00152128
48	3	0.25	0.0225351	3.58e-06
49	8	0.27	0.260138	0.029127
50	4	0.29	4.07e-05	1.13e-05
51	3	0.14	1	1
52	10	0.29	1.21e-06	3.89e-11
53	12	0.35	4.72e-11	2.67e-15
54	4	0.36	0.0070488	0.00047966

Table 9 Hematopoietic and lymphoid drug network module statistics. Each row represents a clustered module in the drug network.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	14	0.07	0	0
2	15	0.13	0	0
3	6	0.09	0	0
4	6	0.16	0	0
5	15	0.08	5.32e-12	0
6	4	0.1	0	0
7	3	0.14	0	0
8	6	0.12	0	0
9	7	0.11	4.27e-08	8.58e-09
10	3	0.08	1.64e-05	9.19e-06
11	7	0.16	0	0
12	3	0.11	5.13e-12	2.77e-12
13	6	0.16	0	0
14	6	0.16	0	0
15	3	0.13	1.92e-07	2.10e-07
16	3	0.08	0.00050301	0.00029115
17	6	0.14	0	0
18	3	0.1	0.133376	0.129384
19	3	0.14	1	1
20	7	0.18	3.81e-05	3.56e-05
21	7	1	0	0
22	3	1	0	0
23	7	1	0	0
24	3	1	0	0

Table 10 Large intestine drug network module statistics. Each row represents a clustered module in the drug network.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	10	0.02	0.186503	2.21e-07
2	4	0.04	0	0
3	5	0.03	9.57e-07	5.37e-09
4	10	0.01	0.999625	1
5	4	0.03	9.47e-05	8.17e-05
6	4	0.03	0.989295	0.710205
7	4	0.02	0.357353	0.135326
8	3	0.04	0.00150324	0.00125364
9	13	0.02	2.18e-08	2.90e-15
10	11	0.02	0.00680177	0.00053472
11	11	0.03	8.23e-11	1.14e-13
12	7	0.04	2.69e-12	0
13	4	0.02	0.998103	0.948496
14	4	0.06	1.33e-08	1.10e-08
15	14	0.03	0.0005284	0.00019257
16	5	0.02	0.707335	0.681246
17	5	0.02	0.99471	0.803602
18	13	0.03	0.970937	0.925808
19	5	0.04	0.266776	0.208291
20	6	0.1	0	0
21	3	0.05	0.017976	0.0136504
22	13	0.05	0.159213	0.00879935
23	7	0.04	0.982697	0.839345
24	3	0.09	6.28e-06	3.86e-06
25	7	0.06	0.030158	0.0299761
26	12	0.06	0.141314	0.156973
27	5	0.09	0.00866474	0.00748602
28	7	0.06	0.659764	0.630083
29	7	0.18	0	0
30	3	0.06	0.999918	0.999938
31	15	0.08	0.999537	0.972333
32	3	0.19	0	0
33	15	0.11	0.573053	0.605238
34	12	0.12	0.999992	0.646473
35	9	0.16	0.999939	0.99993
36	7	0.12	1	1
37	5	0.13	1	1
38	12	0.16	1	1

39	13	0.17	0.915604	0.352232
40	4	0.11	0.171863	0.181508
41	10	0.18	1	1
42	5	0.13	1	1
43	9	0.22	0.00017154	1.20e-07
44	15	0.26	0.0744473	3.07e-05
45	11	0.26	0.034897	1.75e-06
46	12	0.23	0.92408	0.900807
47	3	0.18	1	1
48	4	0.26	1.39e-05	4.35e-06
49	3	0.27	0.656219	0.619115
50	15	0.3	2.39e-08	0
51	5	0.25	0.329217	0.325568
52	4	0.38	9.73e-05	1.39e-05
53	12	0.35	3.53e-09	9.16e-14
54	3	0.33	0.0438672	0.0445675

Table 11 Lung drug network module statistics. Each row represents a clustered module in the drug network.

Cluster ID	Nodes	Average Edge Weight	*c-score*	*b-score*
1	3	0.05	7.84e-07	4.19e-07
2	14	0.08	0	0
3	13	0.06	6.49e-06	1.83e-13
4	7	0.12	0	0
5	12	0.09	0	0
6	4	0.13	0	0
7	4	0.27	0	0
8	5	0.13	0	0
9	8	0.13	0	0
10	10	0.11	0	0
11	14	0.07	2.63e-06	5.25e-10
12	8	0.15	0	0
13	3	0.07	0.00829008	0.00810006
14	3	0.11	0.00268105	0.00319062
15	7	0.18	0	0
16	14	0.23	0	0
17	8	0.07	0.00134095	3.53e-07
18	6	0.22	0	0
19	8	0.12	0	0
20	12	0.09	0.999999	0.0325464
21	5	0.09	3.17e-05	3.24e-05
22	4	0.26	0	0
23	5	0.13	0	0
24	9	0.11	1.12e-10	7.17e-15
25	5	0.16	0.810523	0.820397
26	4	0.26	0.00042842	0.000104
27	3	0.11	0.999992	0.999549
28	5	0.17	8.33e-08	8.93e-08
29	3	1	0	0

Table 12 Skin drug network module statistics. Each row represents a clustered module in the drug network.

6.3.3 Compounds Within Modules Share Components and Activities

Some of the stronger modules we identified, as defined by larger size and a high average edge weight, shared many of the same drug members in multiple tissue types. Two examples of modules with similar members in multiple tissues were the modules containing selumetinib compounds (**Table 13**), bromodomain inhibitors (**Table 14**). Other modules were more specific to the respective tissue and members did not cluster well in other tissues. Three such examples are the lung modules for drugs with activities based on inhibition of *EGFR* (lung module 20, **Figure 22A**), nicotinamide phosphoribosyltransferase (*NAMPT*) (lung module 32, **Figure 22B**) and *BCL2*/BCL-xL (lung module 34, **Figure 22C**).

Compound	H and L	Lung	Breast	Activity
selumetinib.UNC0638..4.1.mol.mol.	x	x	x	inhibitor of *MEK1* and *MEK2*;inhibitor of *EHMT2* and GLP methyltransferase
selumetinib.tretinoin..2.1.mol.mol.	x	x	x	inhibitor of *MEK1* and *MEK2*;agonist of retinoid acid receptors
selumetinib.BRD.A02303741..4.1.mol.mol.	x	x	x	inhibitor of *MEK1* and *MEK2*;inhibitor of histone methyltransferases
selumetinib.PLX.4032..8.1.mol.mol.	x	x		inhibitor of *MEK1* and *MEK2*;inhibitor of *BRAF*
PD318088	x	x		inhibitor of *MEK1* and *MEK2*
selumetinib	x	x	x	inhibitor of *MEK1* and *MEK2*
selumetinib.MK.2206..8.1.mol.mol.	x	x	x	inhibitor of *MEK1* and *MEK2*;inhibitor of *AKT1*
selumetinib.decitabine..4.1.mol.mol.			x	inhibitor of *MEK1* and *MEK2*;inhibitor of DNA methyltransferase
selumetinib.GDC.0941..4.1.mol.mol.			x	inhibitor of *MEK1* and *MEK2*;inhibitor of PI3K kinase activity
selumetinib.piperlongumine..8.1.mol.mol.			x	inhibitor of *MEK1* and *MEK2*;natural product; modulator of ROS levels
PD318088			x	inhibitor of *MEK1* and *MEK2*
selumetinib.vorinostat..8.1.mol.mol.			x	inhibitor of *MEK1* and *MEK2*;inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*
serdemetan.SCH.529074..1.1.mol.mol.			x	inhibitor of *MDM2*;activator of mutant p53
erlotinib.PLX.4032..2.1.mol.mol.			x	inhibitor of *EGFR* and *HER2*;inhibitor of *BRAF*
SCH.529074			x	activator of mutant p53

Table 13 Selumetinib drug community in three tissue networks. Presence of drug members in tissue-specific identified communities is indicated by "x" for three tissue drug networks; hematopoietic and lymphoid, lung and breast.

Compound	H and L	Lung	Breast	Activity
JQ.1.MK.0752..1.1.mol.mol.	x	x		inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins;inhibitor of gamma-secretase
JQ.1.navitoclax..2.1.mol.mol.	x		x	inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins;inhibitor of *BCL2*, BCL-xL, and BCL-W
GSK525762A	x	x	x	inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins
JQ.1.vorinostat..2.1.mol.mol.	x			inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins;inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*
I.BET151	x			inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins
JQ.1		x		inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins
I.BET151		x	x	inhibitor of bromodomain (BRD) and extra-C terminal domain (BET) proteins
apicidin	x			inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*
ISOX	x	x		inhibitor of *HDAC6*
BRD.K61166597		x		inhibitor of *HDAC1* and *HDAC2*
vorinostat		x		inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*
vorinostat.carboplatin..1.1.mol.mol.		x		inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*;inducer of DNA damage
LBH.589			x	inhibitor of *HDAC1*, *HDAC2*, *HDAC3*, *HDAC6*, and *HDAC8*
TG.101348	x	x		inhibitor of Janus kinase 2
NVP.BSK805	x			inhibitor of Janus kinase 2
piperlongumine	x			natural product; modulator of ROS levels

curcumin	x			natural product; modulator of ROS; modulator of *NF-kB* signaling
PL.DI		x		dimer of piperlongumine; inducer of ROS
decitabine.navitoclax..2.1.mol.mol.	x			inhibitor of DNA methyltransferase;inhibitor of *BCL2*, BCL-xL, and BCL-W
YK.4.279		x		inhibitor of RNA helicase A (RHA) binding to *EWS-FLI1*; inhibitor of *ERG* and *ETV1* activity
ML311		x		inhibitor of *MCL1*
sunitinib			x	inhibitor of *VEGFRs*, *c-KIT*, and *PDGFR alpha* and *beta*

Table 14 Bromodomain drug community in three tissue networks. Presence of drug members in tissue-specific identified communities is indicated by "x" for three tissue drug networks; hematopoietic and lymphoid, lung and breast.

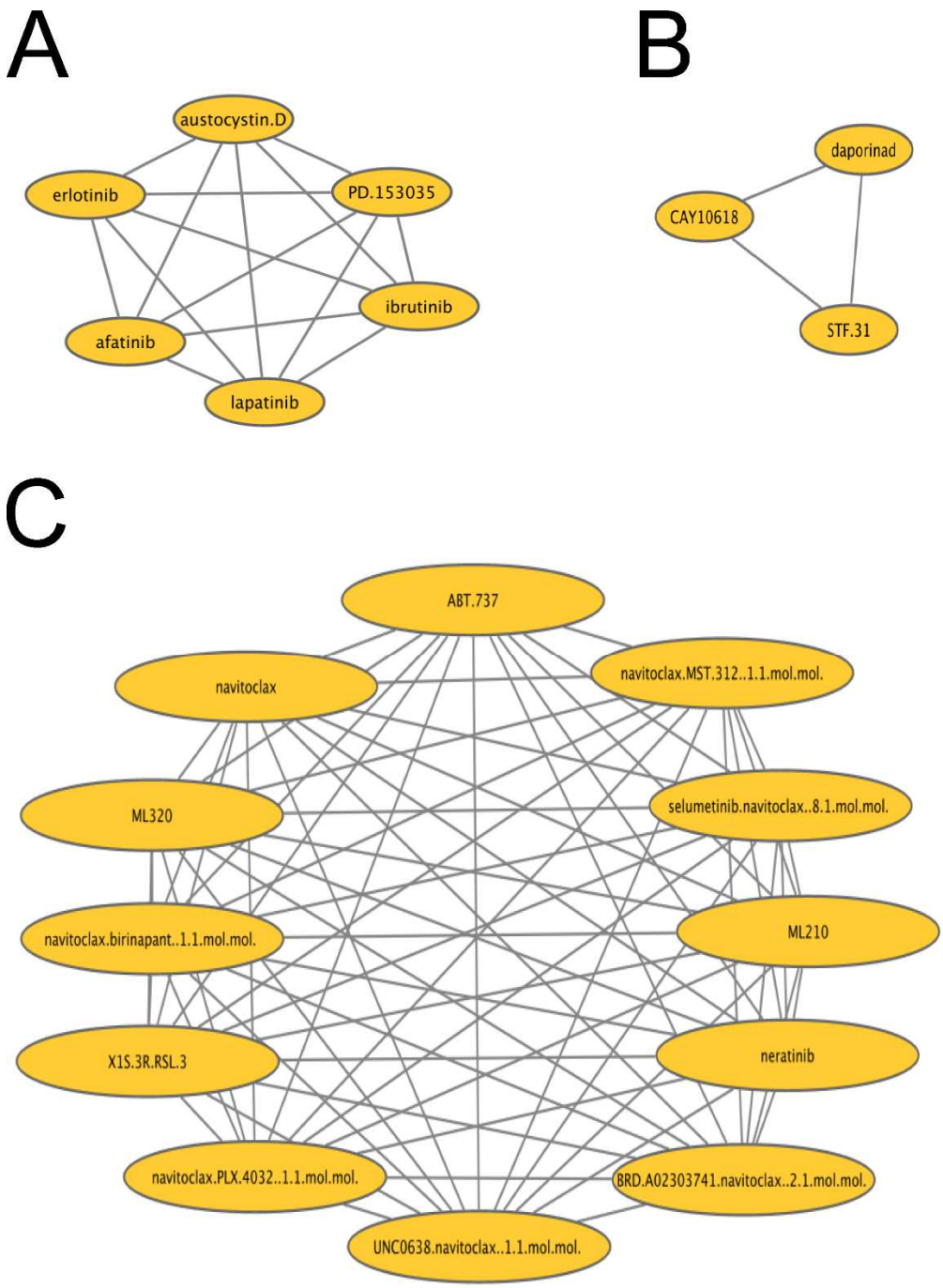

Figure 22 Network diagrams of three drug communities in the lung tissue drug network: A. Module 20; B. Module 32; C. Module 34.

6.3.4 Commonality Among Spliced Events

To understand the relevance of individual spliced events across multiple tissues, we further compared overlapping events from the selumetinib kinase inhibitor and bromodomain inhibitor module groups in hematopoietic and lymphoid, lung and breast tissue; In the selumetinib activity group these were hematopoietic and lymphoid module 30, lung module 29 and breast module 5. In the bromodomain group these were hematopoietic and lymphoid module 47, lung module 53 and breast module 25. We found the selumetinib group did not share any events across all three tissues. However, 13 events were shared between hematopoietic and lymphoid and lung tissues, and two events were shared between hematopoietic and lymphoid and breast tissues. For the bromodomain activity group we found 346 events, originating from 268 genes, common between all three tissues. Despite the extreme number of shared events in the bromodomain activity group, generally we found that although some modules appeared to share similar drug activities across tissues both the members and events connecting them were largely unique. This indicates that while splicing may be important to drug response in multiple tissues, and in some cases events can translate across tissues, for the majority of drug activities the specific events and mechanisms influencing drug response may vary.

6.3.5 Genes, Protein Domains and Mechanisms in Drug Modules

Due to the huge volume of data produced during network analysis, and the time required to manually annotate spliced events, we chose to evaluate the three lung-specific modules that were discussed in section 6.3.3; modules with activities based on inhibition of *EGFR*

(lung module 20), *NAMPT* (lung module 32) and *BCL2*/BCL-xL (lung module 34). Top annotated genes for the modules are in **Tables 15-17**. Both the *EGFR* and *BCL2*/BCL-xL modules shared spliced events in genes related to cellular motility and adhesion. Key genes annotated to these two modules included *CD44*, *CTNND1*, *GSN*, *ADAM15*, *TCF7L2* and *GIT2*.

Next, we annotated module events with protein structure information and discovered multiple spliced exons were spanning regions coding for key protein domains. The transmembrane protein *CD44* is found in leukocytes and regulates transmigration to inflammatory sites [206]. We found multiple events spanning the hyaluronan binding domain; a key region essential for *CD44* functionality [206]. Additionally, we found exons spanning the SH3 binding domain of *ADAM15*, and the *ALG2* binding domain of *SEC31A*. Taken together, these results indicate that differentially spliced exons associated with drug response can have important functional consequences, and that common events can have an influence on multiple groups of drugs with differing activities.

Gene	Event Count	Function
CD44	27	Cell-surface glycoprotein, cell-cell interactions
CTNND1	7	Cell adhesion and signal transduction
GSN	7	Actin filament assembly and disassembly
DDR1	5	Receptor tyrosine kinase activated by collagen
MLPH	4	Binds to myosin 5A, an actin bound transport protein
SEC31A	4	Member of coat protein complex II and is involved in vessicle budding from the endoplasmic reticulum
TCF7L2	4	Transcription factor that has a key role in Wnt signaling
CAST	3	Calapin inhibitor involved with membrane fusion events
COL16A1	3	Alpha chain of type XVI collagen that maintains the extracellular matrix
ENAH	3	Regulates assembly of actin filaments and cellular adhesion

Table 15 Genes annotated to the *EGFR* module in the lung tissue drug network. Event counts are the number of differentially spliced events in the module annotated to each gene. Functions were annotated from GeneCards [205].

Gene	Event Count	Function
CLDND1	4	Claudin Domain Containing 1
LTBP1	4	TGF-beta binding protein
EEF1D	3	Subunit of elongation factor-1 complex
BCAS1	2	Candidate oncogene amplified in tumors
CD151	2	Cell surface glycoprotein mediating signal transduction and cell adhesion
CD47	2	Membrane protein invoveld in membrane transport and signal transduction
COL16A1	2	Alpha chain of type XVI collagen that maintains the extracellular matrix
COL6A3	2	Alpha-3 chain of type VI collagen
ENOSF1	2	Mitochondrial enzyme and antisense to thymidylate synthase
FGD3	2	FYVE, RhoGEF and PH Domain Containing 3

Table 16 Genes annotated to the *NAMPT* module in the lung tissue drug network. Event counts are the number of differentially spliced events in the module annotated to each gene. Functions were annotated from GeneCards [205]

Gene	Event Count	Function
CD44	16	Cell-surface glycoprotein, cell-cell interactions
ADAM15	10	Transmembrane glycoprotein involved in cell adhesion and processing of cytokines
GIT2	8	GPCR interactor that traffics between cytoplasmic complexes and regulates cytoskeletal dynamics
CTNND1	7	Cell adhesion and signal transduction
TCF7L2	7	Transcription factor that has a key role in Wnt signaling
ARFGAP2	6	ADP Ribosylation Factor GTPase Activating Protein 2
CBWD6	5	COBW Domain Containing 6
HMGN1	5	Binds nucleosomal DNA and transcriptionally active chromatin
ABI1	4	Facilitates signal transduction and regulates actin polymerization and cytoskeletal remodeling
BANP	4	Tumor suppressor gene that negatively regulates p53 transcription

Table 17 Genes annotated to the *BCL2* module in the lung tissue drug network. Event counts are the number of differentially spliced events in the module annotated to each gene. Functions were annotated from GeneCards [205]

6.4 Discussion

In chapter six we show that pre-treatment splicing profiles generated from cell lines based on drug sensitivity can be used to link drugs in tissue-specific networks. Our findings indicate that specific spliced events may impact sensitivity to multiple compounds and that treatment of many cancers, arising across a wide array of tissues, could be influenced by differentially spliced transcripts. Additionally, we found that the specific events connecting drugs, and members of connected drug modules, differ by tissue.

We found that the network structure, and expectedly the total number of modules identified, was directly dependent on the number of cell lines used to construct the networks. Despite this, we still identified a similar number of significant modules in each tissue that reflect the similarities in drug activities and the relevance of pre-treatment splicing profiles in drug response. We anticipate that if more cell lines were available for analysis the power of our network splicing approach would improve. Given the potential for additional network connectivity, it is possible that many relationships between drugs defined by differentially spliced events remain undetected. Additionally, having a data set that originated from a single source, as opposed to an integrated data set, might improve the consistency of the data and reduce technical noise that could be impeding our analysis.

Chapter 7 Conclusions and Future Directions

7.1 Conclusions

7.1.1 Conclusions on Differential Gene Expression Analysis of Concussed Athletes

Given the prevalence of automobile accidents, household mishaps and participation in adolescent sports, concussion continues to be a major public health concern. Currently there is a lack of treatment options for concussed individuals which leaves medical professionals with little recourse. Implementation of advanced experimental techniques and integration of multi-omics data is desperately needed to improve our understanding of concussions. In our study, we characterized differentially expressed genes in the largest concussion cohort to date. We identified similarities in the gene expression profiles of injured participants that matched pathophysiology of concussion. We also identified alterations in key immune signaling pathways that specify the type of immune and wound healing responses following injury. Deconvolution analysis also indicated a difference in immune cell type proportions pointing to neutrophils as key players in concussion immune response. We determined that known blood-based biomarkers cannot be detected in gene expression data, however there is potential to use transcriptome analysis following injury to monitor the short-term immune response and possibly inform researchers on the best course of action once new treatment options become available.

7.1.2 Conclusions on Predictive Modeling of Drug Response with Splicing Data

Our study demonstrated that differentially spliced events can separate cell lines by drug sensitivity before treatment. We also demonstrated predictive power across hundreds of compounds which suggested that splicing does impact drug response at a broad level. We

did not heavily optimize our modeling pipeline or test more advanced machine learning techniques like neural networks. We anticipate there is significant room for improvement in predictive modeling with splicing data and integration with other data types. Additionally, our model was based solely on skipped-exon splicing data. However, even though we chose to use one type of spliced event, we still established a strong link between splicing and drug response. Finally, our goal was to separate cell lines by drug response with splicing data, although it is likely that our doxorubicin model separated cell lines by both drug response and cell type related features. We do not see this as a failure though since splicing does play a key role in defining cell types and the biological nature of some cells will make them more susceptible to certain drugs.

7.1.3 Conclusions on Quasi-binomial GLM for Differential Splicing Analysis

Quasi-binomial generalized linear modeling is a powerful tool for analysis of modern, large volume splicing datasets. In our study we processed splicing data from almost one thousand cell lines in mere minutes; that scale of analysis on older methods would have taken weeks, assuming the massive computational resources required could be made available. We also confirmed our method identified relevant spliced events, given the condition we were investigating, by annotating events to genes that were previously implicated in epithelial-mesenchymal transition. Additionally, an event we identified contained sequences that coded for an essential protein domain; this established a connection between individual exons and functional consequences that could impact drug response. Still, it may be possible to improve the performance of QBGLM on smaller

datasets by adding additional model parameters. Despite this, we expect our model will have wide applicability and accelerate analysis of large modern datasets.

7.1.4 Conclusions on Analysis of Drug Splicing Networks

Through network analysis of drug splicing data we determined that compounds sharing the same drug components or activities can be clustered by differentially spliced events. Additionally, our analysis shows that splicing influences many compounds across multiple tissues, that in some cases events mediating drug response can be relevant to multiple tissues, and that in other cases underlying spliced events can be tissue-specific. Finally, we annotated key genes and specific protein domains to differentially spliced exons that could help researchers investigate a mechanism explain differences in drug response among tissue sub-groups. Our work provides a vast dataset that hold tremendous potential for advancing the current understanding of drug response dynamics. More work is needed to thoroughly catalogue the relationships between drugs and exons hiding in the networks.

7.2 Future Directions

7.2.1 Future Directions of Research in Predictive Biomarkers for Concussed Athletes

More research into concussion response and healing is needed to understand the potential medical interventions that can benefit concussion patients. To further investigate changes in the transcriptome, additional samples with full time courses are required to analyze changes in individuals over time. Additional samples would enable longitudinal analysis

and improve the power of differential gene expression detection methods. A much larger cohort, with whole-genome sequencing data, is necessary to investigate the potential association between genetics and delayed recovery. Overall, knowledge on concussion is improving; for example we have learned that alcohol use is a major factor influencing delayed recovery, sex based differences lead to longer recovery times in women and there is a substantial connection between repetitive low-impact events and neurodegeneration later in life. Based on this some future directions for analysis in our data set include integrating our results with those from studies in alcohol use disorder, investigating sex-specific differences in our cohort and long-term follow-up of medical data in contact and non-contact controls to compare differences in neurodegeneration rates. Splicing of transcript isoforms following injury can also be investigated, although deeper sequencing would be required for solid transcriptome coverage. CARE continues to enroll new participants and has secured funding for expanding their research goals in the coming years. Part of the next phase of the CARE initiative will be single-cell RNAseq analysis, which will separate the biological signals we observed in bulk data and confirm changes in immune cell types following injury that we found in our deconvolution analysis.

7.2.2 Future Directions of Research in Differentially Spliced Exons Mediating Drug Response Sensitivity

Existing public data sets, that researches rely on to investigate drug response, were collected before RNAseq became the dominant transcriptome analysis technology. Larger drug response data sets accompanied by RNAseq and other multi-omic data are needed to improve consistency and predictive modeling capabilities. We expect that our drug

response modeling strategy could be improved in four ways: one, additional event types besides skipped-exon events could be included to increase the number of features investigated; two, integration with other omics data will provide a more comprehensive biological information; three, further optimization of the feature selection and training procedure will improve overall performance; and four, advanced machine learning methods such as neural networks could be applied to the data. In the coming years we anticipate many researchers will integrate isoform-specific information into drug response models.

Until whole-isoform sequencing becomes affordable enough to perform en masse, QBGLM could be relevant to process new experimental data. We expect that QBGLM could be improved by adding variance controlling functionality that adjusts model parameters based on observed data metrics and sample counts; this would allow for accurate analysis of small sample sets and enhance performance in larger data sets as well. Additionally, incorporating a method to identify unannotated events would help account for incomplete annotation of reference genomes. Finally, repackaging QBGLM into an easily accessible tool, compatible with existing splicing annotation, would help researchers quickly implement the method in their research.

Due to the size of the integrated data set we used for network analysis, there is still a substantial amount of work that can be done to investigate splicing relationships between communities of drugs. Additionally, our network analysis could be improved by testing additional edge weighting schemes and module identification algorithms. Our study

would also benefit from validation of underlying events using additional data sets or wet lab experiments. Finally, we could extend the utility of our networks by investigating other biological questions such as splicing mediated drug toxicity and developing a strategy to find complementary drug pairs.

Lastly, a substantial issue that surfaced repeatedly during our research was that currently no resource exists which catalogues, annotates, and visualizes differentially spliced events in a centralized and user-friendly manor. We see this as a major detriment to studying splicing in transcript isoforms and we anticipate development of a centralized splicing-oriented resource would provide a valuable service to the research community.

References

[1] Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet 2001;17:100–7. https://doi.org/10.1016/S0168-9525(00)02176-4.

[2] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5. https://doi.org/10.1038/ng.259.

[3] Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. JAMA 2021. https://doi.org/10.1001/jama.2021.5469.

[4] Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene 2014;33:5311–8. https://doi.org/10.1038/onc.2013.533.

[5] Sciarrillo R, Wojtuszkiewicz A, Assaraf YG, Jansen G, Kaspers GJL, Giovannetti E, et al. The role of alternative splicing in cancer: From oncogenesis to drug resistance. Drug Resist Updat 2020;53:100728. https://doi.org/10.1016/j.drup.2020.100728.

[6] Annalora AJ, Marcus CB, Iversen PL. Alternative Splicing in the Cytochrome P450 Superfamily Expands Protein Diversity to Augment Gene Function and Redirect Human Drug Metabolism. Drug Metab Dispos 2017;45:375–89. https://doi.org/10.1124/dmd.116.073254.

[7] Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 2016;17:53. https://doi.org/10.1186/s13059-016-0917-0.

[8] Mokhtari RB, Homayouni TS, Baluch N, Morgatskaya E, Kumar S, Das B, et al. Combination therapy in combating cancer. Oncotarget 2017;8:38022–43. https://doi.org/10.18632/oncotarget.16723.

[9] Wang X, Zhang H, Chen X. Drug resistance and combating drug resistance in cancer. Cancer Drug Resist 2019;2:141–60. https://doi.org/10.20517/cdr.2019.10.

[10] Godskesen T, Nygren P, Nordin K, Hansson M, Kihlbom U. Phase 1 clinical trials in end-stage cancer: patient understanding of trial premises and motives for participation. Support Care Cancer 2013;21:3137–42. https://doi.org/10.1007/s00520-013-1891-7.

[11] Tator CH. Concussions and their consequences: current diagnosis, management and prevention. CMAJ Can Med Assoc J 2013;185:975–9. https://doi.org/10.1503/cmaj.120039.

[12] Silverberg ND, Iaccarino MA, Panenka WJ, Iverson GL, McCulloch KL, Dams-O'Connor K, et al. Management of Concussion and Mild Traumatic Brain Injury: A Synthesis of Practice Guidelines. Arch Phys Med Rehabil 2020;101:382–93. https://doi.org/10.1016/j.apmr.2019.10.179.

[13] Graham R, Rivara FP, Ford MA, Spicer CM, Youth C on S-RC in, Board on Children Y, et al. Concussion Recognition, Diagnosis, and Acute Management. National Academies Press (US); 2014.

[14] Bey T, Ostick B. Second Impact Syndrome. West J Emerg Med 2009;10:6–10.

[15] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform 2016:bbw065. https://doi.org/10.1093/bib/bbw065.

[16] Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 2014;32:1202–12. https://doi.org/10.1038/nbt.2877.
[17] Amin SB, Yip W-K, Minvielle S, Broyl A, Li Y, Hanlon B, et al. Gene Expression Profile Alone Is Inadequate In Predicting Complete Response In Multiple Myeloma. Leukemia 2014;28:2229–34. https://doi.org/10.1038/leu.2014.140.
[18] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2021:1–16. https://doi.org/10.1038/s41580-021-00407-0.
[19] Birzele F, Csaba G, Zimmer R. Alternative splicing and protein structure evolution. Nucleic Acids Res 2008;36:550–8. https://doi.org/10.1093/nar/gkm1054.
[20] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7. https://doi.org/10.1038/nature11003.
[21] Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013;154:1151–61. https://doi.org/10.1016/j.cell.2013.08.003.
[22] Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin 2019;69:363–85. https://doi.org/10.3322/caac.21565.
[23] Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLOS Biol 2017;15:e2003243. https://doi.org/10.1371/journal.pbio.2003243.
[24] Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. Nat Rev Mol Cell Biol 2017;18:102–14. https://doi.org/10.1038/nrm.2016.139.
[25] Tress ML, Abascal F, Valencia A. Most Alternative Isoforms Are Not Functionally Important. Trends Biochem Sci 2017;42:408–10. https://doi.org/10.1016/j.tibs.2017.04.002.
[26] Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol 2017;18:437. https://doi.org/10.1038/nrm.2017.27.
[27] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative Isoform Regulation in Human Tissue Transcriptomes. Nature 2008;456:470–6. https://doi.org/10.1038/nature07509.
[28] Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci 2011;108:11093–8. https://doi.org/10.1073/pnas.1101135108.
[29] Wu X, Hurst LD. Determinants of the Usage of Splice-Associated cis-Motifs Predict the Distribution of Human Pathogenic SNPs. Mol Biol Evol 2016;33:518–29. https://doi.org/10.1093/molbev/msv251.
[30] Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet 2016;17:19. https://doi.org/10.1038/nrg.2015.3.

[31] Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 2017;541:353–8. https://doi.org/10.1038/nature21031.
[32] López-García P, Moreira D. Eukaryogenesis, a syntrophy affair. Nat Microbiol 2019;4:1068–70. https://doi.org/10.1038/s41564-019-0495-5.
[33] Vosseberg J, Snel B. Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery. Biol Direct 2017;12:30. https://doi.org/10.1186/s13062-017-0201-6.
[34] Lambowitz AM, Zimmerly S. Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harb Perspect Biol 2011;3. https://doi.org/10.1101/cshperspect.a003616.
[35] Keating KS, Toor N, Perlman PS, Pyle AM. A structural analysis of the group II intron active site and implications for the spliceosome. RNA 2010;16:1–9. https://doi.org/10.1261/rna.1791310.
[36] Will CL, Lührmann R. Spliceosome Structure and Function. Cold Spring Harb Perspect Biol 2011;3. https://doi.org/10.1101/cshperspect.a003707.
[37] Zhang X, Zhan X, Yan C, Zhang W, Liu D, Lei J, et al. Structures of the human spliceosomes before and after release of the ligated exon. Cell Res 2019;29:274–85. https://doi.org/10.1038/s41422-019-0143-x.
[38] Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 2003;4:960–70. https://doi.org/10.1038/nrm1259.
[39] Wang Y, Liu J, Huang B, Xu Y-M, Li J, Huang L-F, et al. Mechanism of alternative splicing and its regulation (Review). Biomed Rep 2015;3:152–8.
[40] Busch A, Hertel KJ. Splicing predictions reliably classify different types of alternative splicing. RNA 2015;21:813–23. https://doi.org/10.1261/rna.048769.114.
[41] Fredericks AM, Cygan KJ, Brown BA, Fairbrother WG. RNA-Binding Proteins: Splicing Factors and Disease. Biomolecules 2015;5:893–909. https://doi.org/10.3390/biom5020893.
[42] Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, et al. SRSF1-Regulated Alternative Splicing in Breast Cancer. Mol Cell 2015;60:105–17. https://doi.org/10.1016/j.molcel.2015.09.005.
[43] Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, et al. THE SPLICING FACTOR SRSF1 REGULATES APOPTOSIS AND PROLIFERATION TO PROMOTE MAMMARY EPITHELIAL CELL TRANSFORMATION. Nat Struct Mol Biol 2012;19:220–8. https://doi.org/10.1038/nsmb.2207.
[44] Sundvall M, Veikkolainen V, Kurppa K, Salah Z, Tvorogov D, van Zoelen EJ, et al. Cell Death or Survival Promoted by Alternative Isoforms of ErbB4. Mol Biol Cell 2010;21:4275–86. https://doi.org/10.1091/mbc.E10-04-0332.
[45] Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, et al. An EMT–Driven Alternative Splicing Program Occurs in Human Breast Cancer and Modulates Cellular Phenotype. PLOS Genet 2011;7:e1002218. https://doi.org/10.1371/journal.pgen.1002218.

[46] Pradella D, Naro C, Sette C, Ghigna C. EMT and stemness: flexible processes tuned by alternative splicing in development and cancer progression. Mol Cancer 2017;16. https://doi.org/10.1186/s12943-016-0579-2.
[47] An ESRP-regulated splicing programme is abrogated during the epithelial–mesenchymal transition. EMBO J 2010;29:3286–300. https://doi.org/10.1038/emboj.2010.195.
[48] Fischer KR, Durrans A, Lee S, Sheng J, Li F, Wong STC, et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. Nature 2015;527:472–6. https://doi.org/10.1038/nature15748.
[49] Zheng X, Carstens JL, Kim J, Scheible M, Kaye J, Sugimoto H, et al. EMT Program is Dispensable for Metastasis but Induces Chemoresistance in Pancreatic Cancer. Nature 2015;527:525–30. https://doi.org/10.1038/nature16064.
[50] Saleh T, Bloukh S, Carpenter VJ, Alwohoush E, Bakeer J, Darwish S, et al. Therapy-Induced Senescence: An "Old" Friend Becomes the Enemy. Cancers 2020;12:822. https://doi.org/10.3390/cancers12040822.
[51] Kwon SM, Min S, Jeoun U, Sim MS, Jung GH, Hong SM, et al. Global spliceosome activity regulates entry into cellular senescence. FASEB J 2021;35:e21204. https://doi.org/10.1096/fj.202000395RR.
[52] Paronetto MP, Passacantilli I, Sette C. Alternative splicing and cell survival: from tissue homeostasis to disease. Cell Death Differ 2016;23:1919–29. https://doi.org/10.1038/cdd.2016.91.
[53] Mercatante DR, Bortner CD, Cidlowski JA, Kole R. Modification of Alternative Splicing of Bcl-x Pre-mRNA in Prostate and Breast Cancer Cells: ANALYSIS OF APOPTOSIS AND CELL DEATH *. J Biol Chem 2001;276:16411–7. https://doi.org/10.1074/jbc.M009256200.
[54] Robson EJD, Khaled WT, Abell K, Watson CJ. Epithelial-to-mesenchymal transition confers resistance to apoptosis in three murine mammary epithelial cell lines. Differentiation 2006;74:254–64. https://doi.org/10.1111/j.1432-0436.2006.00075.x.
[55] Rodríguez-Cruz TG, Liu S, Khalili JS, Whittington M, Zhang M, Overwijk W, et al. Natural Splice Variant of MHC Class I Cytoplasmic Tail Enhances Dendritic Cell-Induced CD8+ T-Cell Responses and Boosts Anti-Tumor Immunity. PLoS ONE 2011;6:e22939. https://doi.org/10.1371/journal.pone.0022939.
[56] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24. https://doi.org/10.1038/nature07943.
[57] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Proto-Oncogenes and Tumor-Suppressor Genes. Mol Cell Biol 4th Ed 2000.
[58] Falzone L, Salomone S, Libra M. Evolution of Cancer Pharmacological Treatments at the Turn of the Third Millennium. Front Pharmacol 2018;9:1300. https://doi.org/10.3389/fphar.2018.01300.
[59] Gambardella V, Tarazona N, Cejalvo JM, Lombardi P, Huerta M, Roselló S, et al. Personalized Medicine: Recent Progress in Cancer Therapy. Cancers 2020;12:1009. https://doi.org/10.3390/cancers12041009.
[60] Hansen AR, Bedard PL. Clinical application of high-throughput genomic technologies for treatment selection in breast cancer. Breast Cancer Res BCR 2013;15:R97. https://doi.org/10.1186/bcr3558.

[61] Arruebo M, Vilaboa N, Sáez-Gutierrez B, Lambea J, Tres A, Valladares M, et al. Assessment of the Evolution of Cancer Treatment Therapies. Cancers 2011;3:3279–330. https://doi.org/10.3390/cancers3033279.
[62] Pucci C, Martinelli C, Ciofani G. Innovative approaches for cancer treatment: current perspectives and new challenges. Ecancermedicalscience 2019;13:961. https://doi.org/10.3332/ecancer.2019.961.
[63] Noonan KL, Ho C, Laskin J, Murray N. The Influence of the Evolution of First-Line Chemotherapy on Steadily Improving Survival in Advanced Non-Small-Cell Lung Cancer Clinical Trials. J Thorac Oncol 2015;10:1523–31. https://doi.org/10.1097/JTO.0000000000000667.
[64] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev 2017;9:91–102. https://doi.org/10.1007/s12551-016-0247-1.
[65] Hawkins PCD, Skillman AG, Nicholls A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. J Med Chem 2007;50:74–82. https://doi.org/10.1021/jm0603365.
[66] Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nat Commun 2019;10:1197. https://doi.org/10.1038/s41467-019-09186-x.
[67] Mullard A. $1.3 billion per drug? Nat Rev Drug Discov 2020;19:226–226. https://doi.org/10.1038/d41573-020-00043-x.
[68] Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. Alzheimers Dement Transl Res Clin Interv 2017;3:651–7. https://doi.org/10.1016/j.trci.2017.10.005.
[69] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3:673–83. https://doi.org/10.1038/nrd1468.
[70] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. Science 1991;252:1651–6. https://doi.org/10.1126/science.2047873.
[71] Motameny S, Wolters S, Nürnberg P, Schumacher B. Next Generation Sequencing of miRNAs – Strategies, Resources and Methods. Genes 2010;1:70–84. https://doi.org/10.3390/genes1010070.
[72] Wang H-LV, Chekanova JA. An Overview of Methodologies in Studying lncRNAs in the High-Throughput Era: When Acronyms ATTACK! Methods Mol Biol Clifton NJ 2019;1933:1–30. https://doi.org/10.1007/978-1-4939-9045-0_1.
[73] Reuter JA, Spacek D, Snyder MP. High-Throughput Sequencing Technologies. Mol Cell 2015;58:586–97. https://doi.org/10.1016/j.molcel.2015.05.004.
[74] Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: Overviews and challenges. BioTechniques 2014;56:61–77. https://doi.org/10.2144/000114133.
[75] Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, et al. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. J Biol Chem 2019;294:18220–31. https://doi.org/10.1074/jbc.RA119.010676.

[76] Xu L, Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer. J Hum Genet 2020;65:25–33. https://doi.org/10.1038/s10038-019-0679-0.
[77] Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun 2019;10:3359. https://doi.org/10.1038/s41467-019-11272-z.
[78] Zascavage RR, Thorson K, Planz JV. Nanopore sequencing: An enrichment□free alternative to mitochondrial DNA sequencing. Electrophoresis 2019;40:272–80. https://doi.org/10.1002/elps.201800083.
[79] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010;38:1767–71. https://doi.org/10.1093/nar/gkp1137.
[80] Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. National Center for Biotechnology Information n.d. https://www.ncbi.nlm.nih.gov/ (accessed October 12, 2021).
[81] The European Bioinformatics Institute < EMBL-EBI n.d. https://www.ebi.ac.uk/ (accessed October 12, 2021).
[82] Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, et al. The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 2018;46:D762–9. https://doi.org/10.1093/nar/gkx1020.
[83] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio 2013.
[84] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
[85] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.
[86] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4.
[87] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18:1851–8. https://doi.org/10.1101/gr.078212.108.
[88] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.
[89] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 2011;8:469–77. https://doi.org/10.1038/nmeth.1613.
[90] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28:511–5. https://doi.org/10.1038/nbt.1621.

[91] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 2010;7:1009–15. https://doi.org/10.1038/nmeth.1528.
[92] Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A 2014;111:E5593–601. https://doi.org/10.1073/pnas.1419161111.
[93] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
[94] What the FPKM? A review of RNA-Seq expression units. The Farrago 2014. https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/ (accessed September 13, 2021).
[95] de Jong TV, Moshkin YM, Guryev V. Gene expression variability: the other dimension in transcriptome analysis. Physiol Genomics 2019;51:145–58. https://doi.org/10.1152/physiolgenomics.00128.2018.
[96] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8.
[97] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.
[98] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881–7. https://doi.org/10.1093/bioinformatics/btm453.
[99] Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008;9:321–32. https://doi.org/10.1093/biostatistics/kxm030.
[100] Chen Y, Lun A, Smyth G. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR, 2014. https://doi.org/10.1007/978-3-319-07212-8_3.
[101] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40:4288–97. https://doi.org/10.1093/nar/gks042.
[102] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106. https://doi.org/10.1186/gb-2010-11-10-r106.
[103] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013;31:46–53. https://doi.org/10.1038/nbt.2450.
[104] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 2010;28:503–10. https://doi.org/10.1038/nbt.1633.
[105] Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, et al. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics 2015;31:2400–2. https://doi.org/10.1093/bioinformatics/btv034.

[106] Zou H, Hastie T. Regularization and variable selection via the Elastic Net. J R Stat Soc Ser B 2005;67:301–20.
[107] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[108] Liaw A, Wiener M. Classification and Regression by RandomForest. ResearchGate 2001;23.
[109] Velliangiri S, Alagumuthukrishnan S, Thankumar joseph SI. A Review of Dimensionality Reduction Techniques for Efficient Computation. Procedia Comput Sci 2019;165:104–11. https://doi.org/10.1016/j.procs.2020.01.079.
[110] Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLOS Comput Biol 2019;15:e1006907. https://doi.org/10.1371/journal.pcbi.1006907.
[111] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 1901;2:559–72. https://doi.org/10.1080/14786440109462720.
[112] Rosenthal S. Data Imputation. Int. Encycl. Commun. Res. Methods, American Cancer Society; 2017, p. 1–12. https://doi.org/10.1002/9781118901731.iecrm0058.
[113] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33:159–74. https://doi.org/10.2307/2529310.
[114] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13. https://doi.org/10.1038/nrg1272.
[115] Farine DR, Whitehead H. Constructing, conducting and interpreting animal social network analysis. J Anim Ecol 2015;84:1144–63. https://doi.org/10.1111/1365-2656.12418.
[116] Emmons S, Kobourov S, Gallant M, Börner K. Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. PLOS ONE 2016;11:e0159161. https://doi.org/10.1371/journal.pone.0159161.
[117] Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, et al. Assessment of network module identification across complex diseases. Nat Methods 2019;16:843–52. https://doi.org/10.1038/s41592-019-0509-5.
[118] Lancichinetti A, Radicchi F, Ramasco JJ. Statistical significance of communities in networks. Phys Rev E 2010;81:046110. https://doi.org/10.1103/PhysRevE.81.046110.
[119] World Health Organization(WHO). The ICD-10 classification of mental and behavioural disorders. Genève, Switzerland: World Health Organization; 1993.
[120] DePadilla L. Self-Reported Concussions from Playing a Sport or Being Physically Active Among High School Students — United States, 2017. MMWR Morb Mortal Wkly Rep 2018;67. https://doi.org/10.15585/mmwr.mm6724a3.
[121] Helmick KM, Spells CA, Malik SZ, Davies CA, Marion DW, Hinds SR. Traumatic brain injury in the US military: epidemiology and key clinical and research programs. Brain Imaging Behav 2015;9:358–66. https://doi.org/10.1007/s11682-015-9399-z.
[122] Centers for Disease Control and Prevention. Surveillance Report of Traumatic Brain Injury-related Hospitalizations and Deaths by Age Group, Sex, and Mechanism of Injury-United States, 2016 and 2017 2021.

[123] McCrea M, Hammeke T, Olsen G, Leo P, Guskiewicz K. Unreported Concussion in High School Football Players: Implications for Prevention. Clin J Sport Med 2004;14:13–7.
[124] McCrea M, Broglio SP, McAllister TW, Gill J, Giza CC, Huber DL, et al. Association of Blood Biomarkers With Acute Sport-Related Concussion in Collegiate Athletes: Findings From the NCAA and Department of Defense CARE Consortium. JAMA Netw Open 2020;3:e1919771–e1919771. https://doi.org/10.1001/jamanetworkopen.2019.19771.
[125] Broglio SP, McCrea M, McAllister T, Harezlak J, Katz B, Hack D, et al. A National Study on the Effects of Concussion in Collegiate Athletes and US Military Service Academy Members: The NCAA–DoD Concussion Assessment, Research and Education (CARE) Consortium Structure and Methods. Sports Med Auckl Nz 2017;47:1437–51. https://doi.org/10.1007/s40279-017-0707-1.
[126] Care Consortium n.d. http://www.careconsortium.net/ (accessed October 12, 2021).
[127] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30. https://doi.org/10.1093/bioinformatics/btt656.
[128] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 1995;57:289–300.
[129] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J Integr Biol 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
[130] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21:3439–40. https://doi.org/10.1093/bioinformatics/bti525.
[131] Durinck S, Spellman PT, Birney E, Huber W. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009;4:1184–91. https://doi.org/10.1038/nprot.2009.97.
[132] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[133] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40. https://doi.org/10.1093/bioinformatics/btr260.
[134] Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell-type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37:773–82. https://doi.org/10.1038/s41587-019-0114-2.
[135] Giza CC, Hovda DA. The New Neurometabolic Cascade of Concussion. Neurosurgery 2014;75:S24–33. https://doi.org/10.1227/NEU.0000000000000505.
[136] Flanagan JL, Simmons PA, Vehige J, Willcox MD, Garrett Q. Role of carnitine in disease. Nutr Metab 2010;7:30. https://doi.org/10.1186/1743-7075-7-30.

[137] Bazarian JJ, Biberthaler P, Welch RD, Lewis LM, Barzo P, Bogner-Flatz V, et al. Serum GFAP and UCH-L1 for prediction of absence of intracranial injuries on head CT (ALERT-TBI): a multicentre observational study. Lancet Neurol 2018;17:782–9. https://doi.org/10.1016/S1474-4422(18)30231-X.
[138] Blaylock RL, Maroon J. Immunoexcitotoxicity as a central mechanism in chronic traumatic encephalopathy—A unifying hypothesis. Surg Neurol Int 2011;2:107. https://doi.org/10.4103/2152-7806.83391.
[139] Hinson HE, Rowell S, Schreiber M. Clinical evidence of inflammation driving secondary brain injury: A systematic review. J Trauma Acute Care Surg 2015;78:184–91. https://doi.org/10.1097/TA.0000000000000468.
[140] Livingston WS, Gill JM, Cota MR, Olivera A, O'Keefe JL, Martin C, et al. Differential Gene Expression Associated with Meningeal Injury in Acute Mild Traumatic Brain Injury. J Neurotrauma 2017;34:853–60. https://doi.org/10.1089/neu.2016.4479.
[141] Israelsson C, Bengtsson H, Kylberg A, Kullander K, Lewén A, Hillered L, et al. Distinct Cellular Patterns of Upregulated Chemokine Expression Supporting a Prominent Inflammatory Role in Traumatic Brain Injury: [1]. J Neurotrauma 2008;25:959–74. http://dx.doi.org.proxy.ulib.uits.iu.edu/10.1089/neu.2008.0562.
[142] Khuman J, Meehan WP, Zhu X, Qiu J, Hoffmann U, Zhang J, et al. Tumor necrosis factor alpha and Fas receptor contribute to cognitive deficits independent of cell death after concussive traumatic brain injury in mice. J Cereb Blood Flow Metab 2011;31:778–89. https://doi.org/10.1038/jcbfm.2010.172.
[143] Patterson ZR, Holahan MR. Understanding the neuroinflammatory response following concussion to develop treatment strategies. Front Cell Neurosci 2012;6:58. https://doi.org/10.3389/fncel.2012.00058.
[144] Di Battista AP, Rhind SG, Richards D, Hutchison MG. An investigation of plasma interleukin-6 in sport-related concussion. PLoS ONE 2020;15:e0232053. https://doi.org/10.1371/journal.pone.0232053.
[145] Ciccone M, Calin GA, Perrotti D. From the Biology of PP2A to the PADs for Therapy of Hematologic Malignancies. Front Oncol 2015;5:21. https://doi.org/10.3389/fonc.2015.00021.
[146] Pico AR, Kelder T, Iersel MP van, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway Editing for the People. PLOS Biol 2008;6:e184. https://doi.org/10.1371/journal.pbio.0060184.
[147] Nishimura D. BioCarta. Biotech Softw Internet Rep 2001;2:117–20. https://doi.org/10.1089/152791601750294344.
[148] Petrone AB, Gionis V, Giersch R, Barr TL. Immune biomarkers for the diagnosis of mild traumatic brain injury. NeuroRehabilitation 2017;40:501–8. https://doi.org/10.3233/NRE-171437.
[149] Merchant-Borna K, Lee H, Wang D, Bogner V, van Griensven M, Gill J, et al. Genome-Wide Changes in Peripheral Gene Expression following Sports-Related Concussion. J Neurotrauma 2016;33:1576–85. https://doi.org/10.1089/neu.2015.4191.
[150] Farina AR, Cappabianca L, Sebastiano M, Zelli V, Guadagni S, Mackay AR. Hypoxia-induced alternative splicing: the 11th Hallmark of Cancer. J Exp Clin Cancer Res 2020;39:110. https://doi.org/10.1186/s13046-020-01616-9.

[151] Dehm SM. mRNA Splicing Variants: Exploiting Modularity to Outwit Cancer Therapy. Cancer Res 2013;73:5309–14. https://doi.org/10.1158/0008-5472.CAN-13-0444.
[152] Zammarchi F, Stanchina E de, Bournazou E, Supakorndej T, Martires K, Riedel E, et al. Antitumorigenic potential of STAT3 alternative splicing modulation. Proc Natl Acad Sci 2011;108:17779–84. https://doi.org/10.1073/pnas.1108482108.
[153] Hernandez-Lopez HR, Graham SV. Alternative splicing in human tumour viruses: a therapeutic target? Biochem J 2012;445:145–56. https://doi.org/10.1042/BJ20120413.
[154] Pawellek A, McElroy S, Samatov T, Mitchell L, Woodland A, Ryder U, et al. Identification of Small Molecule Inhibitors of Pre-mRNA Splicing. J Biol Chem 2014;289:34683–98. https://doi.org/10.1074/jbc.M114.590976.
[155] Bauman JA, Kole R. Modulation of RNA splicing as a potential treatment for cancer. Bioeng Bugs 2011;2:125–8. https://doi.org/10.4161/bbug.2.3.15165.
[156] Niedermeier M, Hennessy BT, Knight ZA, Henneberg M, Hu J, Kurtova AV, et al. Isoform-selective phosphoinositide 3′-kinase inhibitors inhibit CXCR4 signaling and overcome stromal cell–mediated drug resistance in chronic lymphocytic leukemia: a novel therapeutic approach. Blood 2009;113:5549–57. https://doi.org/10.1182/blood-2008-06-165068.
[157] Cesi G, Philippidou D, Kozar I, Kim YJ, Bernardin F, Van Niel G, et al. A new ALK isoform transported by extracellular vesicles confers drug resistance to melanoma cells. Mol Cancer 2018;17. https://doi.org/10.1186/s12943-018-0886-x.
[158] Peng H, Peng T, Wen J, Engler DA, Matsunami RK, Su J, et al. Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. Bioinformatics 2014;30:1899–907. https://doi.org/10.1093/bioinformatics/btu133.
[159] Ashley EA. Towards precision medicine. Nat Rev Genet 2016;17:507–22. https://doi.org/10.1038/nrg.2016.86.
[160] Ogilvie LA, Wierling C, Kessler T, Lehrach H, Lange BMH. Predictive Modeling of Drug Treatment in the Area of Personalized Medicine. Cancer Inform 2015;14:95–103. https://doi.org/10.4137/CIN.S1933.
[161] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
[162] Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform 2007;2:59–77.
[163] Vickers AJ. Prediction models in cancer care. CA Cancer J Clin 2011;61:315–26. https://doi.org/10.3322/caac.20118.
[164] Shen S, Wang Y, Wang C, Wu YN, Xing Y. SURVIV for survival analysis of mRNA isoform variation. Nat Commun 2016;7:11548. https://doi.org/10.1038/ncomms11548.
[165] Safikhani Z, Smirnov P, Thu KL, Silvester J, El-Hachem N, Quevedo R, et al. Gene isoforms as expression-based biomarkers predictive of drug response in vitro. Nat Commun 2017;8:1126. https://doi.org/10.1038/s41467-017-01153-8.

[166] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22.
[167] Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training. 2016.
[168] Cox J, Weinman S. Mechanisms of doxorubicin resistance in hepatocellular carcinoma. Hepatic Oncol 2016;3:57–9. https://doi.org/10.2217/hep.15.41.
[169] Eliaz RE, Nir S, Marty C, Szoka FC. Determination and Modeling of Kinetics of Cancer Cell Killing by Doxorubicin and Doxorubicin Encapsulated in Targeted Liposomes. Cancer Res 2004;64:711–8. https://doi.org/10.1158/0008-5472.CAN-03-0654.
[170] Minderman H, Linssen PC, Wessels JM, Haanen C. Doxorubicin toxicity in relation to the proliferative state of human hematopoietic cells. Exp Hematol 1991;19:110–4.
[171] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740–54. https://doi.org/10.1016/j.cell.2016.06.017.
[172] Pozdeyev N, Yoo M, Mackie R, Schweppe RE, Tan AC, Haugen BR. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. Oncotarget 2016;7:51619–25. https://doi.org/10.18632/oncotarget.10010.
[173] Gillet J-P, Varma S, Gottesman MM. The Clinical Relevance of Cancer Cell Lines. JNCI J Natl Cancer Inst 2013;105:452–8. https://doi.org/10.1093/jnci/djt007.
[174] The Cancer Cell Line Encyclopedia Consortium, The Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature 2015;528:84–7. https://doi.org/10.1038/nature15736.
[175] Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, et al. Prediction and Quantification of Splice Events from RNA-Seq Data. PLOS ONE 2016;11:e0156132. https://doi.org/10.1371/journal.pone.0156132.
[176] Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 2017;18:583. https://doi.org/10.1186/s12864-017-4002-1.
[177] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
[178] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8. https://doi.org/10.1093/bioinformatics/btr064.
[179] Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013;499:172. https://doi.org/10.1038/nature12311.
[180] González-Mariscal L, Lechuga S, Garay E. Role of tight junctions in cell proliferation and cancer. Prog Histochem Cytochem 2007;42:1–57. https://doi.org/10.1016/j.proghi.2007.01.001.
[181] Martin TA, Jiang WG. Loss of tight junction barrier function and its role in cancer metastasis. Biochim Biophys Acta BBA - Biomembr 2009;1788:872–91. https://doi.org/10.1016/j.bbamem.2008.11.005.

[182] Tsukita S, Yamazaki Y, Katsuno T, Tamura A, Tsukita S. Tight junction-based epithelial microenvironment and cell proliferation. Oncogene 2008;27:6930. https://doi.org/10.1038/onc.2008.344.
[183] Provenzano PP, Keely PJ. Mechanical signaling through the cytoskeleton regulates cell proliferation by coordinated focal adhesion and Rho GTPase signaling. J Cell Sci 2011;124:1195–205. https://doi.org/10.1242/jcs.067009.
[184] Parri M, Chiarugi P. Rac and Rho GTPases in cancer cell motility control. Cell Commun Signal CCS 2010;8:23. https://doi.org/10.1186/1478-811X-8-23.
[185] Rhyu MS, Jan LY, Jan YN. Asymmetric distribution of numb protein during division of the sensory organ precursor cell confers distinct fates to daughter cells. Cell 1994;76:477–91. https://doi.org/10.1016/0092-8674(94)90112-0.
[186] Zhao C, Chen A, Jamieson CH, Fereshteh M, Abrahamsson A, Blum J, et al. Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukaemia. Nature 2009;458:776–9. https://doi.org/10.1038/nature07737.
[187] Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 2016;16:413. https://doi.org/10.1038/nrc.2016.51.
[188] Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation. Mol Cell 2013;52:720–33. https://doi.org/10.1016/j.molcel.2013.11.010.
[189] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.
[190] Wang K, Ubriaco G, Sutherland LC. RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours. BMC Genomics 2007;8:348. https://doi.org/10.1186/1471-2164-8-348.
[191] Saint-Geniez M, Jiang A, Abend S, Liu L, Sweigard H, Connor KM, et al. PGC-1α Regulates Normal and Pathological Angiogenesis in the Retina. Am J Pathol 2013;182:255–65. https://doi.org/10.1016/j.ajpath.2012.09.003.
[192] UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45:D158–69. https://doi.org/10.1093/nar/gkw1099.
[193] Makeyev AV, Liebhaber SA. The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. RNA 2002;8:265–78.
[194] Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, et al. RBFOX2 Is an Important Regulator of Mesenchymal Tissue-Specific Splicing in both Normal and Cancer Tissues. Mol Cell Biol 2013;33:396–405. https://doi.org/10.1128/MCB.01174-12.
[195] Wen J, Toomer KH, Chen Z, Cai X. Genome-wide analysis of alternative transcripts in human breast cancer. Breast Cancer Res Treat 2015;151:295–307. https://doi.org/10.1007/s10549-015-3395-2.
[196] Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KCH, et al. Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. Mol Cell 2017;65:539-553.e7. https://doi.org/10.1016/j.molcel.2017.01.011.

[197] Seguin L, Kato S, Franovic A, Camargo MF, Lesperance J, Elliott KC, et al. A β3 integrin-KRAS-RalB complex drives tumor stemness and resistance to EGFR inhibition. Nat Cell Biol 2014;16:457–68. https://doi.org/10.1038/ncb2953.
[198] Ma J li, Zeng S, Zhang Y, Deng G lu, Shen H. Epithelial–mesenchymal transition plays a critical role in drug resistance of hepatocellular carcinoma cells to oxaliplatin. Tumor Biol 2016;37:6177–84. https://doi.org/10.1007/s13277-015-4458-z.
[199] Shang Y, Fan XC and D. Roles of Epithelial-Mesenchymal Transition in Cancer Drug Resistance. Curr Cancer Drug Targets 2013;13:915–29. https://doi.org/10.2174/15680096113136660097.
[200] Salt MB, Bandyopadhyay S, McCormick F. Epithelial-to-Mesenchymal Transition Rewires the Molecular Path to PI3K-Dependent Proliferation. Cancer Discov 2014;4:186–99. https://doi.org/10.1158/2159-8290.CD-13-0520.
[201] Ye X, Weinberg RA. Epithelial–Mesenchymal Plasticity: A Central Regulator of Cancer Progression. Trends Cell Biol 2015;25:675–86. https://doi.org/10.1016/j.tcb.2015.07.012.
[202] David CJ, Chen M, Assanah M, Canoll P, Manley JL. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature 2010;463:364–8. https://doi.org/10.1038/nature08697.
[203] Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 2021;49:D458–60. https://doi.org/10.1093/nar/gkaa937.
[204] Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 2018;46:D493–6. https://doi.org/10.1093/nar/gkx922.
[205] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinforma 2016;54:1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5.
[206] Teriete P, Banerji S, Noble M, Blundell CD, Wright AJ, Pickford AR, et al. Structure of the Regulatory Hyaluronan Binding Domain in the Inflammatory Leukocyte Homing Receptor CD44. Mol Cell 2004;13:483–96. https://doi.org/10.1016/S1097-2765(04)00080-2.

Curriculum Vitae

Edward Ronald Simpson Jr.

Education

Ph.D. Informatics
Indiana University (January, 2022)
Indiana University-Purdue University Indianapolis, Indianapolis, IN
M.S. Bioinformatics
Indiana University (May, 2016)
Indiana University-Purdue University Indianapolis, Indianapolis, IN
B.S. Genetics & Microbiology
Purdue University (May, 2008)
Purdue University, West Lafayette, IN

Research and Work Experience

October ’18 – Present	Research Associate Center for Medical Genomics IU School of Medicine High-throughput Sequencing and Analysis
January ’15 – January ’22	Ph.D. Candidate, Analyst Center for Computational Biology and Bioinformatics IU School of Medicine Downstream and Advanced Analysis of Next-Generation Sequencing Data
June ’15 – December ’15	Graduate Research Assistant IU Department of Obstetrics and Gynecology Physiologically Based Pharmacokinetic Modeling
May ’14 - July ’16	Laboratory Technician IU Molecular Genetics Diagnostic Laboratory Testing and Analysis for Human Genetic Disorders
December ’09 – May ’14	Laboratory Technician Indiana State Department of Health Research, Design and Implement Molecular Assays
September ’08 – December ’09	Laboratory Technician A-tek, Biowatch Program Environmental Testing

March ’07 – July ’08	Undergraduate Researcher Dr. Richard Kuhn, Purdue University Develop Reagents for Viral Research

Awards and Honors

Institutional Candidate for the PhRMA foundation pre-doctoral fellowship	(2017)
Nominated for Sherry Queener Graduate Student Excellence Award	(2016)

Conference Posters and Talks

Concussion results in immediate gene expression changes that trigger signaling pathways related to injury and immune system response in peripheral blood samples: Findings from the NCAA-DOD CARE Consortium
Military Health System Research Symposium (Kissimmee, FL 2021), Accepted Talk

Alternative mRNA Splicing-based Drug Response Networks Yield Interactive and Mechanistic Insights
Complex Networks (Lisbon, Portugal 2019), Poster

Basal Splicing Profiles Predict Doxorubicin Response in Cancer Cell Lines
American Association of Cancer Research (Chicago, IL 2018), Poster

Publications

[in preparation] E Simpson, K N H Nudelman, Jie Ren, J Harezlak, J L Reiter, T M Foroud, A Saykin, S Broglio, M A McCrea, P F Pasquina, T W McAllister, Y Liu. RNAseq Analysis of Concussed Collegiate Athletes Reveals Activation of Immune Signaling Processes: A Retrospective Cohort Study by the NCAA-DOD CARE Consortium.

[in press] E Simpson, S Chen, J Reiter, Y Liu. Differential splicing of skipped-exons predicts drug response in cancer cell lines. Genomics, Proteomics & Bioinformatics. doi:10.1016/j.gpb.2019.08.003.

TB Ladd, JA Johnson Jr, CL Mumaw, HJ Greve, X Xuei, E Simpson, MA Barnes, BJ Green, TL Croston, C Ahmed. Aspergillus versicolor Inhalation Triggers Neuroimmune, Glial, and Neuropeptide Transcriptional Changes. ASN Neuro. Jan-Dec 2021; 13:17590914211019886. doi:10.1177/17590914211019886.

AM Patterson, PA Plett, CH Sampson, E Simpson,Y Liu, LM Pelus, C Orschell. Prostaglandin E 2 Enhances Aged Hematopoietic Stem Cell Function. Stem Cell Rev Rep. 2021; May 11. doi:10.1007/s12015-021-10177-z.

GG Grecco, DL Haggerty, EH Doud, BM Fritz, F Yin, H Hoffman, AL Mosley, E Simpson, Y Liu, AJJ Baucum. A multi□omic analysis of the dorsal striatum in an animal model of divergent genetic risk for alcohol use disorder. Neurochem. 2020. doi:10.1111/jnc.15226.

X Chen, Y Liu, C Xu, L Ba, Z Liu, X Li, J Huang, E Simpson, H Gao, D Cao. QKI is a critical pre-mRNA alternative splicing regulator of cardiac myofibrillogenesis and contractile function. Nature Commun. 2021; 12:89. doi:10.1038/s41467-020-20327-5.

R L Konger, E Derr-Yellin, TA Zimmers, T Katona, X Xuei, Y Liu, HM Zhou, E Simpson, MJ Turner. Epidermal PPARγ Is a Key Homeostatic Regulator of Cutaneous Inflammation and Barrier Function in Mouse Skin. Int J Mol Sci. 2021; Aug 11;22(16):8634. doi:10.3390/ijms22168634.

MD Davis, TM Clemente, OK Giddings, K Ross, RS Cunningham, L Smith, E Simpson, Y Liu, K Kloepfer, IS Ramsey. A Treatment to Eliminate SARS-CoV-2 Replication in Human Airway Epithelial Cells Is Safe for Inhalation as an Aerosol in Healthy Human Subjects. J Neurochem, Oct 2020. doi:10.1111/jnc.15226.

SA Peck Justice, MP Barron, GD Qi, HRS Wijeratne, JF Victorino, E Simpson, JZ Vilseck, AB Wijeratne, AL Mosley, Amber L. Mutant thermal proteome profiling for characterization of missense protein variants and their associated phenotypes within the proteome. J Biol Chem. 2020; Nov 27;295(48):16219-16238. doi:10.1074/jbc.RA120.014576.

PB Nakshatri, B Kumar, E Simpson, KK Ludwig, ML Cox, H Gao, Y Liu, H Nakshatri. Breast Cancer Cell Detection and Characterization from Breast Milk–Derived Cells. Cancer Res. 2020; 80 (21):4828-4839. doi:10.1158/0008-5472.CAN-20-1030.

N Marino, R German, X Rao, E Simpson, S Liu, J Wan, Y Liu, G Sandusky, M Jacobsen, M Stoval. Upregulation of lipid metabolism genes in the breast prior to cancer diagnosis. NPJ Breast Cancer. 2020; 6:50. doi:10.1038/s41523-020-00191-8.

S Krishnan, RS Stearman, L Zeng, A Fisher, EA Mickler, BH Rodriguez, ER Simpson, T Cook, JE Slaven, M Ivan. Transcriptomic modifications in developmental cardiopulmonary adaptations to chronic hypoxia using a murine model of simulated high-altitude exposure. Am J Physiol Lung Cell Mol Physiol 2020; Sep 1;319(3):L456-L470.

W Wu, F Syed, E Simpson, CC Lee, DL Eizirik, R Mirmira, Y Liu, C Evans-Molina. 81-OR: The Impact of Proinflammatory Cytokines on Human Pancreatic Islet Alternative Splicing Patterns. Diabetes 2020 Jun; 69(S1). doi:10.2337/db20-81-OR.

S Marino, DN Petrusca, E Simpson, Edward, JL Anderson, XQ Xie, Y Liu, J Chirgwin, GD Roodman. Targeting the p62-ZZ/N-End Rule Pathway in Multiple Myeloma Overcomes Proteasome Inhibitor-Resistance Via Induction of Necroptosis and Enhances

the Bone Anabolic Effects of Proteasome Inhibitors. Blood 2019; 134 (S1): 4391. doi:10.1182/blood-2019-131865.

Y Wang, B He, Y Zhao, JL Reiter, SX Chen, E Simpson, W Feng, Y Liu. Comprehensive Cis-Regulation Analysis of Genetic Variants in Human Lymphoblastoid Cell Lines. Front. Genet., 10 September 2019. doi:10.3389/fgene.2019.00806.

M Anjanappa, Y Hao, E Simpson, P Bhat-Nakshatri, JB Nelson, SA Tersey, RG Mirmira, AA Cohen-Gadol, MR Saadatzadeh, L Li, F Fang, KP Nephew, KD Miller, Y Liu, H Nakshatri. A System for Detecting High Impact-low Frequency Mutations in Primary Tumors and Metastases. Oncogene Jan 2018; 37, 185-196. doi:10.1038/onc.2017.322.

PBS Celestino-Soper, ESimpson, D Tumbleson Brink, TC Lynnes, S Dlouhy, M Vatta, J Yeley, C Brown, S Bai. Intragenic CFTR Duplication and 5T/12TG Variant in a Patient with Non-Classic Cystic Fibrosis. Scientific Reports 2016; 6, 38776. doi:10.1038/srep38776.

RF Relich, RM Humphries, HR Mattison, JE Miles, E Simpson, IJ Corbett, BH Schmitt, M May. Francisella philomiragia Bacteremia in a Patient with Acute Respiratory Insufficiency and Acute-on-Chronic Kidney Disease. Journal of Clinical Microbiology Dec 2015; 53,3947. doi:10.1128/JCM.01762-15.