NETWORK-BASED ALTERNATIVE SPLICING SIGNATURES OF DRUG RESPONSE IN AML

by

Julian A. Egger

A DISSERTATION

Presented to the
Deparment of Medical Informatics and Clincial Epidemiology
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Bioinformatics and Computational Biomedicine

December 2020

I dedicate this dissertation to my wife Angela for all her love and support during my PhD candidacy.

Abstract

Background Contributing to proteomic diversity, alternative splicing of pre-mRNA is widespread in the human transcriptome and can greatly influence regulation of both normal and disease-related cellular phenotypes. Similar to gene expression, alternative splicing does not occur independently, but in a coordinated fashion throughout the transcriptome in order to maintain proper cellular function. Gene co-expression networks have been widely used as an approach to elucidate coordinated regulatory patterns of gene transcription. Studies have shown that genome-wide expression can occur in the form of network modules consisting of highly co-expressed genes operating within specific cellular pathways. Such modules are often well-preserved across similar biological systems and associated with various phenotypes. Here we demonstrate a framework for de novo network inference of co-splicing in the form of modules consisting of complex alternative splicing variants.

Results Network inference methods can be used to characterize coordination of complex alternative splicing variants in the form of co-splicing modules. We utilize graph-based splicing quantification methods to annotate and quantify complex splicing variants from short read RNA-sequencing (RNA-seq) data and formulate them in a way suitable for a module-based network approach. This framework allows us to identify groups of complex splicing variants who undergo coordinated regulation and are statistically associated with various phenotypes. We applied this framework on various tissue types from the Genotype-Tissue Expression (GTEx) project and identified a set of consensus modules consisting of complex splicing variants highly co-spliced across tissue types. Consensus tissue modules also exhibit module-level splicing values that are highly tissue-specific. We then applied our framework to infer a co-splicing network of acute myeloid leukemia and identified co-splicing modules strongly correlated with drug response for multiple targeted therapies.

Conclusions Our proposed framework for de novo network inference of co-splicing can help characterize transcriptome-wide coordination of complex splicing variation in various biological systems and identify groups of splicing variants operating within functional pathways. Our module-based approach can be further applied to other RNA-seq datasets to identify groups of complex splice variants that are both highly co-spliced and associated with various phenotypes of interest.

List of Figures

Contents

Chapter 1: Introduction

Chapter 1 serves as an introductory review for the topics addressed in this dissertation. First, we review the underlying biology of alternative splicing in basic cellular regulation along with the role of aberrant splicing in human disease. We then discuss the current state of technology for characterizing and quantifying alternative splicing, namely using RNA-sequencing. Finally, we provide an overview of de nevo network inference methods typically used for the study of gene co-expression.

RNA Splicing in Biology & Disease

Alternative Splicing Ribonucleic acid (RNA) splicing is a post-transcriptional process in which one or more genomic regions of an expressed gene are excised out before translation can occur. Known as introns, these excised regions are removed by a dynamic cellular protein complex called the spliceosome (B. D. Wang and Lee 2018; Zhou and Chng 2017). Composed of multiple proteins and small RNA, the spliceosome links together the remaining non-excised pre-messenger RNA (pre-mRNA) regions, known as exons, to form a processed mRNA transcript now capable of being translated into a functional protein product (Zhou and Chng 2017; Kim et al. 2018). Further, the specific selection of introns removed from the transcribed pre-mRNA during splicing can occur in different combinations. This can in turn lead to the production of multiple, distinct mature RNA products from a single expressed gene (Figure 1). This phenomenon is known as alternative splicing and the most recent of whole-transcriptome sequencing (RNA-Seq) studies show that over 90% of multi-exon human genes produce alternatively spliced transcripts (Necochea-Campion et al. 2016; B. D. Wang and Lee 2018; Zhou and Chng 2017).

The ability for a single gene to produce multiple transcript products is an obvious mechanism for protein diversity given that the number of cataloged protein products

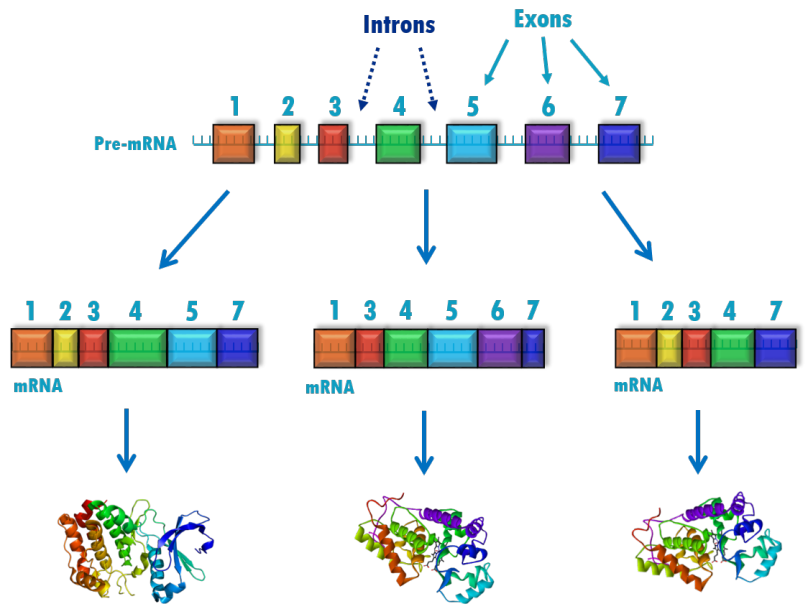

Figure 1: Diagram of Alternative Splicing. During gene transcription, introns are removed from the precursor RNA and the remaining exons are stiched together by the splicing machinery. The specific selection of exons included in the processed transcript may vary, leading to the production of alternatively spliced transcripts.

is far greater than the number of protein coding genes [Friedman, Hastie, and Tibshirani (2008). While many alternatively spliced transcripts do result in protein products with different functionality, other transcripts contribute specifically to regulation. In some cases, alternatively spliced transcripts do not encode a final protein product at all and are often targeted for nonsense-mediated decay (NMD). The relative expression of each alternative transcript is often highly specific and differential splicing levels can be observed across different cell types. Thus, along with gene expression, alternative splicing plays a key role in regulating cellular phenotypes (e.g., tissue specificity) and driving fundamental cellular programs (e.g., cellular differentiation) (Friedman, Hastie, and Tibshirani 2008).

Aberrant Splicing in Cancer Alternative splicing of pre-mRNA is a tightly regulated process and similar to gene regulation, splicing does not occur independently, but rather in a coordinated fashion throughout the transcriptome in

order to drive proper cellular function (Fagnani et al. 2007; Dai et al. 2012). Loss of such control can lead to aberrant splicing events either by deregulating the expression of normal splice variants or mis-splicing of pre-mRNA by the splicing machinery (B. D. Wang and Lee 2018). As such, alternative splicing has been demonstrated as a key transcriptional mechanism contributing to the formation of a variety of human diseases (Kim et al. 2018).

Aberrant splicing dynamics can occur alongside deregulated gene expression as an additional transcriptional mechanism for cells to develop cancer-related cellular phenotypes (Necochea-Campion et al. 2016; B. D. Wang and Lee 2018; Kim et al. 2018). Pan-cancer transcriptome studies have identified numerous cancer-specific alternative splicing variants absent in healthy tissues and numerous cancer splicing variants are specific to different cancer types and subgroups (B. D. Wang and Lee 2018). Cancer-specific splicing variants can occur in known oncogenes and tumor suppressor genes such as HRAS and KLF65. Deregulated splicing can also contribute to some of the key hallmarks of cancer described by (Hanahan 2000) including sustained proliferative signaling, induced angiogenesis, metastasis and lack of apoptosis (B. D. Wang and Lee 2018; Kim et al. 2018; Park et al. 2018). For example, alternative splicing of the Bcl-X gene produces two distinct isoforms with opposing functions. The first isoform, BCL-xS, is pro-apoptotic in function while the second isoform BCL-xL is anti-apoptotic. Aberrant splicing regulation of the BCL-X gene in tumor cells can lead to an increase in the relative expression of anti-apoptotic BCL-xL, further promoting the survival of cancer cells (B. D. Wang and Lee 2018; Kim et al. 2018).

Splicing factor genes are trans-acting regulators of alternative splicing decisions. Deregulation of splicing factor expression can lead to aberrant splicing events and contribute to cancer development and progression (Necochea-Campion et al. 2016; B.

D. Wang and Lee 2018; Zhou and Chng 2017). For example, the proto-oncogenic group of SRSF splicing factors genes are found to be overexpressed in several cancer types, producing a variety of aberrant splicing events leading to oncogenic protein isoforms (B. D. Wang and Lee 2018). Mutations in splicing factor genes can also lead to the production of aberrant splicing. Somatic mutations in the U2AF1 and SF3B1 splicing factor genes, for example, have been shown to promote cancer progression and aggressiveness. Further, previous studies indicate that aberrant splicing can have a significant impact on drug response with changes in the relative expression of specific splicing variants leading to decreased sensitivity towards a variety of cancer treatments (Necochea-Campion et al. 2016; B. D. Wang and Lee 2018; Zhou and Chng 2017). Several examples of splice variants leading to poor response have been identified and studied (Necochea-Campion et al. 2016; B. D. Wang and Lee 2018; Zhou and Chng 2017). Some of these aberrant splicing variants occur in well-known cancer-associated genes such as the BCR-ABL fusion gene in which increased expression of the BCR-ABL35INS variant leads to resistance of the tyrosine kinase inhibitor (TKI) imatinib in chronic myeloid leukemia (CML) patients (B. D. Wang and Lee 2018). These studies have identified splicing variants as potential therapeutic targets and have also pushed the way for the development of therapeutic strategies that target cancer-related splicing events (B. D. Wang and Lee 2018; Zhou and Chng 2017).

RNA-Sequencing

Characterization and understanding of genomic regulation in health and disease has increased substantially since the introduction of massively parallel high-throughput sequencing technologies in the mid-2000s (Hanahan 2000). Whole-transcriptome or RNA-sequencing (RNA-seq) platforms utilize the high-throughput approach of DNA-sequencing (DNA-seq) to comprehensively capture the transcriptome from a

population of cells. In a typical RNA-seq experiment, RNA is isolated from samples of interest using either an mRNA enrichment or ribosomal depletion strategy. The isolated RNA is then synthesized to cDNA and attached with adapters prior to amplification. Sequencing is then performed in a massively parallel fashion, producing millions of short sequencing reads representing the current cellular state of transcription. Thus, RNA-seq not only captures RNA at the sequence level, but also provides a snapshot of the amount of RNA being expressed across the transcriptome (Hanahan 2000).

The most common use of RNA-seq is differential gene expression in which the amount of RNA from each gene is quantified and compared across conditions. RNA-seq, however, captures the expression of all transcripts after splicing has occurred and can also be utilized for characterizing changes in alternative splicing (Hanahan 2000; Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010). Further, previous methods for measuring alternative splicing, including splicing-sensitive microarrays, were limited to only known splicing variants (Hanahan 2000). RNA-seq, on the other hand, has the ability to quantify novel transcripts as well, assuming the transcripts are expressed with enough abundance and the sample of interest is sequenced at a reasonable depth (Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010; Hanahan 2000).

Co-expression Networks

Proper cellular function is maintained by the dynamic transcription of coding and non-coding RNA across the transcriptome (Fagnani et al. 2007; Dai et al. 2012). Transcription of each gene, however, does not occur independently. Instead, transcriptional activation is a highly coordinated process under tight regulatory control with proteins and functional RNA interacting within cellular pathways to drive biological functions. These pathways are extremely complex and the functionality of genes and their corresponding interactions are often poorly

understood (Dam et al. 2018; Gaiteri et al. 2014). Co-expression networks provide an effective de novo inference approach for studying mechanisms of transcriptional regulation on a systematic level. Utilizing gene expression measurements from microarray and RNA-sequencing platforms, co-expression networks characterize transcriptome-wide coordination of gene expression based on their dynamic transcription across various conditions (Dam et al. 2018).

An underlying assumption of gene co-expression networks is that genes which are highly co-expressed are often involved in similar biological processes (Dam et al. 2018). This assumption can be extremely useful for a variety genomic applications involving transcriptional regulation. For example, gene co-expression networks can provide insight on the functionality of less annotated genes based on which genes they are highly co-expressed with (Dam et al. 2018). The also provide insight on regulatory potential by identifying genes having a significantly large number of interacting partners (Dam et al. 2018). Finally, co-expression networks can identify groups of co-expressed genes that are highly involved in driving a particular biological processes or phenotype (Langfelder and Horvath 2007).

Constructing and Analyzing Gene Co-expression Networks Gene co-expression networks describe pairwise relationships between genes based on their expression in a given dataset (Dam et al. 2018; Gaiteri et al. 2014; Langfelder and Horvath 2008; Sanati et al. 2018; Horvath 2011). In a given network, nodes represent genes and edges represent pairwise relationships (connections) between genes. The pairwise edges, however, are unknown prior to network construction and must be directly inferred using gene expression measurements from samples of interest (Figure 2). Various methods can be utilized to define pairwise relationships in gene co-expression networks. A commonly used correlation measure for inferring co-expression network edges is the Pearson correlation (or sample correlation). This

correlation represents a scaled version of a cosine correlation in which the vectors of sample expression for each gene are scaled by subtracting the mean of the vector from each value of the vector divided by the variance of the vector (Horvath 2011). The Pearson correlation is then computed by taking the cosine distance between the two scaled vectors for genes x and y:

$$cor(x, y) = cosineCor(scale(x), scale(y)).$$

Other correlation measures can also be utilized for inferring co-expression networks including the Spearman correlation measure (Horvath 2011). The Spearman correlation is more robust to outliers compared to the Pearson correlation and is determined by calculating the Pearson correlation based on the ranks of the vectors rather than the vectors themselves. The Spearman correlation, however, tends to be overly conservative. Authors of the popular weighted gene co-expression network analysis (WGCNA) framework have demonstrated the use of the biweight midcorrelation measure which utilizes the strengths of both the Pearson correlation and Spearman correlation, taking advantage of the high power aspect of Pearson correlation alongside the robustness of the Spearman correlation (Horvath 2011).

Network edges can then be defined using the resulting similarity values and are represented in the form of an n x n symmetric matrix known as an adjacency matrix $A = (a_{ij})$. Each pairwise entry of the adjacency can be in the form of a weighted or unweighted connection. In unweighted networks, relationships for each gene pair are represented as binary connections in which two genes are either connected (1) or not connected (0) (Dam et al. 2018; Gaiteri et al. 2014; Langfelder and Horvath 2008; Sanati et al. 2018; Horvath 2011). A basic approach for defining unweighted networks from resulting correlation measures is to utilize a hard-thresholding technique in which an edge is defined between two genes if their correlation is above a certain

value (e.g. ≥ 0.9). A more sophisticated approach for defining binary co-expression networks is the use of graphical lasso, a type of Gaussian graphical model (Friedman, Hastie, and Tibshirani 2008). Graphical lasso utilizes a regularization parameter similar to LASSO regression and attempts to estimate the precision matrix which is the inverse of a covariance matrix initially computed from the gene expression data. The resulting precision matrix represents pairwise dependencies between all gene pairs. Any non-zero entry represents a conditional dependency between two genes and thus a binary edge is added to the network (Friedman, Hastie, and Tibshirani 2008).

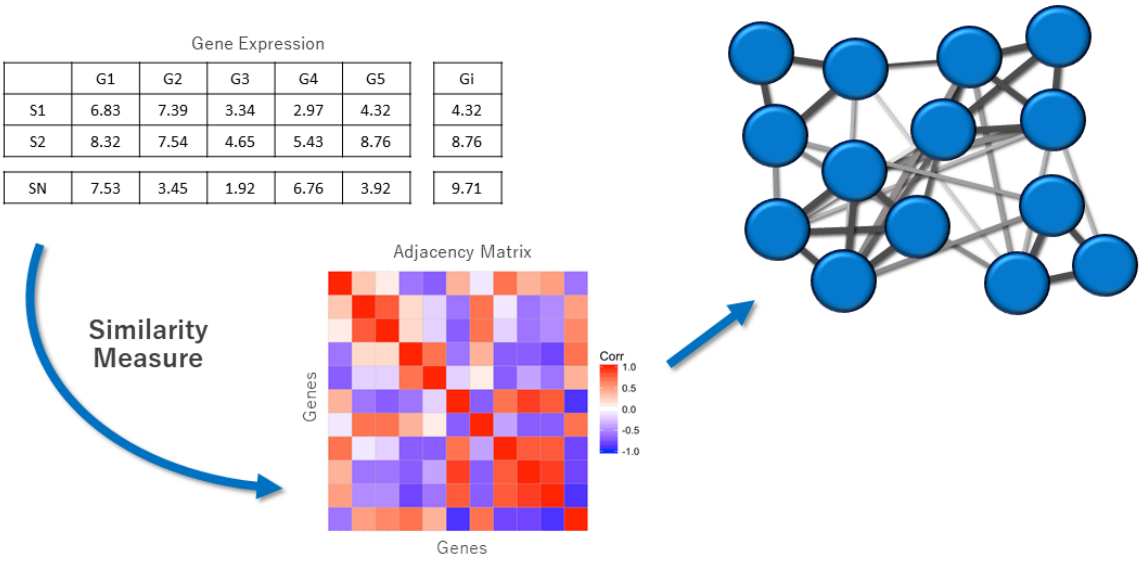

Figure 2: Overview of De Novo Co-expression Network Inference. In de novo network inference the network structure is unkown and must be inferred using sample expression data. A similarity measure is used to infer pairwise relationships between nodes and are stored in a symmetric adjacency matrix. A hard- or soft-thresholding technique and a possible edge transformation is applied to define edges of the network.

In weighted co-expression networks, an edge is present between every pair of nodes in the network. A value is then assigned to every edge and represents the strength of relationship between two genes based on their expression throughout each sample of the dataset (Horvath 2011). Initially, pairwise correlation values will be between -1 and 1, with edges closer to 1 representing stronger positive correlations and edges closer to -1 representing stronger negative relationships. In a typical co-expression

network, network edges are positive and thus a transformation is used to ensure all edge weights are between 0 and 1. Typically, a network is transformed into either a signed or unsigned network. Unsigned networks transform the edge weights using their absolute values, resulting in negatively correlated genes grouping together with their positively correlated counterparts. Signed networks, however, scale the edge weights so that negative correlation values are represented between 0 and 0.5 and positive correlation values are between 0.5 and 1. The signed network transformation approach has been shown to produce more biologically relevant networks than unsigned networks, but unsigned networks have also been successfully applied to certain biological applications (Horvath 2011).

It is critical to understand that utilizing any form of similarity measure (e.g. correlation) for de novo inference of co-expression has the potential to introduce spurious or noisy edge weights. The extent of spurious co-expression estimates is unknown and will vary depending on the dataset(s) used for the analysis. As previously stated, binary networks are often created using a hard-thresholding technique in which an edge is added only if two genes are highly correlated, however the appropriate choice of threshold is difficult to determine. Graphical lasso utilizes a more sophisticated approach for hard thresholding in which an edge is created between two genes if they are conditionally dependent (having a non-zero in the resulting precision matrix). The resulting precision matrix, however, is still dependent on the choice of regularization parameter during inversion of the covariance matrix (Friedman, Hastie, and Tibshirani 2008).

The Weighted Gene Co-expression Network Analysis (WGCNA) framework utilizes a soft-thresholding technique for weighted network edges. Here, the resulting adjacency matrix will undergo a power transformation (AFpower) in which the resulting edge weights are raised to the power of a constant value (β). Given that the choice of β is

unknown, WGCNA borrows from the assumption of biological networks having a close to scale-free topology. In a scale-free topological network, the connectivity of the network follows a power-law distribution in which a small number of nodes have an extremely high level of connectivity while the remaining nodes are less connected. Recent studies have shown that early claims of all real-world networks following a close to scale-free topology may be false and that scale-free networks are potentially quite rare (Broido and Clauset 2019). However, much evidence still indicates that biological networks (e.g. gene regulatory networks) follow a close to scale-free property. Nonetheless, the scale-free property is used as an approximation to guide the choice of β for power transformation in order to reduce the effect of spurious correlations. With an R^2 fitting index equal to 1 indicating a perfect power-law distribution, a β value can be selected that so that the R^2 fitting indicates approximate scale-freeness (e.g. $R^2 > 0.85$).

Module-based Co-expression Network Analysis Previous studies of transcriptional regulation using de novo network inference approaches have demonstrated gene expression occurring in the form of highly co-expressed modules (Langfelder and Horvath 2007). Using a community detection procedure to group network nodes into subgraphs of highly correlated genes, co-expression modules are often highly preserved across similar biological systems. The identification of network modules from a co-expression network can help characterize distinct groups of genes that may operate within specific biological pathways. Thus, the use of network modules provide a more comprehensive biological understanding of phenotype-specific regulation than traditional method for characterizing gene expression such as differential testing (differential expression) of individual genes. Further, the use of a module-based approach of co-expression provides an inherent data reduction technique (Langfelder and Horvath 2007). The expression of individual genes can be

summarized to the module level and one can then focus the analysis on a relatively small set of modules as opposed to thousands of genes, thus mitigating the issue of multiple testing. Whereas differential expression techniques identify extreme fold changes of the expression of individual genes across two phenotypes, a co-expression module statistically associated with two phenotypes may contain genes with both large expression changes as well as genes with less significant expression changes, yet are still highly relevant within the biological pathway in which the differentially expressed genes are involved (Langfelder and Horvath 2007).

Chapter 2: Creating a Module-based Network Framework for De Novo Co-splicing Inference of Complex Splicing Variants

Introduction

Chapter 2 describes an algorithm for de novo inference of co-splicing, the foundation of this dissertation. As the ability to infer systems-level coordinated splicing is highly dependent on the methods used for measuring alternative splicing variants, the chapter begins with a general review of the methods currently available for annotating and quantifying splicing from short RNA-sequencing reads. This is followed by a discussion on the important considerations when inferring co-splicing relationships and how the process differs from that of gene co-expression inference. Finally, the chapter presents an algorithm for formulating complex alternative splicing events quantified from RNA-seq samples in a manner suitable for a module-based network approach of co-splicing.

Methods for Quantifying Splicing From RNA-sequencing

A key application for RNA-seq is to measure differences in the expression of genes and splice variants across different cell types (e.g., tissue types) or conditions (e.g., healthy and disease). In doing so, the expression of genes and transcripts must be accurately quantified using sequencing reads each sample. Compared to gene expression, measuring changes in alternative splicing is far more difficult due to the fact that RNA-seq reads are typically between 50 and 150bp long and are much shorter than that of exons and introns (Hanahan 2000; Liu, Loraine, and Dickerson 2014; Anders, Reyes, and Huber 2012; Shen et al. 2014). This produces significant complexity issues when trying to accurately annotate and quantify multiple

alternatively spliced transcripts of the same gene (Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010; Hanahan 2000; Anders, Reyes, and Huber 2012; Shen et al. 2014). When the samples of interest belong to a well-studied organism such as human or mouse, reference-based assembly methods can accurately align short RNA-seq reads to their genomic location of origin (Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010; Hanahan 2000). When quantifying expression at the gene level, all reads aligning within the region of a particular gene are counted towards the overall expression of that gene. Quantifying alternatively spliced transcripts of the same gene, however, is a more difficult task. Transcripts resulting from alternative splicing will typically share multiple exons or exon regions with one another. Thus, when a read aligns entirely to a genomic region (exon) shared by two or more alternative transcripts, there is no way to distinguish the transcript of origin for that particular read (Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010; Hanahan 2000; Anders, Reyes, and Huber 2012; Shen et al. 2014).

A plethora of methods for quantifying alternative splicing using short read sequencing data have been developed since RNA-seq was introduced in the late 2000s. Many of these methods are also designed to test for significant changes in the expression of splice variants between two or more conditions (Liu, Loraine, and Dickerson 2014; Trapnell et al. 2010; Hanahan 2000; Anders, Reyes, and Huber 2012; Shen et al. 2014). Changes in the expression of splice variants are often described as differential splicing and is somewhat analogous to differential gene expression (Anders, Reyes, and Huber 2012; Shen et al. 2014). However, whereas differential expression describes the change in the total expression of a particular gene, differential splicing typically describes the change in relative abundance of each splice variant (or alternative spliced transcript) contributing to the total gene expression. Further, a gene may be significantly differentially spliced between two conditions without having a significant change in total expression (Liu, Loraine, and Dickerson 2014). The majority of

methods that have been developed for quantifying splice variants and testing for differential splicing between two or more conditions can be classified into to one of three categories: full isoform resolution models, exon-based models, and event-based models (Liu, Loraine, and Dickerson 2014).

Full Isoform Resolution Models Isoform resolution models first attempt to assemble short sequencing reads into full length transcripts and then estimate their relative abundance (Trapnell et al. 2010; Hanahan 2000). In the assembly step, methods such as Cufflinks will utilize a graph-based structure to try and resolve the minimal set of transcripts that best explain the data. The abundance of each assembled transcript is then quantified using statistical modeling. Often with the use of a maximum-likelihood estimation, each read is assigned to a potential transcript before the final transcript abundances are estimated. This is performed using a statistical modeling approach that accounts for uncertainty in each read's transcript of origin along with cross-replicate variability estimates. The final transcript abundance estimates are then used to test for significant changes in the relative expression of each transcript between biological conditions (Trapnell et al. 2010; Hanahan 2000). These methods are beneficial for characterizing splicing in that they provide full length transcripts along with their relative abundance and change in abundance between conditions. The results are intuitive for studying alternative splicing and are thus easily interpretable for further analysis on splicing changes in different conditions. A significant drawback of these methods, however, are the complexities of full length transcript assembly and abundance estimation using short sequencing reads. First, the resulting number of transcripts from the assembly step may inaccurately represent the number and structure of the truly expressed transcripts for a given gene. Second, transcript abundance estimation can introduce a high degree of uncertainty given that a significant proportion of reads can potentially

belong to multiple transcripts. This degree of uncertainty can increase significantly when the number of possible alternative transcripts increases (Liu, Loraine, and Dickerson 2014).

Exon and Event-based Splicing Quantification Methods A second class of methods for measuring splicing changes try to circumvent the transcript complexity issue by focusing only on the expression of individual exons (Anders, Reyes, and Huber 2012). These exon-based approaches utilize the fact that many transcripts share multiple exons. Rather than trying to identify which transcript a read belongs to when it aligns to an exon shared by multiple transcripts, it quantifies the total expression of the exon while ignoring the transcript of origin. To do this, exon-based methods such as DEXSeq take the union of all possible exons for a given gene's set of transcripts (provided within a gene structure annotation file) and compresses them into non-overlapping exonic bins. The expression of each exon bin is quantified using a negative binomial model for read counts. DEXSeq then tests for differential splicing of a gene between two or more conditions using a generalized linear model (GLM) that tests for significant differential exon usage of each counting bin relative to the total expression of all exon bins belonging to that gene (Anders, Reyes, and Huber 2012).

The third class of methods extends upon the approach of exon-based methods to avoid the complexities of transcript assembly and abundance estimation while incorporating more biological relevance towards alternative splicing of pre-mRNA (Shen et al. 2014). Event-based methods utilize a compress and count approach similar to exon-based methods, but instead of compressing transcripts into a union of individual exons, regions of transcripts are flattened into individual alternative splicing events. Exon structures for each transcript belonging to a given gene model are checked to see if they participate in one of the five basic splicing event types. Each splicing event identified can be broken into a set of binary outcomes (e.g., the

skipping or inclusion of a cassette exon). The two outcomes for each splicing event are quantified using junction spanning reads that support either of the two cases. Splicing event quantities are typically represented using a metric known as percent spliced in (PSI or Ψ) which represents the proportion of one splicing event outcome over the total of both outcomes. Differential splicing is performed by testing for significant change in the PSI of splicing events between biological conditions (Shen et al. 2014).

Splicing Graph Methods A more recent class of methods for characterizing alternative splicing from RNA-seq data utilize a somewhat hybrid approach, borrowing from full length transcript models while incorporating localization from that of exon and exon-based methods. Splicing graph methods utilize a graph-based view of the possible splicing variations to which a transcribed gene may undergo. First, a splicing graph is created for each gene representing the inclusion and exclusion of all possible exonic segments. Splice graphs can be constructed from previously annotated gene models or from split read alignments of splice junction spanning reads. This approach is similar to that of full isoform resolution models such as Cufflinks which first assembles a graph of all possible transcript variations. However, unlike full transcript models, splice graph quantification models will avoid transcript estimation and instead quantify localized variants found within the graph. Most often, methods will first assemble the splice graph from each gene and then search the graph for the presence of the basic alternative splicing events types (e.g. exon skipping). Splice graphs that incorporate junction spanning read alignments during graph construction have an advantage in that novel splicing variants can be annotated prior to quantification.

Splicing graph-based approaches provide a more comprehensive representation of the potential splicing complexity for a transcribed RNA. Like the event-based approaches that strictly utilize known gene models, graph-based approaches that annotate

splicing in the form of basic splicing events fail to capture the full complexity to which splicing variants can entail. Some recently developed graph based approaches, however, attempt to annotate complex splicing variants prior to the quantification process. Methods such as Leafcutter and MAJIQ redefine what constitutes and alterantive splicing event. Their annotation approaches can in turn quantify splicing at varying degrees of complexity such as non-binary events or compound events involving more than one event type. Recent characterizations of complex splicing in mammalian transcriptomes have found that nearly 30% of splicing occurs in the form of complex splicing events (Li et al. 2018; Vaquero-Garcia et al. 2016).

The Modeling Alternative Junction Inclusion Quantification (MAJIQ) framework first develops per-gene splicing graphs using a combination of previously annotated gene models and de novo junction spanning RNA-seq reads (Vaquero-Garcia et al. 2016). Splicing variants are then annotated by formulating local splicing variants (LSVs) defined from nodes with multiple ingoing or outgoing edges of each splice graph. Each LSV consists of two or more junctions and are quantified using marginal percent selected index (PSI or Ψ) representing their relative expression to that of all junctions of the LSV (Vaquero-Garcia et al. 2016). In a typical experiment, the relative change in PSI ($\Delta\Psi$) would be tested between two conditions and such an analysis is analogous to that of differential expression of genes between two conditions (Figure 3).

Previous Work for Network Inference of Alternative Splicing

Most genomic applications of de novo network inference have focused on gene-level co-expression. Some methods, however, have been developed to study coordinated transcription beyond just co-expressed genes and instead focus on the co-expression of individual transcripts or exons. (Saha et al. 2017) developed gene co-expression networks from 16 tissues types using RNA-seq data from the Genotype-Tissue Expression (GTEx) project. Along with co-expressed genes, their networks also

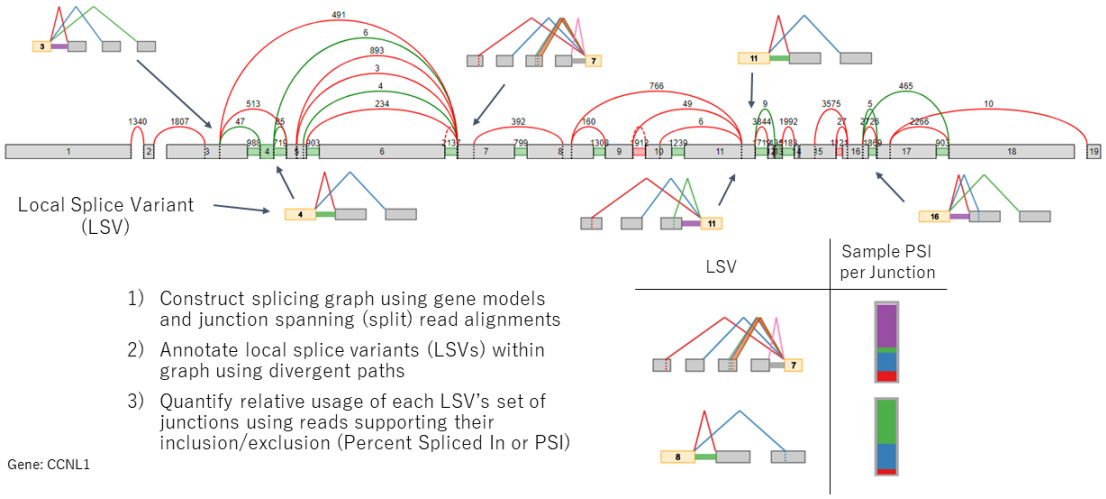

Figure 3: Overview of the MAJIQ Framework for Complex Splicing Quantification. Gene splicing graphs are constructed using previously annotated transcripts and junction spanning read alignments from RNA-seq samples. Complex splicing is annotated and quantified from each graph in the form of local splice variants (LSVs). The relative usage of each junction within an LSV is measured using reads supporting either their inclusion of exclusion.

contained the relative expression of transcripts, creating a single network with nodes representing both genes and transcripts and edges defining the correlations between gene-pairs, transcript-pairs, and gene-transcript pairs. They utilized this network to identify potential regulatory genes such as transcription factors and splicing factors (Saha et al. 2017). This network, however, is potentially limited in that correlations for defining edges between transcript-transcript pairs and transcript-gene pairs are dependent on the accurate estimation of transcript expression. As previously discussed, transcript abundance estimation is a highly difficult and non-trivial task using short RNA-sequencing reads and can often lead to inaccurate estimations of transcript-level expression. Further, the incorporation of individual transcripts into a network is also dependent on accurate transcript assembly, another difficult task using short read data. The number of transcripts assembled by many short read assembly methods may not accurately represent the true number of alternatively spliced transcripts that were expressed in a specific condition and could lead to either an incomplete or inaccurately represented network.

Other methods have demonstrated the use of co-expression networks at exon-level resolution. (Dai et al. 2012) developed a tensor-based algorithm to identify exon clusters across multiple networks. Each network was constructed using the expression of individual exons as nodes with edges representing the correlation between the expression of each exon-exon pair across samples. (Iancu et al. 2015) developed a gene co-expression network based on correlations of aggregated exon-level expression. Using mammalian brain RNA-seq data, this network was built by defining gene-pair edges based on Mantel correlations between gene matrices. Each gene matrix represented the difference in expression of all exons for all samples of that gene. This approach was successful in identifying unique, yet biologically relevant gene modules only detectable using exon-level correlations between genes.

Network Inference of Co-splicing Modules Consisting of Complex Splice Variants

It is well known that gene expression is a tightly regulated process with genes interacting in a coordinated fashion within cellular pathways. Gene co-expression networks provide a systems-level framework for inferring both conserved and phenotype-specific coordination of transcription in different biological systems. Alternative splicing of both coding and non-coding RNA also occurs in a highly coordinated manner and recent studies have begun to utilize de novo network inference approaches to elucidate transcriptome-wide coordination of alternative splicing regulation (co-splicing). A systems-level framework that integrates the full complexity of alternative splicing variation in the form of network modules in a manner suitable for the use of studying their association with phenotypic traits, however, has not been demonstrated.

Inferring coordinated alternative splicing (co-splicing) at a network-level is difficult due to the ubiquitousness of alternatively spliced transcripts. A network consisting of isoforms may contain spurious and unreliable network edges due to inaccuracies when estimating full length transcript expression using short RNA-seq reads. Co-splicing inference methods that utilize exon-level gene correlations to mitigate the issues of transcript estimation do not capture complex alternative splicing variants at a level of granularity to that of recent splicing graph approaches. Thus, we sought to develop a co-splicing network approach that 1) characterizes transcriptome-wide coordination of complex splicing variants, 2) formulates complex splicing in a manner suitable for a module-based network approach, 3) allows for a modular-level analysis of splicing variation across one or more biological systems. In this study we first describe an approach that formulates complex splicing variants from splice-graphs annotated using the MAJIQ framework in a manner suitable for a module-based network

analysis of co-splicing. We then demonstrate the use of our formulated co-splicing modules within the popular Weighted Gene Co-expression Network Analysis (WGCNA) framework traditionally used for studying gene co-expression. We applied our co-splicing module approach in two applications. First, we characterize co-splicing variation in the form of network modules across ten human tissues types. We then construct a co-splicing network of acute myeloid leukemia (AML) and identify co-splicing modules predictive of drug response for a variety of small molecule inhibitors.

In a gene co-expression network, nodes represent genes and edges represent pair-wise relationships (connections) between genes. In de novo inference of gene co-expression, the edges are unknown prior to network construction and therefore must be directly inferred from the data using a similarity or correlation measure. Compared to gene co-expression, inferring system-wide relationships between localized splicing variants is less straightforward. A single gene may and often will contain multiple splice variants. Splicing variants originating from the same gene are biologically plausible to be highly correlated, however the extent of which is difficult to ascertain given the uncertainty as to whether the change in relative splicing of both variants is due to changes in the relative expression of the same transcripts. Co-splicing networks inferred using exon-level correlations aggregated to gene level do not include within-gene splicing relationships. However, two splicing variants from one gene may have different correlations with a variant from a second gene. Biologically speaking this may represent an isoform (and resulting protein) having different functional interactions with two isoforms of the same gene and such an event would not be captured in a gene-level co-splicing approach.

We find in our data that pairs of LSVs originating from the same gene indeed have stronger correlations than those of different genes (Figure 4A). A number of

within-gene LSV pairs, however, are found within close proximity to one another with partial or even fully shared exonic regions involved in each variant. Such LSVs are often found within clusters of skipped exons, retained introns, and alternative splice sites to which the inclusion and exclusion of specific RNA segments can occur in a variety of forms. We find that correlations of LSV pairs containing overlapping exonic regions are significantly greater than same gene LSV pairs that do not overlap (Figure 4B). Such correlations are often highly redundant and may contain potential noise due to inconsistent depth of coverage for splice junction spanning reads. Therefore, we propose a pre-processing step that summarizes groups of overlapping LSVs from a splice-graph prior to network construction.

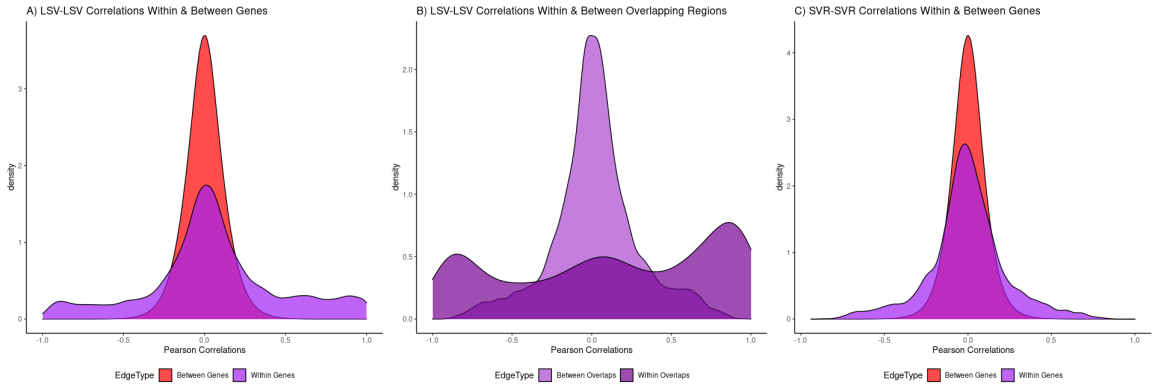

Figure 4: Distribution of Edge Weight Correlations. Density plots represent distribution of Pearson correlations computed using splicing features quantified from AML RNA-seq samples. A) Correlation distributions for between-gene and within-gene LSVs. B) Correlation distributions for overlapping and non-overlapping within-gene LSVs. C) Correlation distributions for between-gene and within-gene SVRs.

Splice Variant Regions (SVRs)

We first identify for each LSV, the most up- and down-stream positions among all of its corresponding junctions. We then iteratively merge all LSVs having overlapping genomic positions until merging can no longer be accomplished. We define such overlapping groups of LSVs as splice variant regions (SVRs) (Figure 5). Given that the set of PSI values for each junction of an LSV are quantified relative to their sum,

we select a single junction to represent the splicing level of each LSV to avoid linear dependency between junctions. For each LSV we select the junction having the greatest variance across all samples of the dataset.

We then define for each SVR a matrix $X^{(I)} = (x_{il}^{(I)})$ where index $i = 1, 2, ...n_I$ corresponds to the LSVs within $SVR^{(I)}$ and index $l = 1, 2, ..., m$ corresponds to the RNA-seq samples in the dataset. We summarize the splicing levels of all LSVs belonging to $SVR^{(I)}$ using Singular Value Decomposition (SVD), a commonly used data reduction technique. We denote SVD of $X^{(I)}$ as

$$X^{(I)} = UDV^T$$

where $U^{(I)}$ and $V^{(I)}$ are orthogonal matrices and the columns of U and V are the left- and right-singular vectors respectively. Assuming the values of D, a diagonal matrix of singular values, are arranged in decreasing order, we represent the splicing value of $SVR^{(I)}$ using the first column of $V^{(I)}$ as

$$SVR^{(I)} = v_1^{(I)}.$$

This representation is equivalent to performing principal component analysis (PCA) on the matrix $X^{(I)}$ and representing the splicing of $SVR^{(I)}$ using the first principal component.

After constructing SVRs from overlapping LSVs of splice-graphs we find that the difference in correlation distributions between within-gene and between-gene SVRs is highly reduced compared to that of LSV correlations (Figure 4C). The formulation of SVRs prior to network construction prevents the inclusion of redundant and potentially noisy network edges between splicing variants while still retaining variation in splicing levels of complex splice variants across samples. Further, the

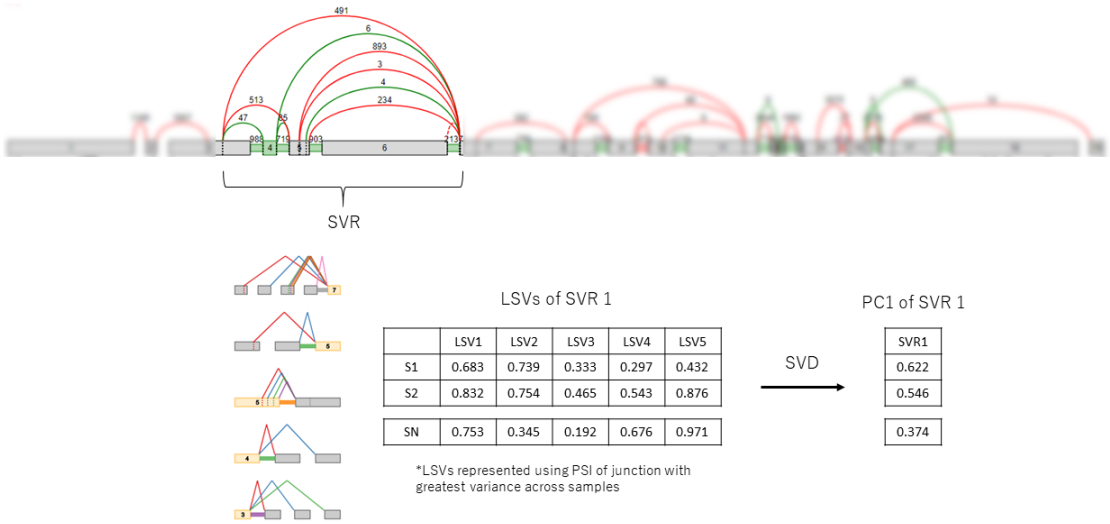

Figure 5: Formulation of Splice Variant Regions (SVRs) From LSVs For Co-splicing Network Inference. SVRs are formulated using all sets of overlapping LSVs from a given splice graph. The splicing value of the SVR is quantified using the 1st principal component after SVD using all LSVs within the SVR.

SVR datatype is also biologically relevant given that SVRs consist of clusters of alternatively used exonic segments undergoing a variety of usage patterns. SVRs capture the variation in splicing levels of such clusters across the sample set and thus provide a suitable data structure for representing complex splicing variants for de novo network inference of co-splicing.

Chapter 3: Characterization of Co-splicing Variation Across Human Tissues (Use case)

Chapter 3 discusses the first use case for the co-splicing algorithm in which we characterize co-splicing variation in the form of network modules across ten human tissue types. The results of the analysis in chapter serve as a proof of principal for the proposed co-splicing module-based network inference framework described in the previous chapter.

Introduction: Tissue-specific Regulation of Transcription and Splicing

Alternative splicing plays an essential role in the regulation of cellular phenotypes. Variation in the expression of specific alternative transcripts can lead to varying gene products between particular tissues. With the advent of RNA-sequencing, the role of alternative splicing in tissue-specificity is becoming more clear. Differentially expressed genes and differential splicing events have been extensively identified between tissue types and stages of cellular differentiation (Grange et al. 2010; The GTEx Consortium 2015). The Genotype Tissue Expression (GTEx) project provides an extensive dataset of well curated genomic data across multiple tissues (The GTEx Consortium 2015). This resource has led to new insights on transcriptional regulation in regards to tissue specificity.

Given the complexities of transcriptome-wide regulation, many de novo network inference approaches have been utilized to study gene co-expression variation across human tissues (The GTEx Consortium 2015; Pierson et al. 2015). Such approaches have identified characteristics of gene co-expression in both a preserved and tissue-specific manner. Likewise, a recent study demonstrated a framework for

integrating transcript expression ratios into a gene co-expression network with the goal of characterizing regulatory mechanisms of gene expression and alternative splicing (Saha et al. 2017). Using data from GTEx, the authors demonstrated the use of Transcriptome-Wide Networks (TWNs) to identify shared and tissue-specific gene co-expression hubs that may regulate the expression of other genes and isoforms. The primary focus of their analysis was the characterization of potential regulatory factors both in terms of gene expression and splicing. A potential limitation to their study was the use of full length transcripts estimated from RNA-seq data from the human GTEx tissue samples. The authors note that care should be taken when making conclusions in regards to specific network edges, which is true for any de novo inference based analysis. This aspect could be of particular concern when using transcripts in a de novo network approach given the known issues of transcript expression estimation methods (Li et al. 2018; Sterne-Weiler et al. 2018; Shen et al. 2014; Vaquero-Garcia et al. 2016).

Studies utilizing gene co-expression networks have demonstrated the presence of co-expression modules, including modules preserved across tissue types (Langfelder and Horvath 2007). Co-splicing in the form of network modules across tissues is less characterized. In this chapter we apply our proposed co-splicing approach using SVRs formulated from alternative splicing graphs quantified using RNA-seq sampes from multiple human tissue types. The use of SVRs provide an efficient means for accurately quantifying splicing in a manner suitable for a network inference approach. Further, groups of SVRs in the form of network modules can be summarized to the module level, allowing for a module-based analysis of co-splicing. We note that the choice of de novo network approach is highly dependent on the question at hand and the authors in the aforementioned study of splicing regulation were addressing a different question in regards to phenotypic regulation. The analysis and results in this chapter focus on groups of highly co-expressed alternative splicing variants and

characterize the variation of these modules across human tissues.

Results

Constructing Co-splicing Networks Across Human Tissues

We aimed to characterize transcriptome-wide coordination of complex splice variants in the form of co-splicing modules across human tissues. To study between-tissue co-splicing variation we first annotated and quantified LSVs from RNA-seq data of ten human tissue types from GTEx using MAJIQ. Only LSVs that met the minimum coverage thresholds (Methods) in all 1,621 donor samples during PSI quantification were included in downstream analysis. This resulted in 6,427 LSVs from 2,147 genes which were then used to formulate 4,147 SVRs. All tissue samples were utilized during summarization of each SVR to ensure the leading eigenvector is consistent across datasets and represents splicing variation across tissue types. Figure 6 shows the result of bi-clustering using the formulated SVRs from the ten human tissue types. The formulated SVRs are able to effectively distinguish the four main tissue groups. Tissue subtypes are less distinct, but subtype-specific splicing is still clearly present from the clustering results and heatmap.

All 4,147 SVRs were utilized for de novo network inference of co-splicing. For each tissue we inferred tissue-specific network edges between SVRs using biweight mid-correlations and the resulting correlation values were linear transformed to a signed network using $s_{signed,i,j} = \frac{1+cor(x_i,x_j)}{2}$. As the use of correlation measures for inferring networks may lead to spurious or noisy edge weights, we raised each of the ten signed adjacency matrices to a power of β. This transformation serves as a soft-thresholding technique that promotes strong correlations while suppressing weak ones. Following techniques from WGCNA, since the choice of β is unknown, we borrow from the assumption that biological networks follow an approximate scale-free

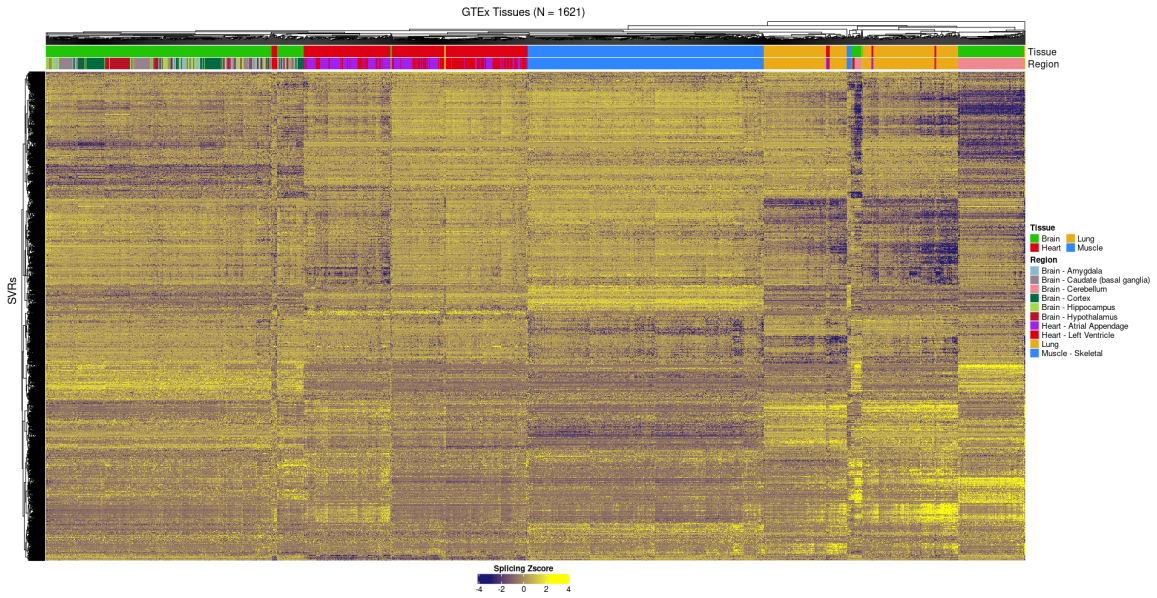

Figure 6: Bi-clustering and Heatmap of SVRs Formulated Using Human Tissue RNA-seq Data. Bi-clustering was performed using complete linkage based on euclidean distance. All 4,147 SVRs from 1,621 tissue samples across ten tissue types were used for clustering.

topology in which there are a small number of nodes having a large number of connections. At $\beta = 5$ all ten tissue co-splicing networks reach an R^2 scale-free fit greater than 0.9 and was thus chosen as the β value for all networks.

Following power transformation we then performed a topological overlap transformation $TOM(A)$ for each tissue adjacency matrix. This transformation provides additional topological relationship characteristics among the nodes of each network. Unlike the initial adjacency matrices where each inferred edge weight is calculated independently from the remaining nodes of the network, the topological overlap measure (TOM) describes relationships between nodes while also accounting for shared relationships among neighboring nodes.

Identifying Consensus Co-splicing Modules Shared Across Tissues

To analyze shared and tissue-specific co-splicing characteristics at a module level we identified a consensus set of co-splicing modules across the ten tissue types.

Consensus network modules represent groups of nodes that are highly correlated and preserved to an appreciable degree in all sample groups (i.e. tissues) of the analysis (Langfelder and Horvath 2007). To identify consensus co-splicing modules we first compute a consensus topological overlap matrix, representing the minimal connectivity of nodes across the ten tissue networks. Using the consensus TOM we then perform a network clustering procedure to identify a set of co-splicing modules shared across the ten tissue types. Many network clustering techniques exist and the choice of clustering is typically driven by the level of granularity necessary for the analysis (Iancu et al. 2014). Under the WGCNA framework it is common to perform average linkage hierarchical clustering on the resulting $TOM(A)$. Hierarchical clustering has an advantage in that the number of clusters does not need to be known beforehand. Other methods, however, have also been proposed for clustering co-expression networks including the use of ensemble approaches. (Botia et al, 2017) first identified gene co-expression modules using average linkage hierarchical clustering, but then performed a re-clustering step using k-means clustering with k being the number of clusters detected in step one (Botia et al. 2017). We extend upon this ensemble approach by first performing hierarchical clustering to detect an initial module count and then re-cluster each network using spectral clustering which has been recently demonstrated as a method for module detection in co-expression networks (Al-Yousef and Samarasinghe 2021).

Following the WGCNA framework, we first computed the consensus topological overlap dissimilarity matrix ($1 - ConsensusTOM$) as input for hierarchical clustering and identified an initial module set. Separately for each tissue, we summarized the expression of each module by performing SVD using all of the corresponding nodes and then merged any closely related modules having a Pearson correlation ≥ 0.75. Nodes not belonging to a proper module (denoted as "grey" in WGCNA) are removed prior to spectral clustering, but remain in the final network

assigned to the improper "grey" module. A total of 13 proper modules remained after merging and k = 13 was thus used for spectral clustering.

Spectral clustering was performed by first transforming the original consensus topological overlap matrix into a degree-normalized Laplacian matrix defined as $L = D - A$, where D is a matrix of the degrees of A and A is the initial adjacency matrix (in this case $Cons(TOM(A))$. Then, the top k eigenvectors are computed on the resulting graph Laplacian and k-means clustering is performed on the network nodes using the resulting eigenvectors. We identified 13 consensus co-splicing modules across the ten tissue types containing 199-544 SVRs from 184-494 unique genes (Figure 7).

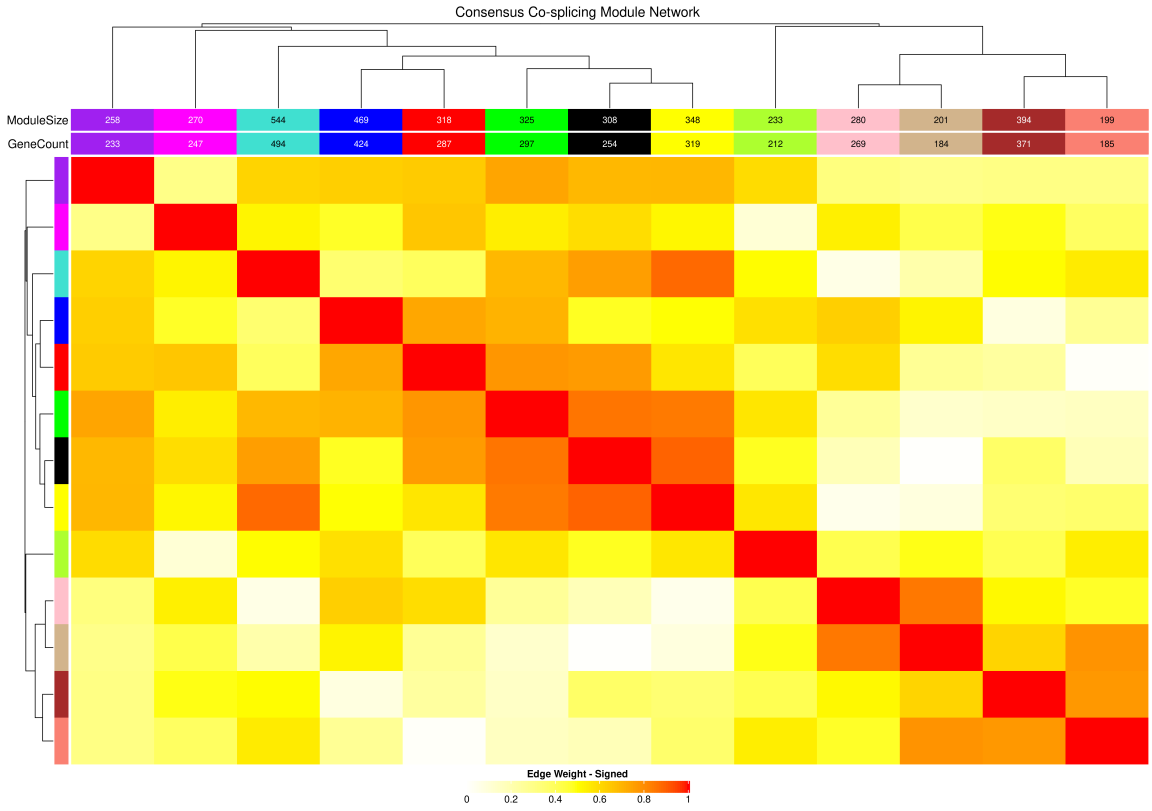

Figure 7: Consensus Module Network of GTEx Tissues. Heatmap values represent inter-modular relationships based on module correlations from all tissue types. The heatmap dendrogram represents the consensus hierarchy of modules across tissue types. Not shown are tissue-specific consensus module clustering and edge weights.

Of the 2,147 genes represented in each tissue co-splicing network, over half of the genes are represented by more than one splicing variant (Figure 8A). Therefore, splicing variants for a single gene may be found in more than one module. Similarly, a single gene may have more than one splicing variant present within the same module. Figures 8B and C show the distributions of genes across modules as represented by their splicing variants. Of the 2,147 unique genes found within the co-splicing network, nearly half of the genes contain one or more splicing variants found within two or more splicing modules. Within each module, genes are more often represented by a single splicing variant, however, a small portion of genes can be found having two or more splicing variants within the same module.

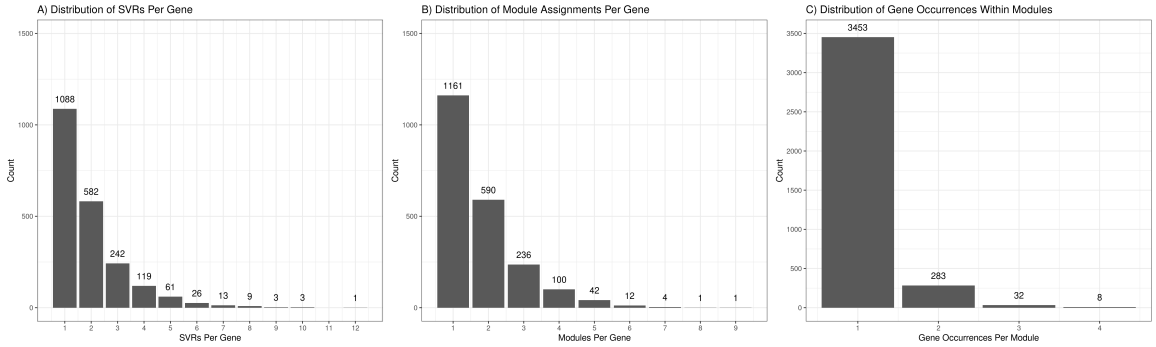

Figure 8: Characterization of SVRs and Genes Across Consensus Modules. A) Distribution showing the number of splice variant regions for each gene of the tissue co-splicing networks. B) Distribution showing the number of modules a gene belongs to given their splicing variants. C) Distribution showing the number of splicing variants belonging to a single gene within each co-splicing module

Quality of consensus module detection was evaluated using a variety of module quality statistics including module density, module connectivity, and overall module summary (Langfelder et al. 2011). For each module statistic we computed a Zscore using the observed module statistic relative to the mean and standard deviation of 10,000 random node to module assignments. Quality statistics for each module were computed using each tissue network separately. Zscores < 2 were implied as being of poor quality for a Z-statistic of a given module with Zscores > 10 implying high

quality modules. All 13 consensus modules were found to have high *Zconnectivity* ($Z > 10$) in all tissues (Figure 9). *Zdensity* values were of moderate ($2 < Z < 10$) to high quality ($Z > 10$) in almost all modules across each tissue, with only the turquoise module having *Zdensity* values < 2 in three of the ten tissue networks (both heart tissues and lung). *Zsummary*, representing the average *Zdensity* and *Zconnectivity* of a given module, was also of high quality ($Z > 10$) for all consensus modules in each tissue. These results indicate that the identified consensus modules represent groups of splicing variants that are indeed highly co-spliced across tissue types.

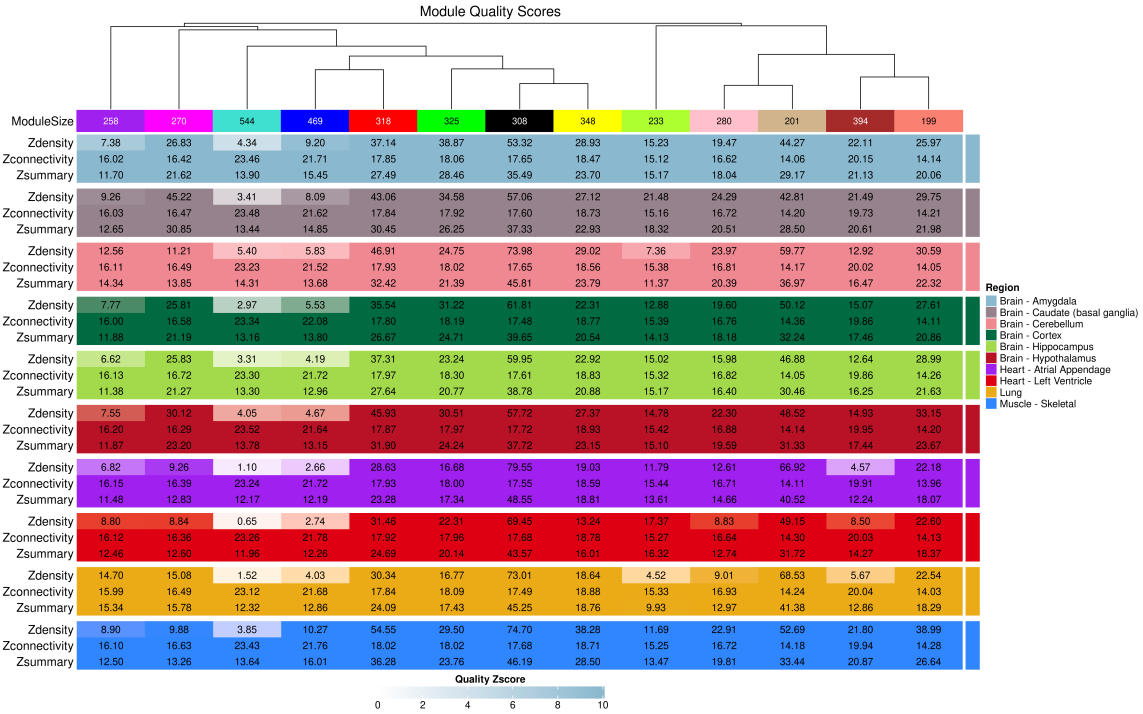

Figure 9: Module Quality Zscores Scores Across Tissues. Modules are ordered based on the consensus module hierachy and Zscores are grouped by tissue. Heatmap intensity is capped at 10 in order to indicate modules in tissues having less than high quality Zscores.

Differential Co-splicing of Network Modules Across Tissue Types

Here, we characterized the intra-modular preservation of co-splicing modules between the ten tissue types. Differences in the hierarchical structure of co-splicing modules

between tissues may reveal tissue-specific differences of splicing regulated pathways. Following WGCNA we first summarize the splicing level of each module separately for each tissue type using SVD. Here, module splicing values (eigengenes in WGCNA) capture the within-tissue variation of SVRs for each module. We then cluster the consensus co-splicing modules using hierarchical clustering based on the within-tissue module splicing values. The dendrogram in figure 7 represents the consensus hierarchy of the consensus modules across the ten tissue types.

As defined by (Langfelder & Horvath 2007) the preservation of two modules between two tissue networks is

$$Preserv_{IJ}^{(1,2)} = 1 - \frac{|cor(M_I^{(1)}, M_J^{(1)}) - cor(M_I^{(2)}, M_J^{(2)})|}{2},$$

where $M_I^{(s)}$ is the module splicing value of the I-th splicing module in tissue s.

The scaled connectivity preservation of a module between two tissue networks is defined as

$$C_I(Preserv^{(1,2)}) = 1 - \frac{\sum_{I \neq J} |cor(M_I^{(1)}, M_J^{(1)}) - cor(M_I^{(2)}, M_J^{(2)})|}{2(N-1)},$$

and describes the preservation of correlation between the I-th module and the remaining modules between the two tissue networks.

The density preservation between two tissue networks is defined as

$$D(Preserv^{(1,2)}) = 1 - \frac{\sum_I \sum_{I \neq J} |cor(M_I^{(1)}, M_J^{(1)}) - cor(M_I^{(2)}, M_J^{(2)})|}{2N(N-1)},$$

and describes the overall preservation of module connectivities between two tissues.

In addition to the consensus module dendrogram, we also perform hierarchical

clustering of co-splicing modules on each tissue separately, thus characterizing the tissue-specific hierarchy of co-splicing modules. Figure 10 shows an example of the three module network preservation statistics along with tissue-specific module dendrograms. For simplicity, the module preservation statistics are applied to four of the ten tissue types (two heart tissues and two brain tissues). Looking at figure 10 it is easy to observe that inter-modular co-splicing is more preserved between similar tissue types than non-related tissues with, edge weights of the module networks showing similar co-splicing patterns between related tissues. Comparing scaled connectivity preservation ($C_I(Preserv^{(1,2)})$) of each module between tissue pairs shows that while most modules remain relatively preserved in scaled connectivity, non-related tissues (e.g. hippocampus and atrial appendage) show decreased module connectivity preservation for the magenta and greenyellow co-splicing modules.

Figure 11 shows a heatmap of the overall preservation ($D(Preserv^{(1,2)})$) values across all ten tissue pairs. As expected, related tissue types were found to have higher levels of co-splicing preservation than un-related tissue types. Clustering of the ten tissue networks using $1 - D(Preserv^{(1,2)})$ as a distance measure results in grouping of similar tissues. The two heart tissues, atrial appendage and left ventricle, showed higher preservation with each other than with any of the remaining tissue types. Of the six brain regions, all but cerebellum showed higher preservation with other brain regions than with non-brain regions.

Functional Enrichment of Consensus Co-splicing Modules

We performed functional enrichment of genes derived from consensus co-splicing modules to characterize their biological function. Co-splicing modules were significantly enriched for numerous GO biological process terms with modules containing both shared (Figure 12) and module-specific (Figure 13) terms. Many of the enriched terms for each module include basic cellular functions to which splicing

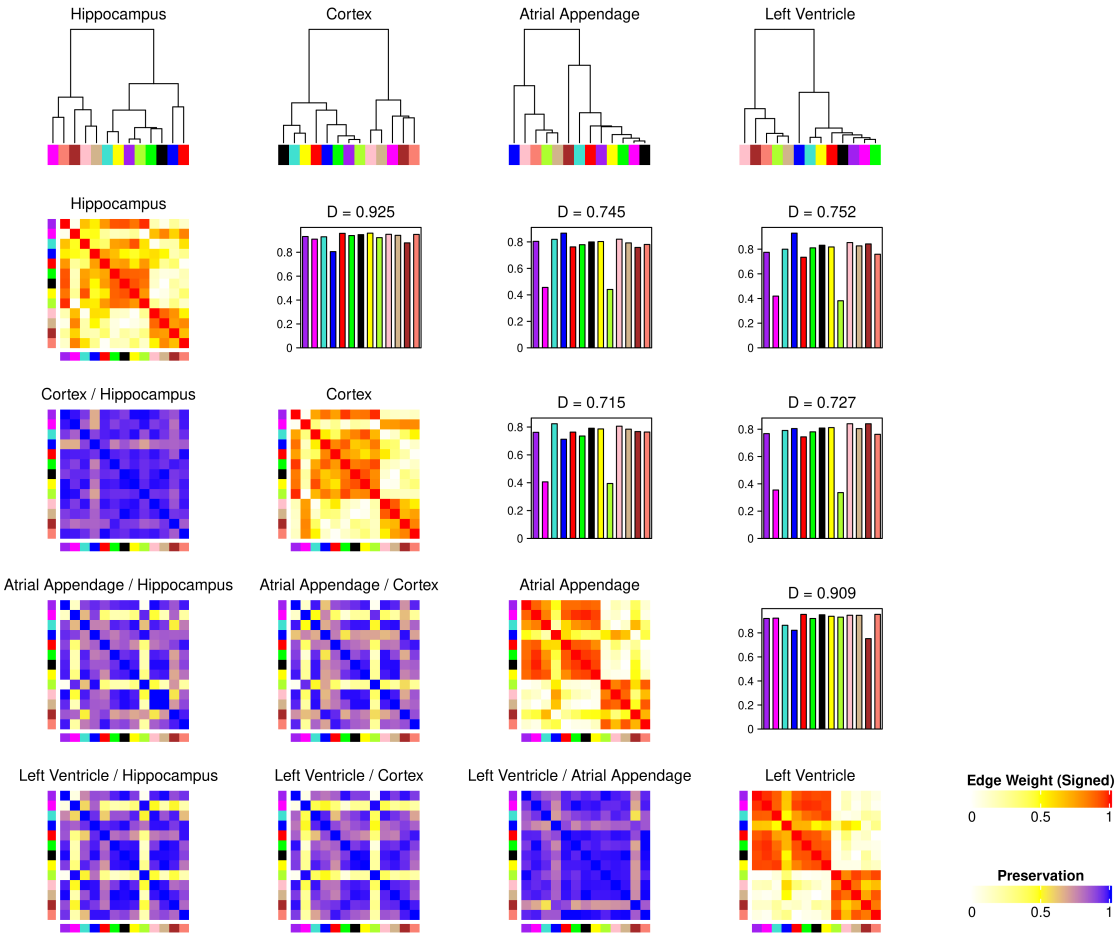

Figure 10: Differential Co-splicing of Network Modules. Two brain tissues (hippocampus and cortex) and two heart tissues (atrial appendage and left ventricle) are shown for example. Top row indicates tissue-specific module clustering of consensus modules. Heatmaps going diagonal from row two to row five represent tissue-specific edge weights of tissue module networks. Yellow and blue heatmaps in the bottom left indicate pairwise preservation values of modules between two tissue types. Bar charts in the upper right corner indicate preservation of module connectivity between two tissue networks. Values above bar charts represent the preservation density (D) between two tissue networks.

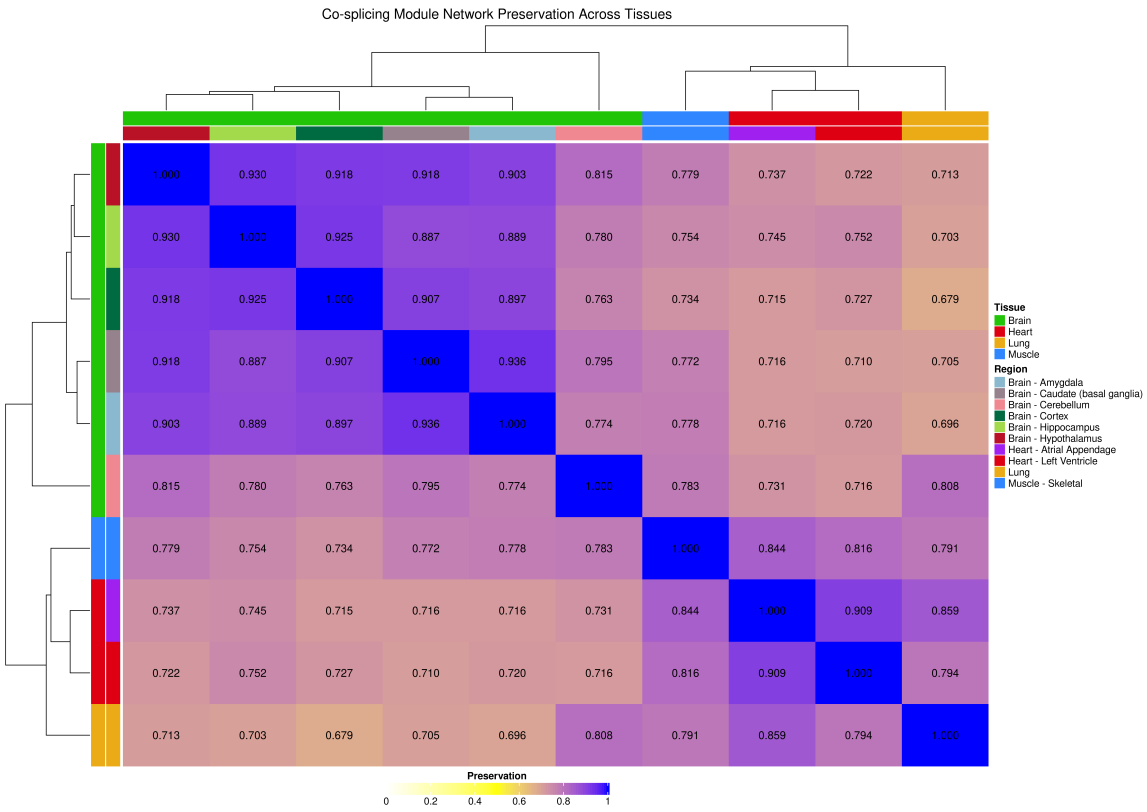

Figure 11: Co-splicing Network Preservation Across Tissues. Values in the heatmap represent module network density preservation between all pairs of tissue networks. Blue values indicate higher levels of network preservation. Clustering of preservation values was performed by subtracting the density preservation value by 1 and performing hierarchical clustering using euclidean distance.

may be contributing in regulation. Enriched terms shared across multiple co-splicing modules include regulation of mRNA stability, translation initiation, and RNA splicing (FDR < 0.01, minimum of 10 genes). Module-specific terms include neutrophil activation and immune response which were both highly enriched in the black co-splicing module, Wnt signaling which was enriched in the green module, and regulation of apoptotic signaling in the blue module.

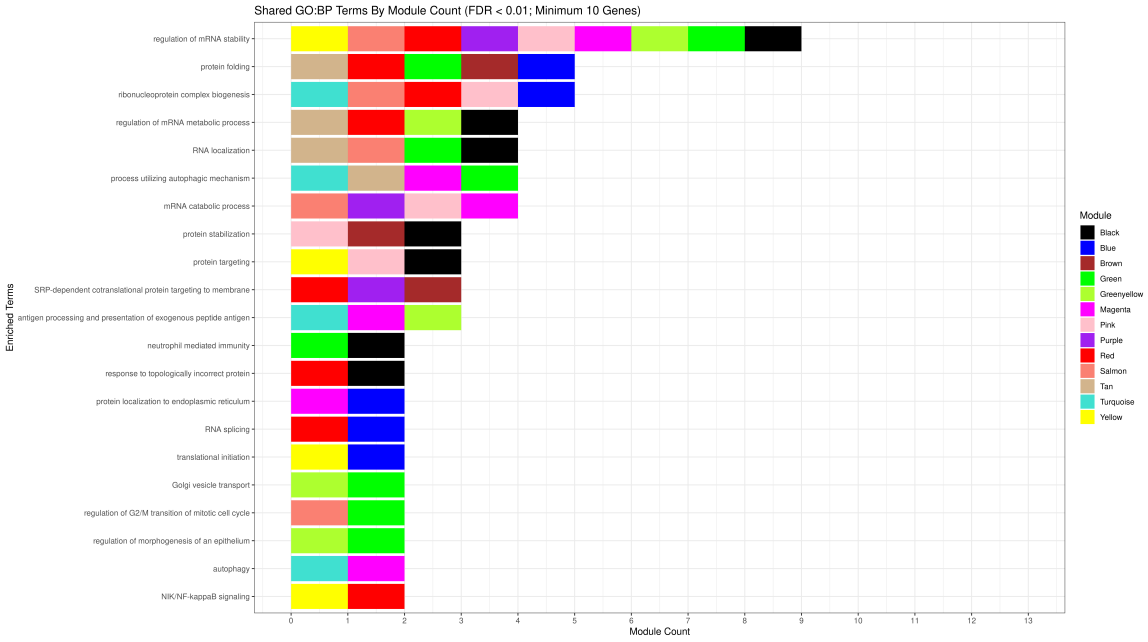

Figure 12: Module Counts of Enriched GO Biological Process Terms From Consensus GTEx Modules.

Co-splicing modules were also enriched for numerous Reactome pathways with several pathways being shared across modules (Figure 14). Infectious disease, regulation of expression of SLITs and ROBOs, and translation were found enriched in at least ten co-splicing modules. Co-splicing modules were also enriched for several module-specific pathways including the turquoise module which was highly enriched for SUMOylation pathways and the magenta module highly enriched for NOTCH signaling (Figure 15).

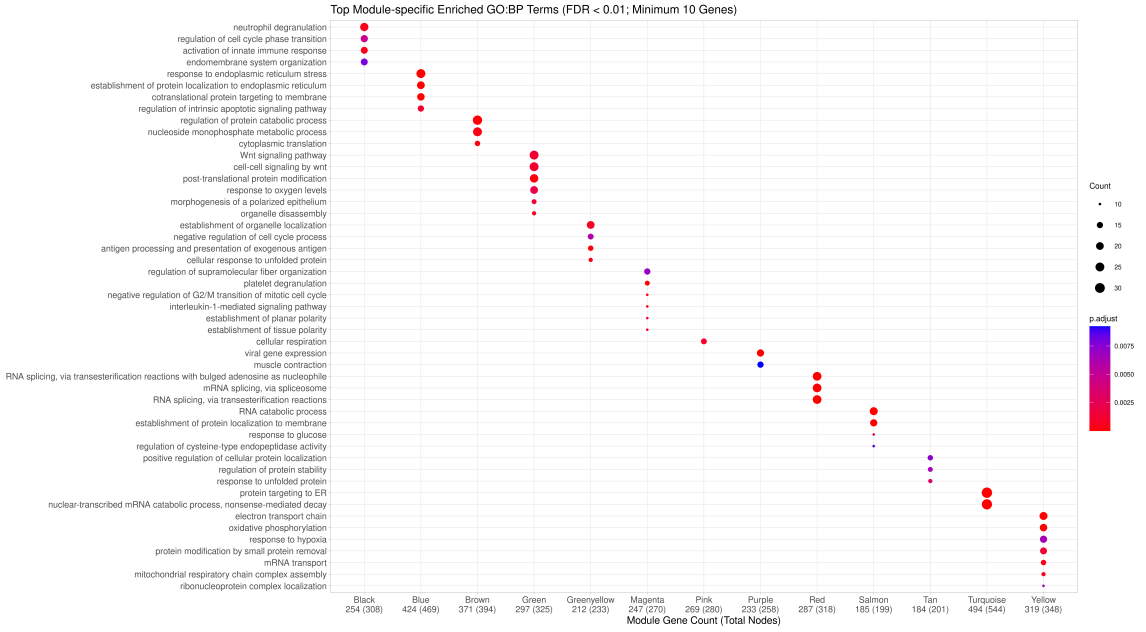

Figure 13: Module-specific Enriched GO Biological Process Terms From Consensus GTEx Modules.

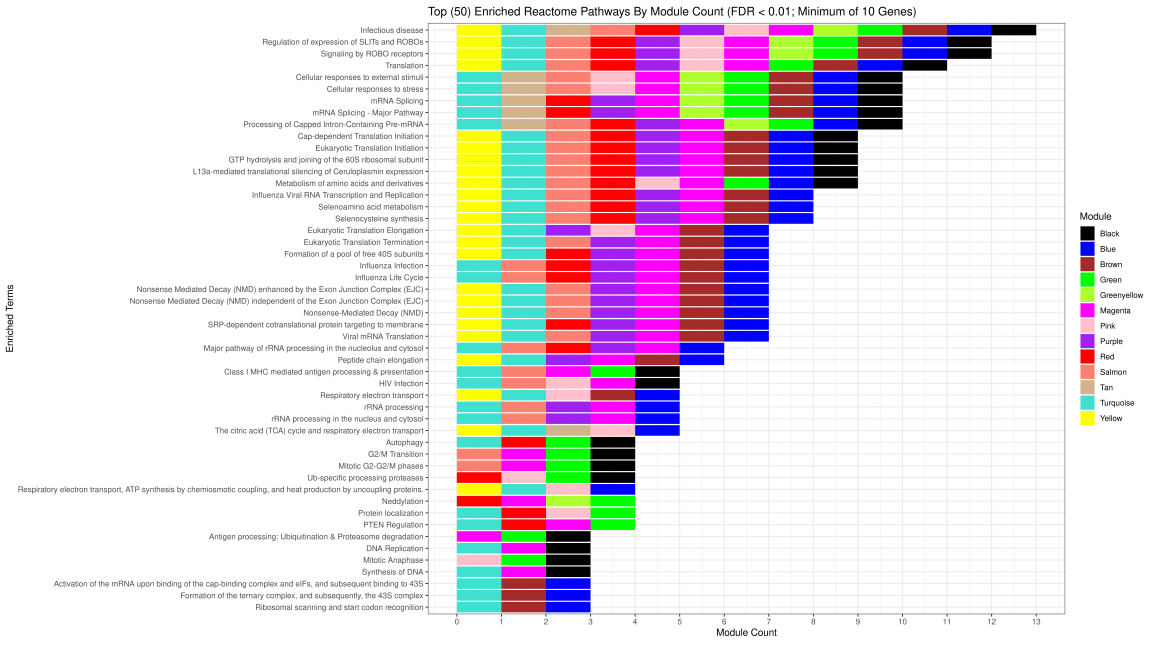

Figure 14: Module Counts of Enriched Reactome Pathways From Consensus GTEx Modules.

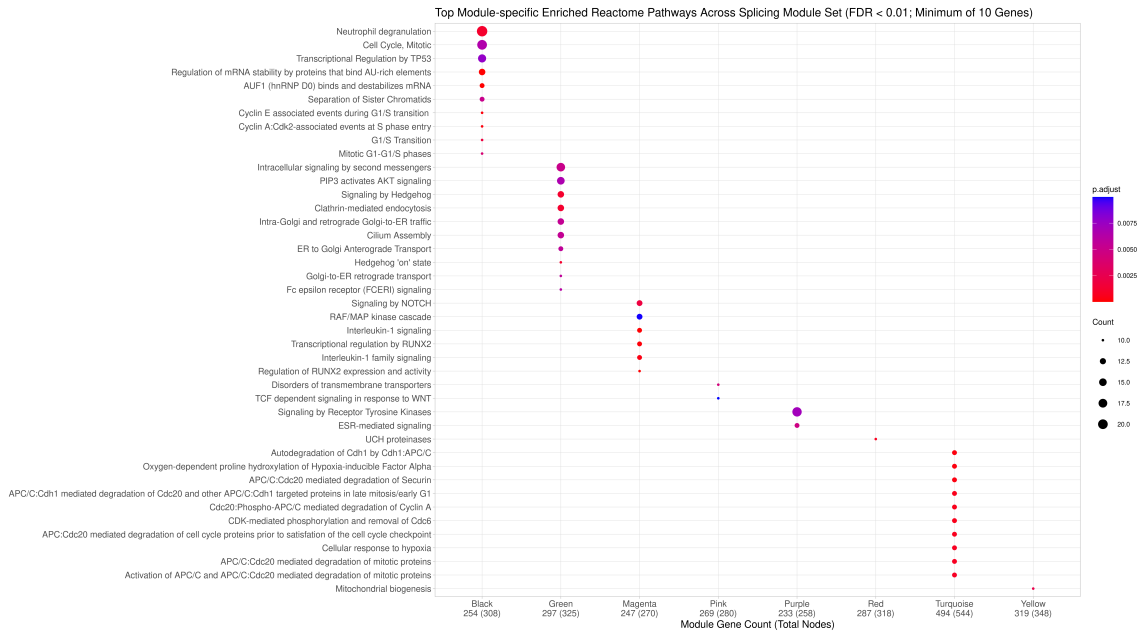

Figure 15: Module-specific Enriched Reactome Pathways From Consensus GTEx Modules.

Characterization of Intra-modular Hub Nodes Across Tissues

Network hubs represent nodes having the highest levels of connectivity within a network. Node connectivity (or weighted degree) is the sum of all edge weights for a given node to all nodes of the network. In a module-based network analysis we focus on intra-modular connectivity of each node which describes the relative connectivity of each node to all nodes within the same module. Intra-modular hub nodes tend to be the most important and functionally relevant nodes for a given module and may include nodes that regulate many of the other nodes within a module. Node module membership (kME) is defined as the correlation of each node with the module splicing value and consistent with gene co-expression studies we found intra-modular connectivity to be highly correlated with module membership across all tissues, indicating that the most central nodes of each module are driving the within module variation (Figure 16).

Intra-modular connectivity of a given node may occur in either a conserved or

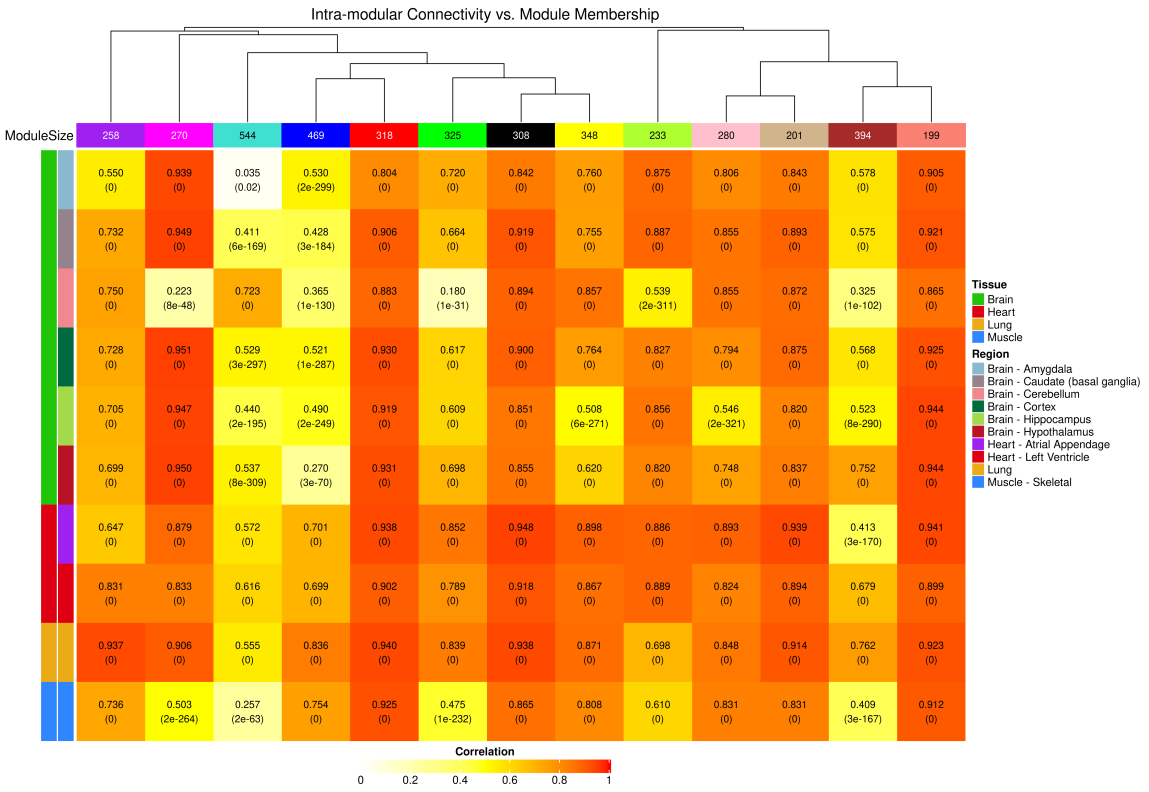

Figure 16: Correlation of Node Intra-modular Connectivity and Module Membership Across Tissue and Consensus Modules.

tissue-specific manner. We first measured the concordance of intra-modular connectivity between the ten co-splicing tissue networks. As expected, similar tissue types were found to contain higher concordance of intra-modular connectivity (Figure 17).

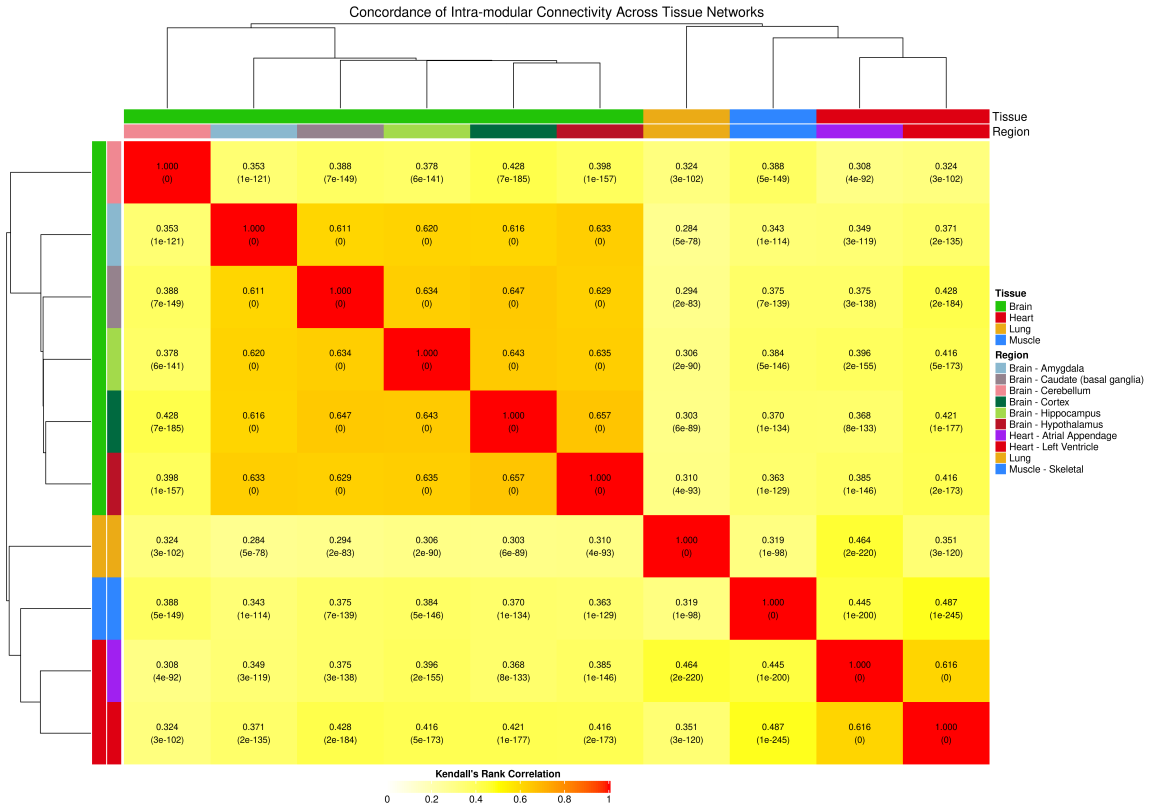

Figure 17: Concordance of Intra-modular Connectivity Across GTEx Networks.

We then identified intra-module hub nodes across each tissue network. We defined intra-modular hubs as nodes found within the top 5% of each module based on intra-modular connectivity, resulting in 213 nodes selected across the 13 consensus modules for each tissue type. Similar to our analysis of the concordance of node degree across tissues, similar tissue types contained a larger overlap of module hub nodes than non-related tissues (Figure 18).

GO enrichment analysis of intra-modular hub genes from each tissue revealed significant enrichment for a variety of terms. Similar to module enrichment,

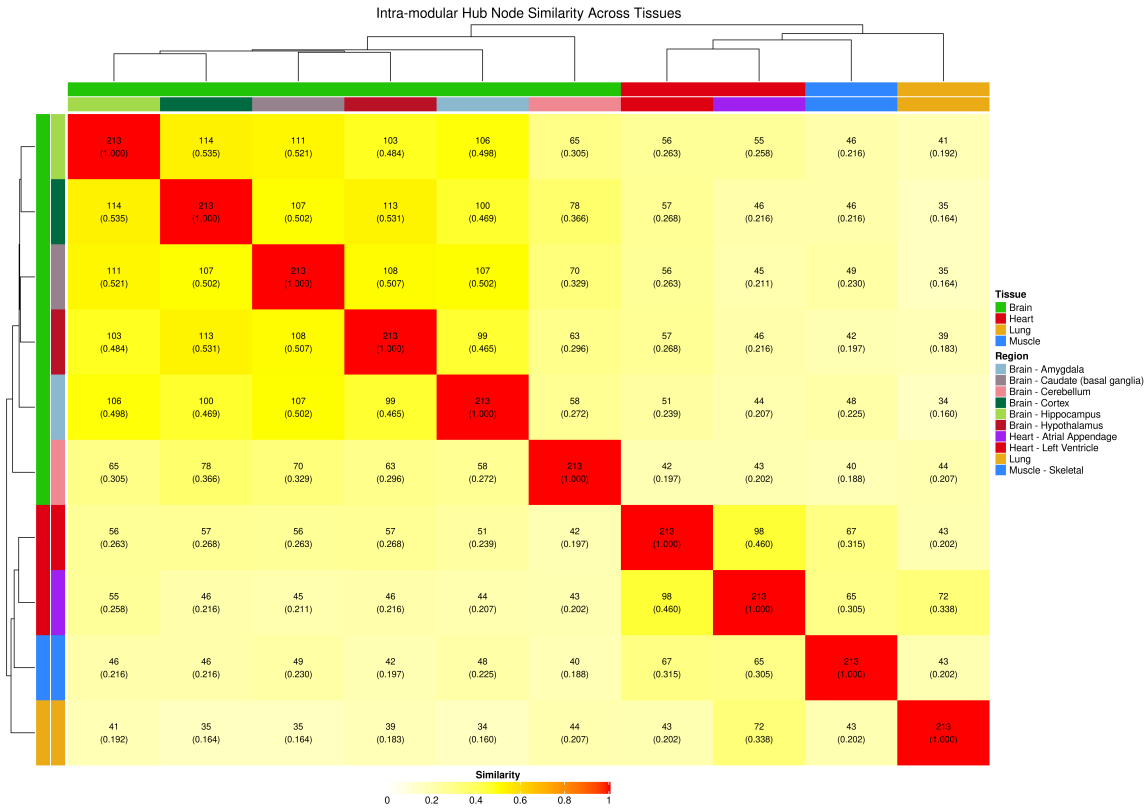

Figure 18: Intra-modular Hub Node Similarity Across GTEx Networks.

enrichment of GO terms in tissue hubs occurred in both a shared and tissue-specific manner (Figures 19 and 20).

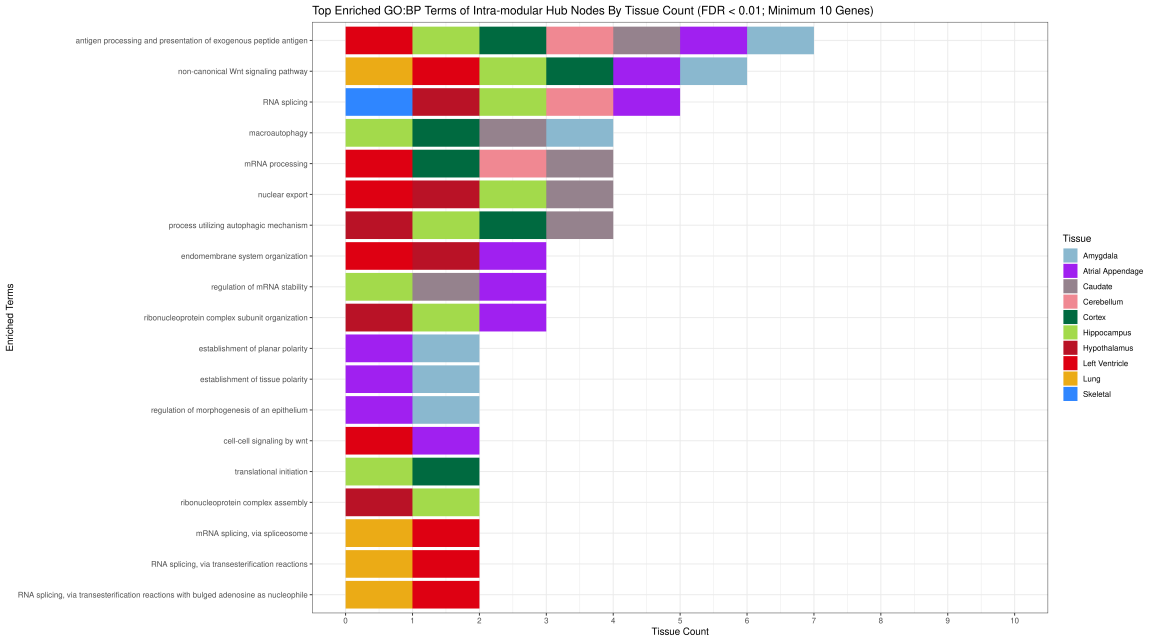

Figure 19: Enriched GO Biological Process Terms of Hub Nodes Shared Across Tissues.

As hub nodes may often contain genes that potentially regulate other nodes within pathways, we looked to see if module hubs were enriched for RNA splicing and RNA binding. Indeed, we found significant enrichment of RNA splicing and RNA binding related GO terms in all ten tissue types and many of the intra-modular hub genes were known splicing factors and RNA binding proteins (Figure 21). . These results indicate that alternative splicing of intra-modular hub genes may serve as a mechanism for regulating splicing of other genes.

We further characterized the level of tissue-specific hubs in each co-splicing module. We defined tissue-specific module hubs as nodes being in the top 5% of intra-modular connectivity in one tissue and not in the top 10% in the remaining nine tissues. All ten tissue types were found to contain tissue-specific hubs in each of the 13 consensus modules with amygdala (brain) showing the highest number of tissue-specific module hubs across co-splicing modules (Figure 22). Eight tissues contained tissue-specific

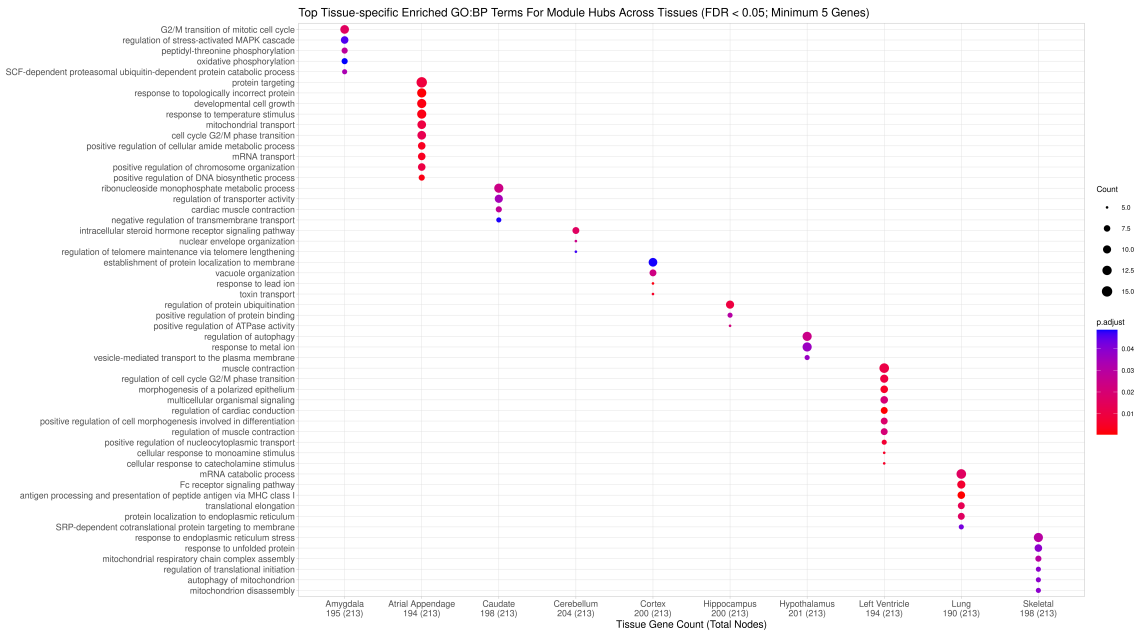

Figure 20: Tissue-specific Enriched GO Biological Process Terms of Hub Nodes

hubs enriched for multiple known GO terms. Five tissues (amygdala, atrial appendage, caudate, cerebellum, and lung) were enriched for unique GO terms including regulation of hematopoietic stem cell differentiation which was enriched in amygdala-specific hubs and intracellular steroid hormone receptor signaling pathway which was enriched in cerebellum hub nodes (Figure 23).

Differential Splicing of Co-splicing Modules Across Tissues

Here we characterized module level splicing differences across tissues types. This analysis looks at splicing variants that are highly co-spliced across tissues but contain tissue-specific module-level splicing variation. These tissue-specific differences in module splicing may reveal co-spliced complex splicing variants that help regulate tissue-specific cellular phenotypes. To characterize tissue-specific splicing variation, we summarize the co-splicing modules using all tissue samples during SVD. Here, module splicing values capture the across-tissue splicing variation of SVRs within each module. We find that all consensus co-splicing modules are moderately to highly

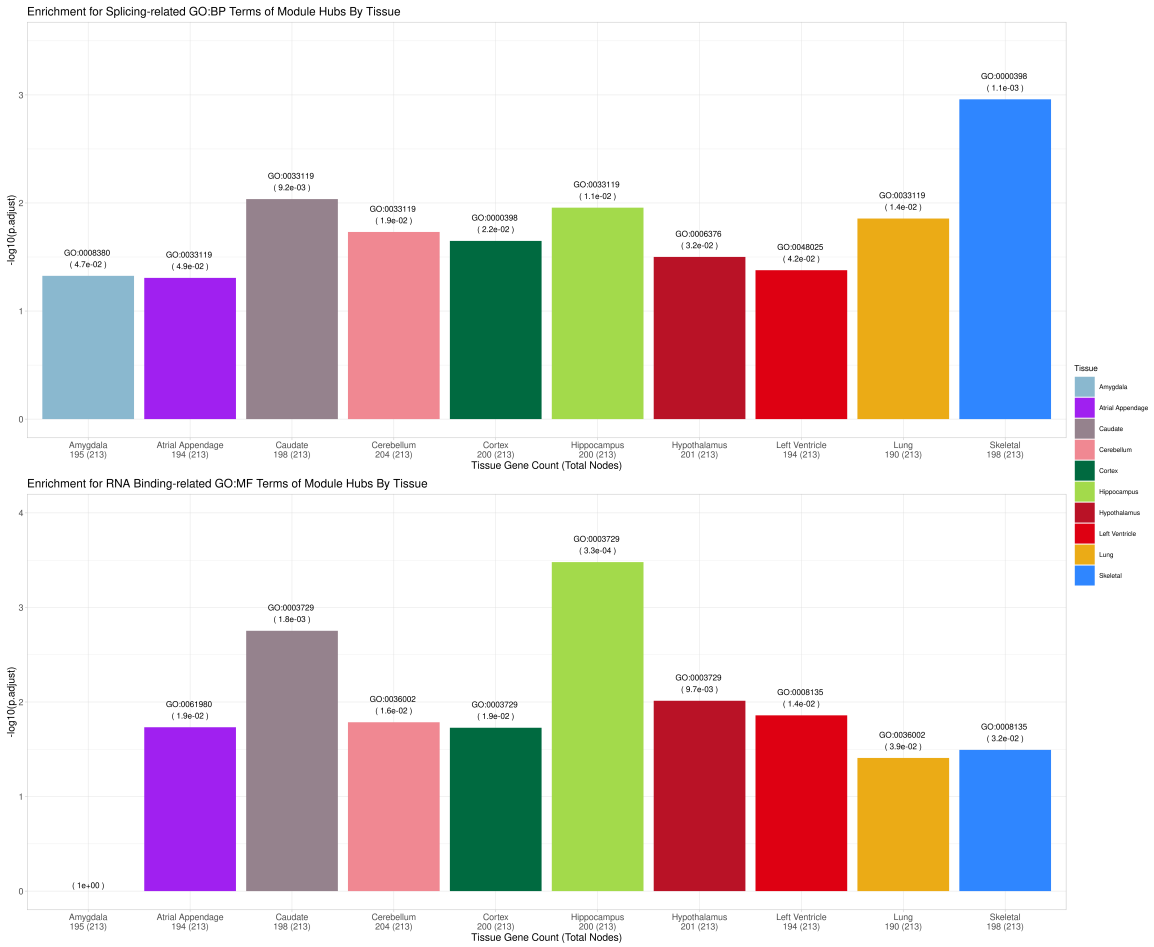

Figure 21: Enrichment for Splicing Regulators in Module Hubs Across GTEx Networks. Bar charts indicate the adjusted p-value for each tissue.. Top) Enrichment of RNA-splicing GO terms using hub nodes from each tissue network. The most significantly enriched GO term for RNA-splicing is indicated above each tissue. Bottom) Enrichment of RNA-binding GO terms using hub nodes from each tissue network. The most significantly enriched GO term for RNA-binding is indicated above each tissue.

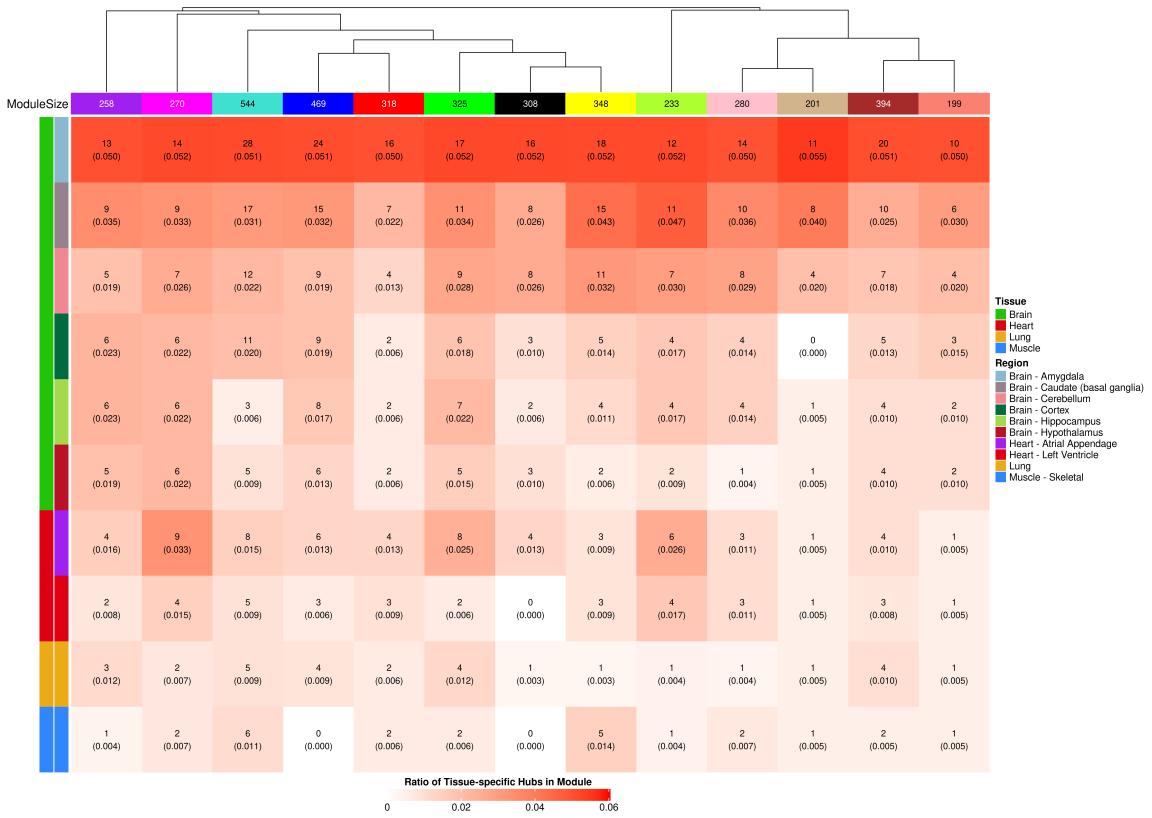

Figure 22: Ratio of Tissue-specific Hubs From GTEx Networks.

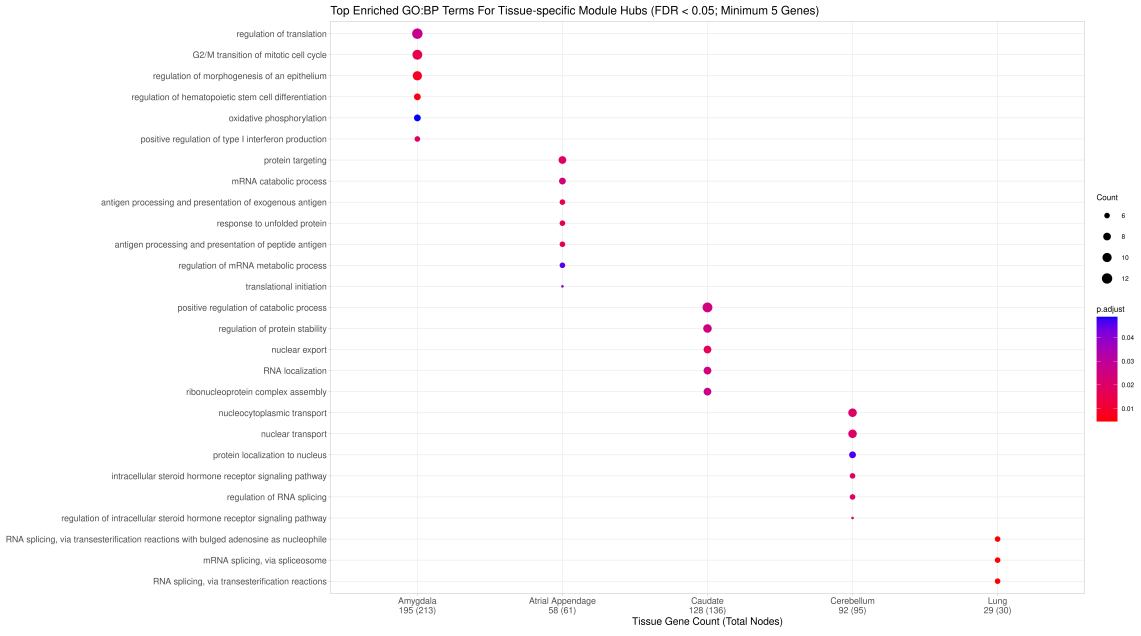

Figure 23: Enriched GO Biological Process Terms of Tissue-specific Module Hubs.

correlated with at least one tissue type and are differentially spliced across tissue groups (Figure 24).

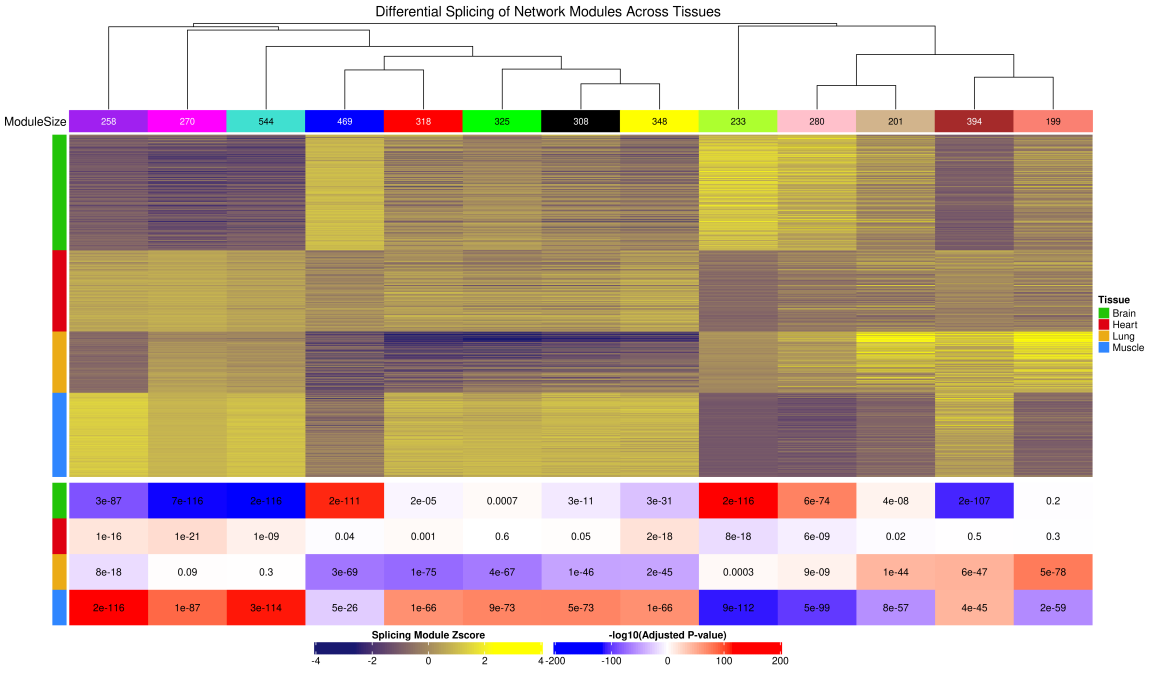

Figure 24: Differential Splicing of Network Modules Across Tissue Groups. Modules are ordered based on the consensus module hierarchy. The top heatmap indicates the module splicing values for each sample with samples grouped by tissue type. The bottom heatmap indicates adjusted p-values from a Wilcoxon test between each tissue group against the remaining tissues in the dataset.

We also characterized module level splicing differences within related tissue groups. Here we looked at splicing variants that are highly co-spliced across tissues, but contain tissue subtype-specific module-level splicing variation. Similar to the previous analysis, this analysis may reveal co-spliced complex splicing variants that help regulate subtype-specific cellular phenotypes. Figure 25 shows a heatmap of module splicing values computed using SVD across the six brain regions. Similar to our previous results of module and node-based network properties, the cerebellum brain tissue region exhibits the most distinctive module-level splicing.

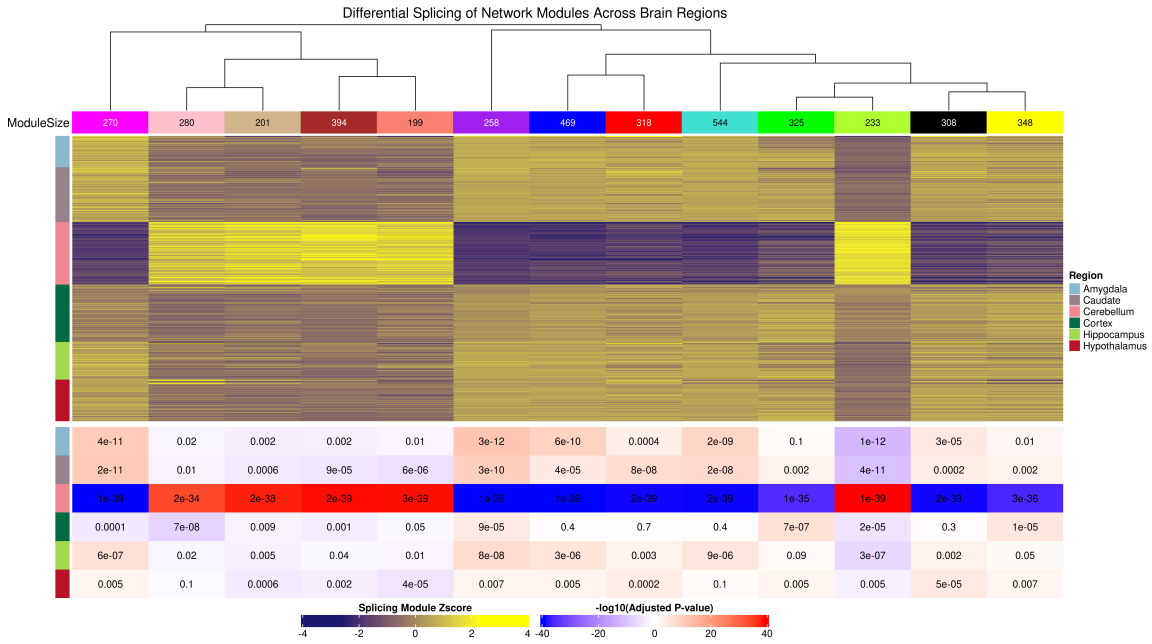

Figure 25: Differential Splicing of Network Modules Across Brain Regions. Modules are ordered based on the consensus module hierarchy between the six brain regions. The top heatmap indicates the module splicing values for each sample with samples grouped by brain region. The bottom heatmap indicates adjusted p-values from a Wilcoxon test between each brain region against the remaining brain tissue types.

Discussion

Using RNA-seq data from the GTEx project, we inferred co-splicing networks on a tissue-specific basis and characterized coordinated alternative splicing variation across tissues in the form of network modules. By formulating SVRs from gene splicing graphs we were able to include complex alternative splicing variants in our networks, thus capturing the complexity of potential splicing variation in a comprehensive and interpretable way. A module-based analysis of co-splicing inherently serves as a useful data reduction technique for transcriptome-wide characterization of splicing regulation, but also has the potential to identify biologically meaningful pathways for which splicing may be involved. Consensus modules represent groups of complex splicing variants that are highly co-spliced in a well preserved manner across multiple systems. By identifying a set of consensus modules, we were able to characterize variation in co-splicing across tissue types at both an inter- and intra-modular level.

The differential co-splicing analysis served as an inter-modular characterization of co-splicing. In this analysis splicing levels of the 13 consensus modules were measured on a per-tissue basis and thus captured the within tissue variation of each module. Hierarchical clustering of the consensus modules was performed for each tissue and revealed tissue-specific characteristics of the module networks similar to that of the original co-splicing networks of SVRs. Inter-modular characteristics (e.g. edge weights between modules) were well preserved across tissues, however tissue-specific differences were clearly present with groups of related tissues having similar network-based characteristics than those of non-related tissues. These results provide a higher-order representation of how alternative splicing is associated with tissue-specific phenotypes. Further, these results also serve as a proof of concept for the use of a network-based analysis of splicing given that the magnitude of variation across co-splicingnetworks fits within the expected degree of variation between related

and non-related tissues.

The intra-modular analyses of co-splicing were also consistent with the across tissue variation observed in the inter-modular analysis. Given the module-focused analysis of co-splicing, we characterized connectivity patterns of the SVRs in relation to their assigned co-splicing modules. Intra-modular node connectivities clearly exhibited tissue-specific differences, but much like the inter-modular connectivities, related tissues exhibited higher similarities of node degree than non-related tissues. These results again promote the utility of this network-based approach for splicnig while also characterizing co-splicing of complex splicing variants across tissues.

Previous co-expression studies have revealed hub nodes to be centrally important in the regulation of other genes within a network. Although de novo inference methods utilizing correlation can not prove direct regularity interactions, subsetting of network nodes exhibiting relatively high connectivity may reveal potential regulatory genes or genes driving a particular phenotype (Saha et al. 2017). With a module-based analysis of co-splicing we characterized splicing variants that are highly connected within co-splicing modules across the ten tissue networks. Given that node connectivity is measured on a per tissue basis, intra-modular selected hub nodes occurred on both a shared and tissue-specific basis with related tissues sharing more hub nodes than non-related tissues. Across tissue types, hub nodes were significantly enriched for GO terms related to RNA splicing and RNA binding. This is consistent with the fact that hub nodes are likely to contain genes involved regulating other genes within the network and in this case, the hub genes may undergo splicing changes of their own to help regulate the splicing of other variants within the module.

Many of the genes in which the splicing hubs were derived were indeed known regulators of RNA splicing. For example, a splicing variant of the Heterogeneous Nuclear Ribonucleoprotein A2/B1 (HNRNPA2B1) gene belonging to the red

consensus module was found to be a top hub node across all ten tissue types. HNRNPA2B1 is known to be an RNA binding gene highly involved in mRNA processing including regulating efficiency of pre-mRNA splicing (Alarcón et al. 2015). Direct splicing variant targets for which HNRNPA2B1 may contribute to post-transcriptional processing cannot be directly proven without further study, but it is likely that HNRNPA2B1 may contribute to the processing of multiple splicing variants throughout the network, particularly within the red co-splicing module. It is also important to note that the splicing variant hub derived from HNRNPA2B1 was in the form of a complex splicing event consisting of a cluster of interchanging exons and retained introns. The SVR formulation of this variant allowed for it to be included in the co-splicing network in a comprehensive and non-redundant manner, thus allowing for it's coordination with other splicing variants to be better characterized.

A common application of a module-based co-expression analysis is to summarize the expression of each module using SVD (e.g. module eigengene) and test if any modules are significantly correlated with a trait of interest. The phenotype of this use-case are different tissue types and while the differential co-splicing analysis primarily focused on inter-modular variation between tissues, we can also characterize how individual modules are differentially spliced across tissues. The primary difference in this analysis is the summarization method use for the module splicing values. Unlike the differential co-splicing analysis, SVD is performed on all tissue samples of the dataset and thus captures the variation of splicing within each module across tissue types. The analysis identified that although the modules remain highly co-spliced across the tissues (as observed with the module quality Zscores), they undergo variation in splicing in a tissue-specific manner. Indeed, every consensus module was found to be differnetially spliced in at least one tissue group. This application is analogous to an approach of measuring module significance in relationship to a phenotype of interest and can be applied to a variety of applications for characterizing splicing changes

across phenotypes on a systems-level. We further demonstrate this approach in the next chapter by identifying co-splicing modules that are highly correlated with drug response in acute myeloid leukemia.

Methods & Materials for Co-splicing Inference Across Human Tissues

Pre-processing of GTEx RNA-Seq Samples Aligned RNA-Seq samples from ten human tissues types were downloaded from the Genotype-Tissue Expression (GTEx) consortium v7 (The GTEx Consortium 2015). The ten tissue types included six brain regions (amygdala, caudate (basal ganglia), cerebellum, cortex, hippocampus, hypothalamus), two heart tissues (atrial appendage, left ventricle), lung tissue and skeletal muscle tissue. Samples were collected, sequenced, and processed according to the online GTEx protocol from which the v7 were downloaded. Briefly, samples were sequenced using seventy-six base pair (bp) paired-end reads with a non-stranded protocol. Splice-aware alignment was performed with STAR using the GRCh37 human reference genome and Ensembl build 74 gene models (Dobin et al. 2013). Only GTEx RNA-Seq samples containing a minimum of 5 million junction spanning read alignments were included in the analysis. Duplicate donor RNA-seq samples of the same tissue-subtype were removed, keeping the donor sample with the highest number of junction spanning read alignments. The final analysis set consisted of 1,621 samples from 492 donors, ages 20-79 across the ten tissue types.

LSV Quantification Annotation and quantification of complex splicing variants from GTEx samples was performed using the Modeling Alternative Junction Inclusion Quantification (MAJIQ) computational framework (Vaquero-Garcia et al. 2016). Splice-graphs were constructed for each gene in both datasets using the MAJIQ build function. GTEx splice-graphs were constructed using the following parameters:

--min-experiments .1 --min-intronic-cov 1.5 --minreads 5 --minpos 3

In order to simplify the resulting splice-graphs and improve splicing quantification, a custom script was used to remove LSV junctions that were covered by a low number of junction spanning read alignments in each dataset before LSV PSI quantification. We removed junctions that contained fewer than 5 reads in more than 1% of samples in the GTEx dataset. After filtering out low coverage junctions, we removed LSVs from eachs plice-graph if they no longer contained at least two LSVs or if the total reads from all junctionswas below 8 in more than 2% of samples. This results in 15,216 LSVs quantified across the 1,621 RNA-seq samples. LSVs from the simplified splice-graphs were then quantified for each sample individually using the MAJIQ PSI function. For a given LSV of a given sample, the PSI of each junction was quantified if at least one junction contained a minimum of 8 reads across 3 unique positions. Only LSVs meeting the minimum read thresholds in all 1,621 samples were used for downstream analysis, resulting in 6,427 LSVs. Given that each LSV contains two or more junctions with each junction's PSI representing it's relative splicing level to that of all junctions of the LSV, a single junction was used to represent each LSV for unsupervised learning methods such as that of de novo network inference. For each LSV of the two datasets, we selected the junction that had the greatest variance across samples to represent it's splicing level for downstream analysis.

Adjusting for Confounding Effects Confounding artifacts were estimated and adjusted for in each dataset using an approach based on (Parsana et al. 2019) for estimating and removing confounding artifacts prior to gene co-expression network analysis. Using the SVA package in R, hidden confounders were estimated in the GTEx LSV set by estimating the number of confounding principal components while keeping tissue type as a fixed outcome. We estimated no confounding principal components in the GTEx LSV dataset with respect to tissue type.

Constructing Co-splicing Networks Across Tissues Signed SVR networks for each of the ten tissue subtypes were constructed using biweight mid-correlations. In order to account for spurious correlations, soft-thresholding was used by raising each network by a constant (β) which promotes strong edge weights and demotes weak ones. The value of β for soft-thresholding was chosen for each network by selecting β values that resulted in an approximate scale-free topology. For simplicity, a β of 7 was chosen for all ten tissue networks, resulting in all co-splicing networks having an R^2 scale-free fit > 0.9. After constructing each network a topological overlap transformation $TOM(A)$ was then performed on each adjacency matrix. Topological overlap matrices characterize the relationships between each node while taking into consideration their shared relationships with other nodes of the network. In order to identify a consensus co-splicing module set for differential co-splicing analysis across tissue types, a consensus topological overlap matrix was constructed using all ten tissue topological overlap matrices. The consensus network across all tissues is defined as the minimum quantile across all network topological overlap matrices and thus represents relationships between nodes that are agreed upon to some degree across all tissue types.

Clustering and Formulation of Consensus Modules For identifying modules of highly co-spliced splicing variants (SVRs) across tissue co-splicing networks, we utilized an ensemble approach of clustering methods performed on the resulting topological overlap matrices. For constructing the final set of modules we performed spherical clustering on the resulting consensus tissue TOM matrix. Using the consensus TOM as input, spherical clustering is performed by first transforming the symmetric matrix into a degree-normalized Laplacian matrix defined as $L = D - A$, where D is a matrix of the degrees of A and A is the initial adjacency matrix (in this case a topological overlap matrix). Then the top k eigenvectors are computed on the

resulting graph Laplacian and k-means clustering is performed using the resulting eigenvectors with k being the resulting number of clusters. Since k is unknown beforehand, we follow a similar approach as (Botia et al. 2017) and first perform hierarchical clustering followed by dynamic tree cutting to determine an initial count of network modules. Here, hierarchical clustering is performed using average linkage distance on a topological overlap dissimilarity matrix, defined as $1 - TOM(A)$. We set the deepSplit parameter to 2 and pamStage to FALSE. Pam cut height was determined based on the 99% level of the resulting dendrogram, resulting in a cut height of 98.5. After the initial modules were constructed from the hierarchical clustering step, the module splicing values were computed and closely connected modules were merged with a merging threshold of 0.25. 15 modules remained after hierarchical clustering and merging and k = 15 was used for spectral clustering. During the k-means clustering of eigenvectors step of spectral clustering, 100 different initial centroid sets were chosen. To ensure reproducibility due to the possibility of varying cluster sets from random initial starting centroids, we repeated the spectral clustering step through 101 iterations selecting the final module sets using majority voting. Finally, module splicing values were again computed on the resulting module set and any closely related modules were again merged. After merging closely related modules a set of 13 modules remained and were used for further analysis.

Comparison of Intra-modular Node Degree Across GTEx Tissues

Similarity of splicing networks across tissues was analyzed using the tissue-specific intra-modular node degrees of each splicing variant. The relative intra-modular degree of each node is calculated as the sum of (weighted) edges to all other nodes belonging to the same modules and is scaled across all nodes of the same module (a module's most connected node has an intra-modular connectivity of 1). Kendall's rank correlations were then computed pairwise for all tissues and hierarchical clustering

was performed using the resulting correlation values $(1 - cor)$ as the distance measure.

Differential Splicing of Modules Across Tissue Types To test for differential splicing of modules across GTEx tissues, module splicing values were recomputed by performing SVD on the set of each module's splice variants across all tissue samples simultaneously. Here, each module's splicing value (i.e. the 1st principal component) captures variance in splicing across all tissue types. After computing module splicing values, for each module we performed a two sided Wilcoxon rank sum test of each tissue versus the remaining nine tissue types. P-values were adjusted for multiple testing using Benjamini-Hochberg method.

Functional Enrichment of Modules and Hub Nodes Functional enrichment of GO terms and Reactome pathways was performed using the clusterProfiler R software library. Due to the low number of unique genes in each co-splicing network, we used the total set of expressed genes as background after removing genes with low read counts. Enrichment was performed using a hypergeometric test and p-values were Benjamini-Hochberg adjusted for multiple testing. In addition, redundant GO terms were removed following GO enrichment based on GO tree similarity distance.

Data Visualization All dendrograms and heatmaps were created using the *ComplexHeatmap* R package (Gu, Eils, and Schlesner 2016). Buzzsaw plots were created using the *circlize* R package (Gu et al. 2014).

Chapter 4: Co-splicing Network Inference of Drug Response in AML (Use case)

Chapter 4 details the second use case for the co-splicing network inference algorithm in which we infer a co-splicing network of AML and identify co-splicing modules predictive of drug response for multiple small molecule inhibitors.

Introduction: Genomic Landscape of Drug Response in AML

AML Acute myeloid leukemia (AML) is characterized by the excessive proliferation of undifferentiated myeloid stem cells and accounts for over 10,000 cancer-related deaths in the U.S. annually. Over 20,000 people are diagnosed with AML each year in the U.S. and the majority of patients will experience relapse after initial treatment followed by progression into a more therapy-resistant cancer. Numerous subtypes of AML exist and cytogenetic profiles can help stratify AML patients into groups of favorable, intermediate, and adverse outcomes (Khwaja et al. 2016). Due to the heterogeneity of AML, however, accurate prognosis has long been difficult. Recent studies have identified various mutational aberrations that can significantly impact patient prognosis and therapies targeting such mutational events have shown to be a promising solution for improving the current standard of care (Tyner et al. 2018).

Cytogenetics of AML Cytogenetic aberrations are highly significant in regards to AML development, diagnosis, prognosis, and treatment decisions. Around half of all AML cases begin as a result of one of the recurrent AML karyotypes (Khwaja et al. 2016). Multiple recurrent cytogenetic events including inversions, deletions and translocations can lead to AML development and often result in the production of AML fusion genes. Some examples of these well characterized abnormalities include the RUNx1-RUNX1T1 fusion gene from the t(8:21)(q22;q22) translocation, the

CBFB-MYH11 fusion gene resulting from either the inv(16)(p13.1q22) inversion or t(16;16)(p13.1;q22) translocation, and the PML-RARA fusion gene from the t(15;17)(q24;q21) translocation. The presence of these specific cytogenetic events help define various AML subtypes. For example, the presence of the PML-RARA fusion defines the acute promyelocytic leukemia (APL) subclass which can be further divided into two subtypes of its own: hypergranular/typical APL and microgranular/hypogranular APL (Khwaja et al. 2016; Hackl, Astanina, and Wieser 2017).

Somatic mutations Somatic mutations are found in nearly all AML patients regardless of the presence of one of the recurrent cytogenetic events (Khwaja et al. 2016; Tyner et al. 2018). The presence of various recurrent mutations in AML also help define the various AML subtypes including AML with mutated NPM1 and mutated CEBPA (De Kouchkovsky and Abdul-Hay 2016). Previous research has suggested that leukemia development requires the presence of mutations from two class types, one that results in deregulation of normal proliferative signaling and the other to deregulate hematopoietic differentiation. More recent studies involving next generation sequencing (NGS) have pointed to other mutation classes of AML including mutations in genes that function as epigenetic modifiers, tumor suppressors, and genes involved in alternative splicing regulation such as splicing factor genes and components of the spliceosome complex (Khwaja et al. 2016).

Recent studies utilizing NGS technology for genomic characterization have revealed that the landscape of somatic mutations in AML is extremely complex and heterogeneous across AML patients (Tyner et al. 2018). Over 1,000 somatic mutations have been identified in AML, however, only around two dozen mutations are known recurrent mutations. Even the most common recurrent mutations in AML are found in a only small subset of patients across AML cohorts. Most AML patients

typically have around a dozen mutated genes, and adult patients have more mutations than children. The exact combination of which genes are found mutated in a given patient tends to vary considerably, but some cases of mutual exclusivity do occur (Khwaja et al. 2016). Research has shown that different mutational aberrations can a have a significant impact on the patient prognosis and response to treatment (Khwaja et al. 2016; Tyner et al. 2018)

Therapeutic Strategies for AML The standard treatment of care for patients with AML has remained relatively consistent over the last few decades. The choice of treatment strategy is usually determined based on patient age and prognosis. Patients under the age of 60 are typically given a more aggressive form of induction chemotherapy. This will often consist of either the "7 + 3" or "10 + 3" strategy in which cytarabine is given for 7 to 10 days along with athracycline for 3 days (Khwaja et al. 2016). The main goal is to achieve complete remission (CR) and around 70% of patients under the age of 60 achieve CR. Therapeutic strategies for elderly patients are less straightforward, and often depend on the general fitness of the patient because intensive chemotherapy may do more harm than good. Elderly patients deemed unfit for standard induction therapy may receive low-dose cytarabine at a less frequent rate than standard therapy (Khwaja et al. 2016). The presence of specific karyotypes should also be considered when choosing the appropriate induction therapy regimen. Patient relapse and minimal residual disease are common after treatment for AML, and consolidation therapy may need to be applied. Selection of the appropriate consolidation therapy approach is based on patient prognosis and may include additional chemotherapy or the use of hematopoietic stem cell transplants (Khwaja et al. 2016).

After decades of stagnation for AML therapeutic options, novel therapeutic strategies involving targeted therapeutic agents have recently been developed. Development of

targeted strategies is in large part due to better characterization of deregulated cellular pathways and genomic profiles in AML (De Kouchkovsky and Abdul-Hay 2016; Tyner et al. 2018). This characterization helps researchers better understand the functional consequences of recurrent mutational events in AML and target the deregulated cellular pathways. Despite recent advancements and promising outlooks for targeted approaches, the heterogeneous nature of AML leaves many patients showing resistance to initial treatment. Further, many studies have demonstrated that patients who relapse after treatment are found to have additional mutational events not present at initial diagnosis and such mutational events may contribute to developing therapeutic resistance (Tyner et al. 2018). Thus, it is essential to better characterize the genetic mechanisms in which cancer cells develop therapeutic resistance in AML.

Alternative Splicing in AML Similar to many other cancers, numerous alternative splicing events have been observed in AML and evidence has supported the notion that aberrant splicing may be one of the mechanisms towards AML progression. A recent study found that over 30% of genes were differentially spliced in AML patients compared to normal samples (Necochea-Campion et al. 2016). Several aberrant splicing events in AML have been observed in known oncogenes and tumor suppressor genes. For example, the FLT3 gene is found differentially spliced in 50% of AML cases and the FLT3-Va splice variant is shown to be upregulated in patients who have relapsed after initial treatment. In additional to aberrantly spliced genes, several splicing factor genes may be deregulated or have somatic mutations in AML patient samples (Tyner et al. 2018; B. D. Wang and Lee 2018; Zhou and Chng 2017). Many of these altered splicing factor genes have potential prognostic significance as well (Necochea-Campion et al. 2016; B. D. Wang and Lee 2018). Aberrant splicing may also contribute to the formulation and maintenance of leukemia stem cells

(LSCs) (B. D. Wang and Lee 2018; Crews et al. 2016; Holm et al. 2015). LSCs are often believed to be one of the key mechanisms for promoting patient relapse and the development of therapy-resistance in AML (X. Wang, Huang, and Chen 2017). As alternative splicing is a fundamental post-transcriptional regulatory program for stem cell maintenance and differentiation, it is not surprising that recent studies have demonstrated a plausible role for deregulated splicing contributing to LSCs' ability to self-renew, avoid therapy, and promote relapse in AML (Zhou and Chng 2017). The full extent of aberrant splicing regulation in AML drug resistance, however, is not fully characterized.

Results

AML Splicing Module Set & Module Quality Statistics

We inferred a network of co-splicing modules in order to identify network-based splicing signatures of drug response in AML. We formulated 8,207 SVRs using 13,354 LSVs quantified from 410 AML RNA-seq samples. Using the 8,207 SVRs, we constructed a signed network using biweight mid-correlations and raised the initial adjacency matrix by a power of 5. After computing a toplogical overlap matrix from the power transformed adjacency matrix, co-splicing modules were identified using hierarchical and spectral clustering, resulting in a set of 13 co-splicing modules containing 378-1,321 SVRs (Figure 27).

Of the 3,742 genes represented in the AML co-splicing network, almost half of the genes are represented by more than one splicing variant (Figure 26A). Therefore, splicing variants for a single gene may be found in more than one module. Similarly, a single gene may have more than one splicing variant present within the same module. Figures 26B and C show the distributions of genes across modules as represented by their splicing variants. Of the 3,742 unique genes found within the

co-splicing network, over half of the genes contain one or more splicing variants found within two or more splicing modules. Within each module, genes are more often represented by a single splicing variant, however, a small portion of genes can be found having two or more splicing variants within the same module.

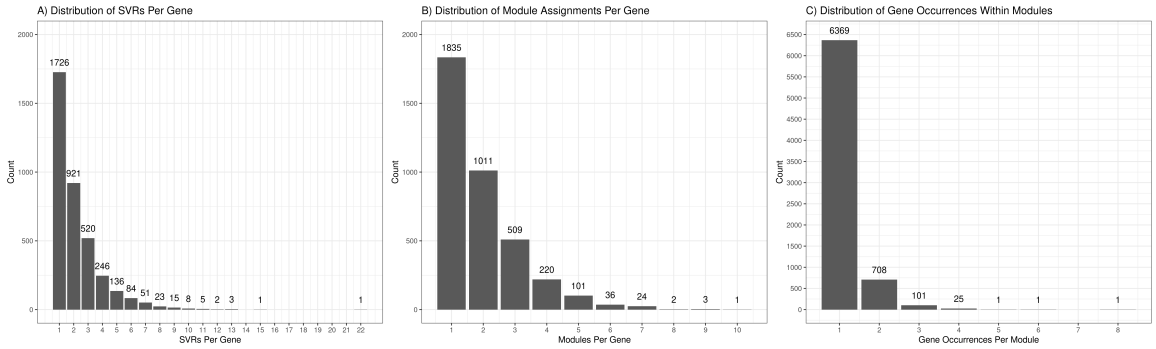

Figure 26: Characterization of SVRs and Genes Across AML Co-splicing Modules. A) Distribution showing the number of splice variant regions for each gene of the co-splicing network. B) Distribution showing the number of modules a gene belongs to given their splicing variants. C) Distribution showing the number of splicing variants belonging to a single gene within each co-splicing module

We evaluated quality of module detection in the AML network using a series of module quality statistics including module density, module connectivity, and overall module summary. For each module statistic we computed a Zscore using the observed module statistic relative to the mean and standard deviation of 10,000 random node to module assignments. Zscores < 2 were implied as being of poor quality for a Z-statistic of a given module with Zscores > 10 implying high quality modules. All 13 AML co-splicing modules were found to have high Zconnectivity, high Zdensity, and high Zsummary ($Z > 10$) values. Module density Zscores can be seen in (Figure 27).

Co-splicing Modules Predictive of Drug Response

Using the AML co-splicing network we aimed to identify network-based splicing signatures associated with drug response for a variety of small molecule inhibitors. Given the clinical relevancy of recurrent mutations in AML, we assessed both

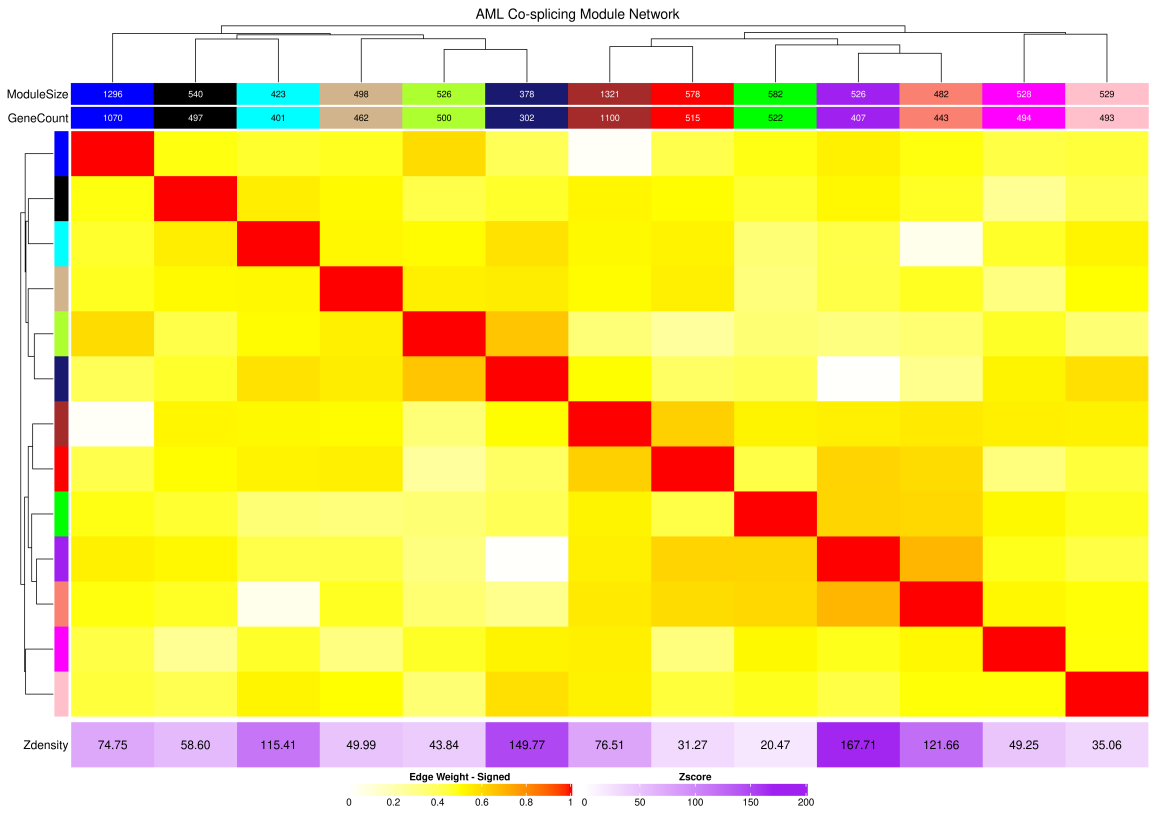

Figure 27: AML Co-splicing Module Network and Module Density Zscores. Dendrogram and heatmap indicate inter-modular relationships between co-splicing modules inferred using AML samples. Density Zscores of each module are presented below the network indicating all modules contain highly co-spliced variants.

co-splicing modules and mutational events in combination in order to characterize their joint contributions for drug response. Using mutational events and module splicing values, we performed iterative least absolute shrinkage and regression operator (LASSO) regression modeling for drug response using inhibitors having response measures for at least 150 samples in our network (105 total inhibitors). Mutational events were included as binary features indicating the presence or absence of a variant for a given gene. The presence or absence of fusion genes resulting from cytogenetic events were also included during modeling. Utilizing LASSO regression modeling for its embedded feature selection property, we trained LASSO models for each drug by first selecting 100 random training sets (75% of samples) and for each training set generated 1,000 random sets with replacement through bootstrapping. We then recorded the count of non-zero coefficients across the 100,000 models for each drug in order to quantify the frequency at which each coefficient was selected as predictive by the LASSO models. Non-zero coefficient frequencies for each inhibitor can be interpreted as correlations with drug response, however, each correlation value is determined with respect to all features utilized during modeling.

Several clinically relevant AML mutations were highly correlated with response to multiple inhibitors including FLT3-ITD, NRAS, and KRAS which all had non-zero coefficient frequencies $> 90\%$ for more than one drug. Of the 13 co-splicing modules, 9 had non-zero coefficient frequencies of at least 50% for one or more inhibitors. The green, salmon, and tan modules had the strongest correlations with drug response among splicing modules, with each having non-zero frequencies of at least 75% for one or more drugs. Of the 105 inhibitors that underwent multivariate modeling, 37 inhibitor models had at least one splicing module with a non-zero coefficient frequency $\geq 50\%$ and 16 inhibitor models had at least one splicing module with a non-zero coefficient frequency $\geq 70\%$.

Of the 13 co-splicing modules, the tan module had the highest number of significant drug response correlations, having non-zero coefficient frequencies of at least 75% for 10 inhibitors and at least 50% for 25 inhibitors. Its highest correlation was for venetoclax response with a non-zero coefficient frequency of 100% followed by selumetinib at 91%. The green co-splicing module was highly correlated with response to both PD173955 and flavopiridol, with correlations of 90% and 83% respectively. A summary of all inhibitor models having drug response to splicing module correlations of at least 70% can be seen in (Figure 28).

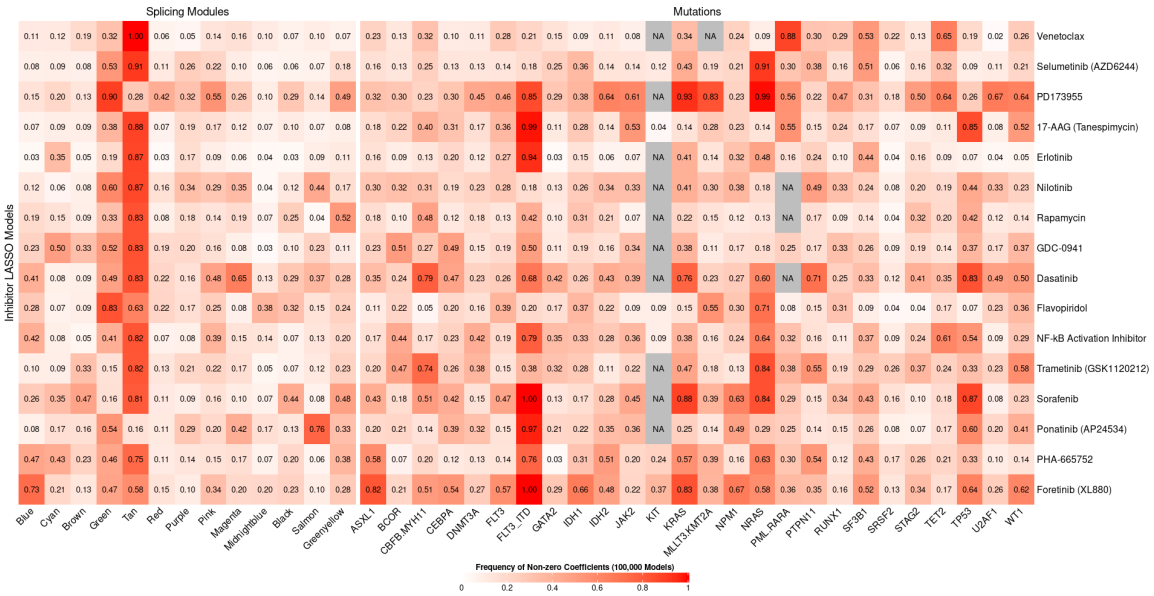

Figure 28: LASSO Non-zero Coefficient Frequencies of Drug Response Models. Inhibitors having co-splicing modules with non-zero coefficient frequences of at least 70% are indicated. Rows indicate iterative LASSO modeling for a given inhibitor. Columns indicate features of each LASSO model. Heatmap values indicate the ratio of non-zero coefficients across 100,000 bootstrap iterations for a given drug. Missing values indicate mutational events not included during modeling due to lack of positive calls in samples with available drug response data.

In addition to looking at feature correlations for each inhibitor model, we also characterized the results of LASSO modeling after grouping inhibitors into drug families. Here we grouped inhibitor models into their corresponding drug families based on the pathways and genes for which they target. For each model coefficient,

we recorded the frequency in which it was selected in more than 50% of LASSO model runs across the drugs for a given family. Only drug families in which two or more inhibitor models had at least one non-zero coefficient greater than 50% were included in the summerization. Along with several mutational events, 9 of the 13 co-splicing modules had correlations greater than 50% for at least one inhibitor model summarized into a corresponding drug family. A summary of LASSO model coefficient frequencies across drug families can be seen in (Figure 29).

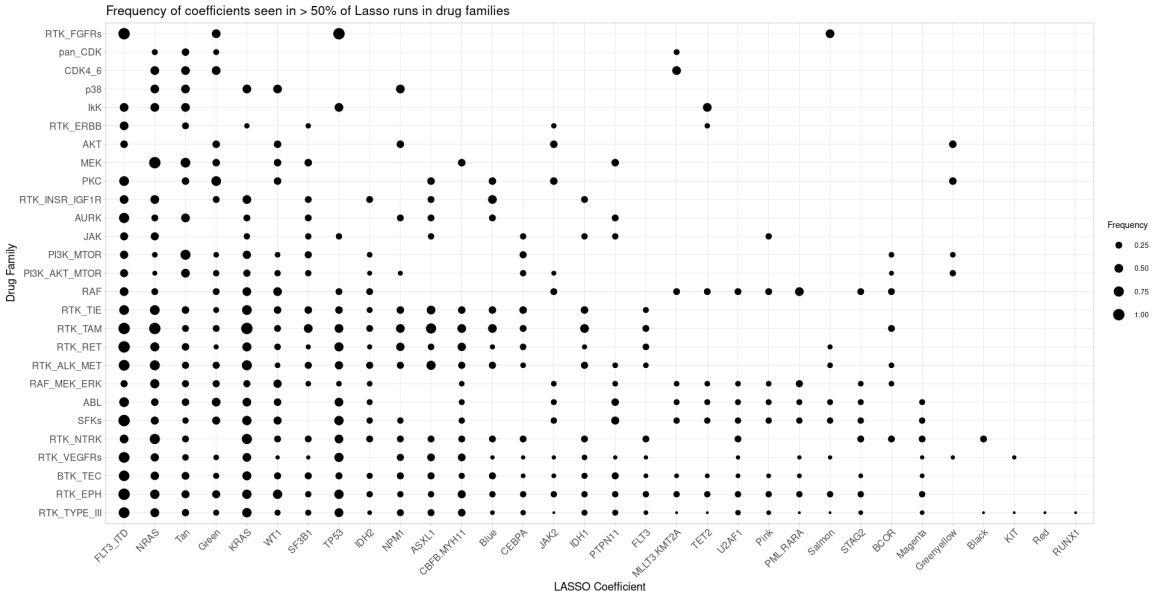

Figure 29: LASSO Coefficient Frequencies Across Drug Families. Dot sizes indicate the ratio of inhibitor models within a given drug family in which a given coefficient had a non-zero correlation frequency greater than 50%. Only drug families having at least two or more inhibitor models with at least one coefficient having a non-zero coefficient frequency greater than 50% are included.

Through LASSO modeling, we identified numerous co-occurrences of splicing modules and mutational events predictive of response for various inhibitors. For example, the tan co-splicing module frequently co-occurred with the PML-RARA fusion gene for response to venetoclax and the green module frequently co-occurred with mutations in both NRAS and KRAS for response to PD173955. Both the green and tan module frequently co-occurred with each other and with NRAS for predicting response to

flavopiridol. The salmon module had a correlation of 0.76 with ponatinib response and frequently co-occurred with FLT-ITD. Buzzsaw plots describing co-occurrence frequencies for PD173955 and ponatinib can be seen in (Figure 30). A buzzplot describing co-occurrence frequencies for venetoclax can be seen in (Figure 32).

In addition to performing LASSO drug response modeling with co-splicing modules and mutational events, we performed the same analysis using gene co-expression modules. Using gene co-expression modules of an AML co-expression network from (Tyner et al. 2018), we ran inhibitor LASSO modeling using the same AML samples used in our analysis of splicing. Consistent with results from (Tyner et al. 2018), numerous gene co-expression modules were highly correlated with response to numerous inhibitors. 17 of 21 co-expression modules had non-zero coefficient ratios of at least 50% for one or more inhibitors. More inhibitors had strong correlations with co-expression modules than with co-splicing modules, with 59 inhibitors having at least one co-expression module with a correlation ≥ 0.7 and 42 inhibitor models containing a co-expression module with a correlation ≥ 0.7. In general, co-expression modules were also highly correlated with inhibitors found to have high splicing module correlations. However, of the 16 inhibitors having splicing module correlations ≥ 0.7, 6 inhibitors still had a splicing module correlation greater than its highest correlation to a gene co-expression module. Results from gene co-expression LASSO modeling for inhibitors with high splicing correlations are shown in (Figure 31).

Consistent with results from (Tyner et al. 2018), numerous gene co-expression modules frequently co-occurred with mutational events for predicting drug response. Similar to the tan co-splicing module, the brown co-expression module had a correlation of 0.95 with venetoclax response and frequently co-occurred with the PML-RARA fusion gene. The greenyellow co-expression module was also highly correlated with venetoclax response having a non-zero coefficient frequency of 78%. A

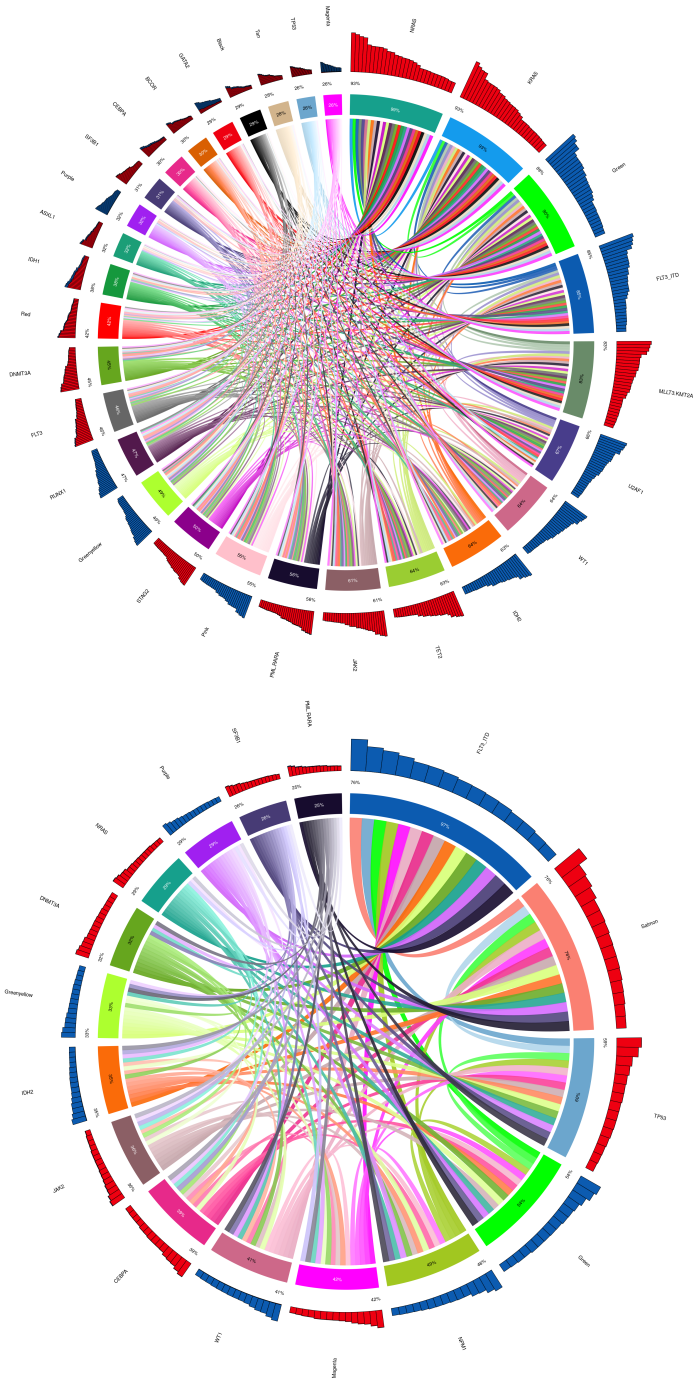

Figure 30: Buzzsaw Plots From Iterative Modeling of PD173955 (Top) and Ponatinib (Bottom) Response. Values in colored segments indicate non-zero coefficient ratios and are relative to segment widths, ordered clockwise from highest to lowest. Links indicate the presence of two coefficients in one or more LASSO runs and bar graphs above each link indicate their positive (red) or negative (blue) co-occurrence frequency. Links are ordered clockwise from highest to lowest within each segment. The highest co-occurence frequency for each segment is indicated as a guide.

comparison of buzzsaw plots for venetoclax models using co-splicing modules and co-expression modules can be seen in (Figure 32).

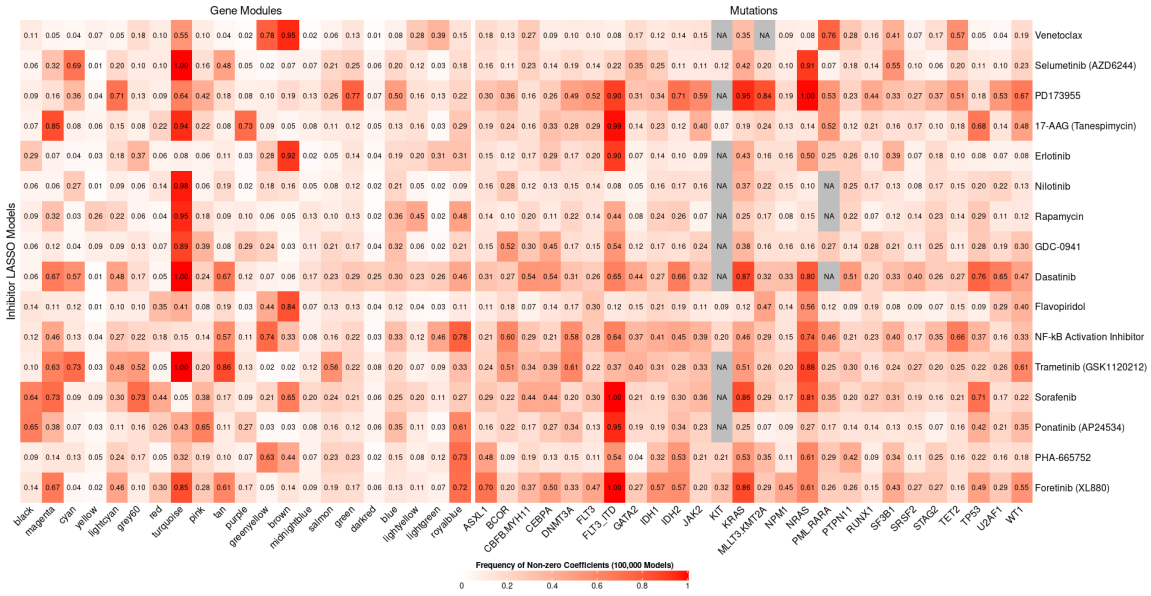

Figure 31: Coefficient Frequencies From LASSO Models Using Gene Co-expression Modules. For the purpose of comparing results of co-expression to that of co-splicing, inhibitors having co-splicing modules with non-zero coefficient frequences of at least 70% are indicated. See figure 28 for description of heatmap.

Functional Enrichment of AML Co-splicing Modules

In order to characterize the biological functions of co-splicing modules predictive of drug response in AML we performed a functional enrichment analysis of their corresponding genes. We identified significantly enriched GO biological process terms in the tan, green, and salmon co-splicing modules. The enriched GO terms occurred in a shared and module-specific manner. Interestingly, the tan co-splicing module, found to be the most predictive module for multiple inhibitors, had the fewest number of enriched GO functions of the significant co-splicing modules. Both the tan and green splicing modules were enriched for both immune and neutrophil-related GO terms (Figure 33).

The three drug response-associated modules were also enriched for various

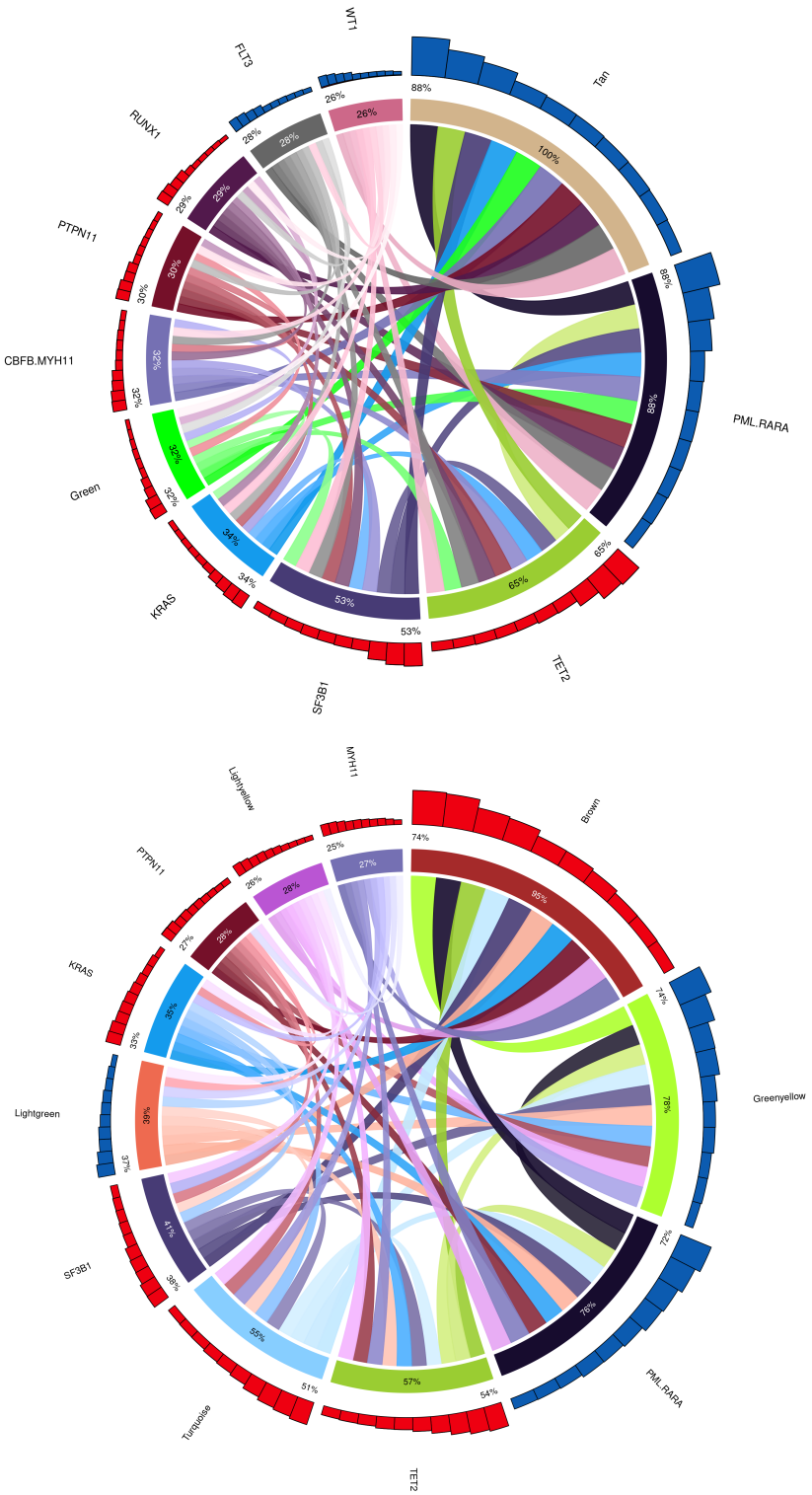

Figure 32: Comparison of Buzzsaw Plots For Venetoclax Response Between Co-splicing (Top) and Co-expression (Bottom) LASSO Models. See Figure 28 for description of buzzsaw plots.

immune-system related pathways from Reactome. The green module had the most number of enriched pathways and were often shared by either the tan or salmon co-splicing modules. Once again, the tan co-splicing module was enriched for the fewest number of Reactome pathways and included neutrophil degranulation and signaling by interleukins (Figure 34).

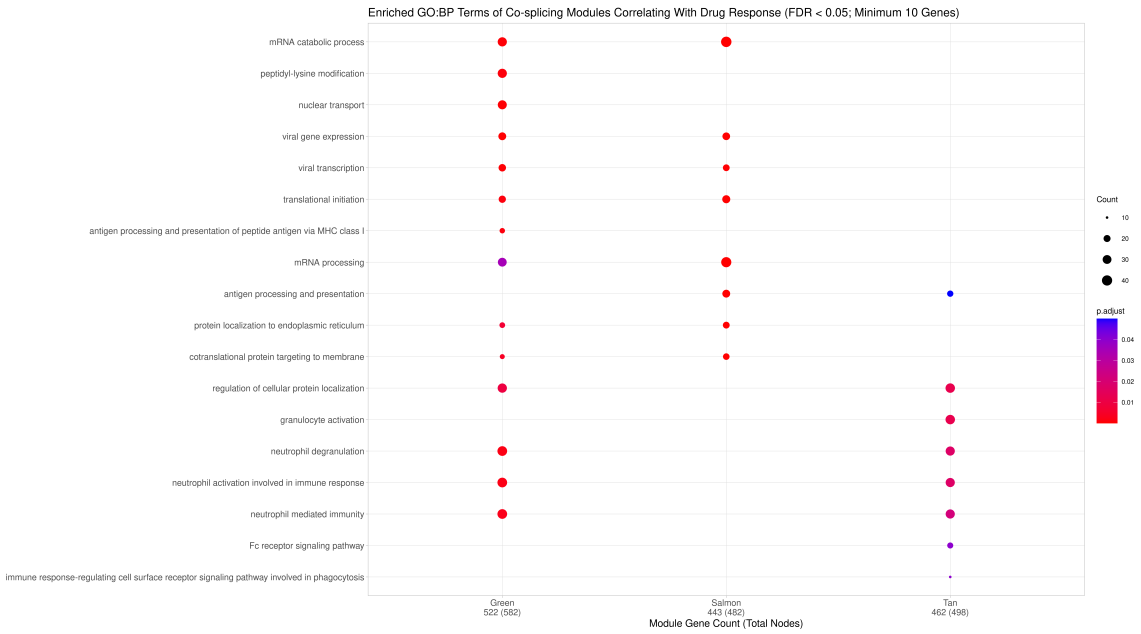

Figure 33: Enriched GO Biological Process Terms of Co-splicing Modules Predictive of Drug Response in AML.

Discussion

Chromosomal rearrangements, gene expression, and recurrent mutational events have become well characterized in AML, however, the role of alternative splicing in AML prognosis is less studied. Previous network inference studies of AML have revealed the existence of co-expression modules that are highly predictive of drug response for various inhibitors. The goal of this analysis was to determine if co-splicing modules also exhibit predictive behavior, perhaps at least in a complementary manner to that of gene expression. Using RNA-seq data from BeatAML we inferred a network of co-splicing in the form of network modules. This analysis allowed us to characterize

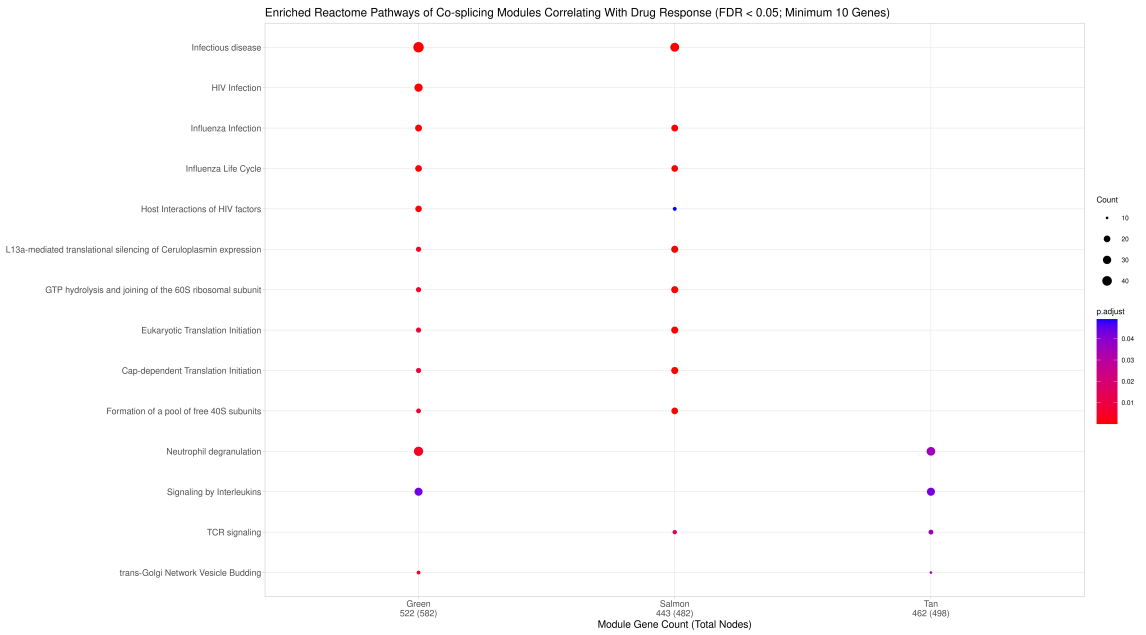

Figure 34: Enriched Reactome Pathways of Co-splicing Modules Predictive of Drug Response in AML.

coordinated splicing across AML and identify groups of complex splicing variants that are highly co-spliced. The AML data utilized in this analysis is genetically heterogeneous, with samples derived from various AML subtypes and cytogenetic profiles. This single network approach thus characterizes co-splicing patterns that are shared across different AML subtypes and prognosis groups.

A common approach for measuring the significance of network modules in relation to a phenotypic trait is to summarize the expression profile of each module's set of genes and correlate the module eigengene with the (often continuous) phenotype of interest. Although an effective method for identifying candidate network signatures, such an approach could be naive in a highly complex system such as that of AML, especially in regards to treatment outcomes which are well known to be highly complex across patients (Tyner et al. 2018). Deregulation of cellular pathways in cancer involve complex changes across the genome and are often orchestrated from a series of genetic events resulting in downstream consequences. Although treatment outcomes for AML

are far from being completely elucidated, it is well known that response to small molecule inhibitors are highly variable depending on the presence of specific cytogenetic and somatic mutational events. Therefore, when studying the associations of gene expression and alternative splicing to that of AML treatment, it is important to do so in a manner that accounts for other biologically relevant effects.

Previously utilized for gene co-expression, the iterative LASSO regression modeling approach allowed us to characterize how alternative splicing is coordinated alongside other genetic events in regards to drug response. Multiple co-splicing modules showed correlations to at least one inhibitor, with the tan co-splicing module being predictive for the greatest number of inhibitors. Of note was the correlation between the tan co-splicing module and drug response to the BCLS apoptosis regulator (BCL2) inhibitor venetoclax. The tan module was found as the most predictive genetic feature and was selected across all iterations of the venetoclax LASSO models. Consistent with another study on co-expression and venetoclax resistance, the tan module was significantly enriched for neutrophil degranduation and immune-related pathways. Surprisingly, evaluation of nodes within the tan module did not reveal a strong presence of splicing variants for BCL-related gene families, however many of the top hub nodes consisted of splice variants derived from genes involved in apoptosis. The top hub node of the tan module was a splicing variant from the SHISA5 gene (Scotin) which is known to interact with p53 during apoptosis (Bourdon et al. 2002). Venetoclax is a BH3 mimetic intended to promote cellular apoptosis by inhibiting activity of the anti-apoptotic BCL2 protein. Evading apoptosis is a one of the hallmarks of cancer and and these results might suggest that coordinated splicing changes contribute to the deregulation of apoptosis pathways. Further study would be needed, however, in order to verify this hypothesis. In addition, the results in this analysis do not necessarily demonstrate that splicing is a direct cause for therapeutic resistance, but rather abberant splicing coordination across the transcriptome

resulting from upstream events (e.g. somatic mutations) may contribute to the drug resistance phenotype.

We further compared the results of LASSO drug response modeling with splicing modules and mutations to results of the same analysis utilizing gene co-expression modules. Unsurprisingly, more drug response models were found to have strong correlations with gene co-expression modules than with co-splicing modules. For example, the lightyellow and magenta gene co-expression modules both had correlations of 0.52 with response to ibrutinib. Using co-splicing modules, the strongest co-splicing module correlation with ibrutinib was from the tan module having a correlation of only 0.24. Drug response for six inhibitors did have stronger correlations with splicing modules than with gene expression modules. For example, the green co-splicing module had a correlation 0.90 with response to PD173955 while the strongest gene co-expression module correlation to PD173955 was from the green co-expression module with a correlation of 0.77. LASSO results for models using either co-splicing or co-expression were similar for the remaining inhibitors that showed high correlations between splicing modules drug response. These results indicate that aberrant splicing may serve a complimentary role alongside gene expression as a transcriptional mechanism used by leukemic cells to develop therapeutic resistance.

In summary we have demonstrated the use of de novo network inference methods for characterizing co-splicing variation in complex disease such as AML. The use of a module-based analysis of co-splicing is advantageous in that it not only reduces the dimensionality of a highly complex transcriptome-wide network of splicing, but can provide biological insight into the cellular pathways for which coordinated splicing variants may be involved. This was the case for the tan co-splicing module which was highly associated with response to venetoclax and enriched for immune-related

pathways. Our results indicate that transcriptome-wide coordination of splicing may indeed serve as a transcriptional mechanism in which cancer cells develop reistance to therapy.

Methods & Materials for Co-splicing Network Inference of Drug Response in AML

Pre-processing of BeatAML RNA-Seq Samples Beat AML is an initiative by the Leukemia & Lymphoma Society (LLS) to utilize advances in genomic research and collaborations from multiple academic institutions and pharmaceutical companies to study and develop improved targeted therapies for treating AML. The initial BeatAML 1.0 dataset consisted of a study cohort of 672 primary specimens from 562 AML patients. RNA-seq, whole-exome, and drug response samples were pre-processed according to (Leek 2017). Briefly, AML samples were prepared for RNA-sequencing use the Agilent SureSelect Strand-Specific RNA Library protocol. Sequencing was then performed using the HiSeq2500 with a 100-cycle paired end protocol which resulted in ninety-six bp (post-trimming) paired-end reads. RNA-seq reads were aligned to the GRCh37 human genome with Ensembl build 75 gene models using the subjunc aligner (1.50-p2). 411 AML samples (earliest from each patient) and 32 healthy control samples were initially considered for co-splicing network analysis. High confident variant and fusion calls selected from (Tyner et al. 2018) were considered for multivariate modeling of drug response. Sample area-under-the-curve (AUC) values for 122 small-molecule inhibitors from an ex vivo drug sensitivity assay were utilized for identifying splicing module predictive of drug response.

LSV Quantification (MAJIQ) Annotation and quantification of complex splicing variants from AML samples was performed using the Modeling Alternative Junction Inclusion Quantification (MAJIQ) computation framework. Splice-graphs

were constructed for each gene in the AML RNA-seq data using the MAJIQ build function with the following parameters:

--min-experiments .08 --min-intronic-cov 1.0 --minreads 12 --minpos 6

In order to simplify the resulting splice-graphs and improve splicing quantification, a custom script was used to remove LSV junctions that were covered by a low number of junction spanning read alignments in each dataset before LSV PSI quantification. We removed junctions that contained fewer than 10 reads in more than 8% of samples in the AML dataset. After filtering out low coverage junctions, we removed LSVs from each splice-graph if they no longer contained at least two LSVs or if the total reads from all junctions was below 20 in more than 16% of samples. This results in 23,735 LSVs quantified across the 433 RNA-seq samples. LSVs from the simplified splice-graphs were then quantified for each sample individually using the MAJIQ PSI function. For a given LSV of a given sample, the PSI of each junction was quantified if at least one junction contained a minimum of 12 reads across 4 unique positions. A single Beat AML showing significantly lower coverage of quantifiable LSVs was further removed from the AML dataset. Only LSVs meeting the minimum read thresholds in all remaining samples (N = 432) were used for downstream analysis, resulting in 13,354 LSVs. Given that each LSV contains two or more junctions with each junction's PSI representing its relative splicing level to that of all junctions of the LSV, a single junction was used to represent each LSV for unsupervised learning methods such as that of de novo network inference. For each LSV of the two datasets, we selected the junction that had the greatest variance across samples to represent its splicing level for downstream analysis.

Adjusting For Confounding Effects Confounding artifacts were estimated and adjusted for using an approach based on (Parsana et al. 2019) for estimating and removing confounding artifacts prior to gene co-expression network analysis. Using

the SVA package in R, hidden confounders were estimated in the Beat AML LSV dataset with no fixed outcome of interest (only a constant intercept term). A single confounding principle component was estimated in the Beat AML LSV dataset which showed a significantly higher proportion of variance explained compared to that of the remaining principle components. We computed Pearson correlations between the 2000 and 5000 most variant LSVs of the AML dataset and each of the tissue subtypes from GTEx. In addition, we also computed Pearson correlations of the same count using various subgroups based on disease and tissue type of the AML samples, including de novo, non-de novo, bone marrow, peripheral blood, and various combinations of tissue and disease type. LSV correlations in each of the 10 GTEx tissue subtypes showed a normal distribution, while LSVs in the full AML dataset as well as the various subgroups showed an obscure tri-modal distribution. After adjusting for the confounding component in the AML dataset using an empirical Bayes linear model (from WGCNA), the correlation of the adjusted LSVs showed a normal distribution. In addition, hierarchical bi-clustering of the corrected LSVs resulted in separation of the healthy control samples from non-healthy samples. Control samples were then removed prior to SVR and network construction.

Network Construction of AML Co-splicing Module Network A signed SVR network was constructed from the 8,207 SVRs using biweight mid-correlations. The parameter maxPoutliers was set to 0.1 due to the heterogeneity of the AML dataset. In order to account for spurious correlations, soft-thresholding was used by raising the initial adjacency matrix by a constant (β) which promotes strong edge weights and demotes weak ones. A β of 5 was chosen resulting in an R^2 scale-free fit > 0.9. A topological overlap transformation was then performed on the adjusted adjacency matrix, which in turn characterizes the relationship between each network node while taking into consideration their shared relationships with other nodes.

Clustering and Formulation of Co-splicing Modules For identifying modules of highly co-spliced splicing variants (SVRs) in AML, we utilized an ensemble approach of clustering methods performed on the resulting topological overlap matrix. Using the TOM matrix as input, spherical clustering is performed by first transforming the symmetric matrix into a degree-normalized Laplacian matrix defined as $L = D - A$, where D is a matrix of the degrees of A and A is the initial adjacency matrix (in this case a topological overlap matrix). Then the top k eigenvectors are computed on the resulting graph Laplacian and k-means clustering is performed using the resulting eigenvectors with k being the resulting number of clusters. Since k is unknown beforehand, we follow a similar approach as (Botia et al. 2017) and first perform hierarchical clustering followed by dynamic tree cutting to determine an initial count of network modules. Here, hierarchical clustering is performed using average linkage distance on a topological overlap dissimilarity matrix, defined as $1 - TOM(A)$. We set the deepSplit parameter to 2 and pamStage to FALSE. Pam cut height was determined based on the 99% level of the resulting dendrogram, resulting in a cut height of 96.5. After the initial modules were constructed from the hierarchical clustering step, the module splicing values were computed and closely connected modules were merged with a merging threshold of 0.25. 13 modules remained after hierarchical clustering and merging and k = 13 was used for spectral clustering. During the k-means clustering of eigenvectors step of spectral clustering, 100 different initial centroid sets were chosen. To ensure reproducibility due to the possibility of varying cluster sets from random initial starting centroids, we repeated the spectral clustering step through 101 iterations selecting the final module sets using majority voting. Finally, module splicing values were again computed on the resulting module set and any closely related modules were again merged. After merging closely related modules a set of 13 modules remained and were used for further analysis.

Statistical modeling of Co-splicing Modules & AML Drug Response Iterative multi-variate regression modeling with lasso was performed using module splicing values and mutation status (0/1) for all available drugs with at least 150 network samples having response values and variant calls. For each drug's sample set, mutations were included if there were at least 10 positive calls. Sample sets for each drug were split into random training (75%) and test (25%) sets. 1,000 bootstrap sets were then generated using the selected training set and a lasso regression model was trained using 10 fold cross-validation for each set of bootstrap samples. Drug response AUC values were predicted for the test samples using the bootstrap models and then averaged. In addition, the entire procedure beginning with the initial training and test split was performed through 100 iterations and coefficient values were recorded for each bootstrap run of each iteration (100,000 total).

Functional Enrichment of Modules and Hub Nodes Functional enrichment of GO terms and Reactome pathways was performed using the *clusterProfiler* R software library (Yu et al. 2012). Due to the low number of unique genes in each co-splicing network, we used the total set of expressed genes as background after removing genes with low read counts. Enrichment was performed using a hypergeometric test and p-values were Benjamini-Hochberg adjusted for multiple testing. In addition, redundant GO terms were removed following GO enrichment based on GO tree similarity distance.

Data Visualization All dendrograms and heatmaps were created using the *ComplexHeatmap* R package (Gu, Eils, and Schlesner 2016). Buzzsaw plots were created using the *circlize* R package (Gu et al. 2014).

Chapter 5: Discussion and Concluding Remarks

Alternative splicing provides an essential contribution towards transcriptomic complexity in mammalian systems, with approximately 95% of multi-exon genes capable of expressing multiple transcripts from a single gene during RNA transcription. The transcriptome is highly dynamic with changes in gene expression and splicing production being a critical process for maintaining phenotype-specific properties such as tissue-specificity. These dynamic mechanisms must remain under tight control to ensure proper regulation of cellular function. Aberrant regulation of both gene expression and splicing can be an underlying mechanism for hallmarks of disease. Transcriptome-wide coordination of such regulatory processes are highly complex and thus it can be advantageous to study cellular regulation under a systems-level view.

Co-expression networks provide an effective means for inferring gene co-expression patterns from raw expression data when the inherent structure of the regulatory network for the system of study is unknown. Although it is known that alternative splicing occurs through transcriptome-wide coordination, a full characterization of systematic splicing across the transcriptome is lacking. Further, inferring a network of co-splicing is more difficult than gene co-expression for multiple reasons. First, although advancements in high-throughput RNA-sequencing have provided a revolution of advancements for transcriptomic studies, alternative splicing remains difficult to annotate and quantify due to short sequencing reads and limited sequencing depth. This is particularly notable when trying to assemble and estimate the expression of full length transcripts. Although this approach provides the most intuitive method for characterizing splicing, methods to do so suffer from inaccurate expression estimation. This can be especially problematic when utilizing inference methods for network construction which already have the potential for introducing

spurious correlations. The full length transcript problem can be mitigated with the use of localized methods, focusing on the relative inclusion or exclusion of individual exons or basic splicing events (e.g. exon skipping). Although effective for measuring variation in splicing levels, such methods fail to capture the full complexity that alternative splicing can entail. Previous studies have estimated that nearly 30% of mammalian splicing variants occur in the form of complex splicing events (Vaquero-Garcia et al. 2016; Sterne-Weiler et al. 2018; Li et al. 2018).

This dissertation demonstrated a framework for de novo inference of co-splicing. The central components of the proposed method in this dissertation were to be able to incorporate the use of complex splicing variants, formulate complex splicing in a manner suitable for a module-based analysis, provide a modular-level characterization of splicing variation across one or more biological systems, and identify co-splicing modules associated with various phenotypic traits. Splicing-graph annotation and quantification methods provide a comprehensive representation of the complexity of RNA splicing while alleviating some of the issues inherited from short RNA-seq reads (Vaquero-Garcia et al. 2016; Sterne-Weiler et al. 2018; Li et al. 2018). Chapter 2 presents and describes an algorithm that formulates complex splicing variants from splice graphs built using the MAJIQ software framework in manner suitable for de novo network inference of co-splicing (Vaquero-Garcia et al. 2016). We show that multiple local splice variants (LSVs) that share genomic regions undergoing alternative usage can be summarized into a single entity which we call splice variant regions (SVRs). LSV splicing values of a given SVR can be characterized using singular value decomposition (SVD), a commonly used data reduction technique. The formulation of SVRs from splice-graphs provide a suitable datatype for inferring networks of complex splicing. Modules of highly correlated splicing variants can be identified and studied in manner fitting within the WGCNA framework for gene co-expression (Langfelder and Horvath 2008). In return, variation in complex splicing

levels can then be characterized across phenotypes at the module-level.

In chapter 3 we first demonstrate the use of SVRs for de novo inference of co-splicing by characterizing variation of co-splicing modules across human tissue types. Using the formulated SVRs from GTEx RNA-seq data, we inferred co-splicing networks for ten human tissue types. We then identified a set of consensus modules consisting of complex variants that are highly co-spliced across tissues. Summarization of splicing levels for each module using SVD allowed us to characterize variation in co-splicing across tissues. We found that co-splicing module relationships are well preserved across independently constructed networks, with similar tissues having the highest preservation values. Functional enrichment of co-splicing modules revealed a variety of biological processes to which alternative splicing regulation may be contributing to proper cellular function. Intra-modular hub nodes were also found in a shared and tissue-specific manner with similar tissues also exhibiting greater overlap of hub nodes. Many hubs were conserved across the majority of tissue types and consistent across tissues was the fact that splicing hubs were highly enriched for genes involved in RNA splicing and binding, indicating that alternative splicing of RNA splicing genes may also contribute to transcriptome-wide splicing regulation. We identified a splicing variant that was found to be a top hub node of the red co-splicing modules across all ten tissue types. The variant of the HNRNPA2B1 gene was a complex splicing variant consisting of interchanging exons and retained introns. Our SVR formulation was able to include such an event in our network inference study in a comprehensive and interpretable manner. Consensus co-splicing modules were also found to be differentially spliced across tissues, indicating that groups of highly co-spliced splicing variants undergo coordinated splicing changes to help drive tissue-specific phenotypes. This demonstration in turn not only characterized across tissue cp-splicing variation at a systems-level, but also provided a proof of princple for the use of such a co-splicing network inference approach. The network characteristics

that were shared across tissues was consistent with the expected biology of study in that similar tissue types exhibited similar network behavior.

In chapter 4 we demonstrate the use of SVRs for studying co-splicing variation in human diseases such as cancer. The well curated and annotated BeatAML dataset allowed us to effectively characterize co-splicing variation in AML. Specifically, we inferred a network of co-splicing modules in acute myeloid leukemia (AML) with the aim of identifying co-splicing modules associated with drug response for various small molecule inhibitors. Given the highly heterogeneous nature of AML in terms of the genomic landscape of cytogenetic events and recurrent somatic mutations, it was important to study co-splicing within proper context. The LASSO regression L1 regularization technique provides an embedded form of feature selection. Through iterative multivariate modeling, we built LASSO regression modules for 105 inhibitors utilizing both the inferred co-splicing modules and mutational events and recorded features selected by LASSO through iteration (i.e. non-zero coefficients). Non-zero coefficient frequencies for each feature can be interpreted as a strength of correlation in regards to response for a given inhibitor and this approach has been previously utilized for identifying gene co-expression and mutations associated with drug response in AML (Tyner et al. 2018; Zhang et al. 2018). With this method we were able to find not only co-splicing modules correlating with drug response, but co-occurrences of both splicing and mutations associated with resistance to therapy.

Several splicing modules from our inferred co-splicing network of AML showed moderate to strong correlations with drug response for multiple inhibitors. Three co-splicing modules had correlations ≥ 0.75 for one or more inhibitors. Most notably, the tan co-splicing module had correlations ≥ 0.75 for thirteen different inhibitors, including a correlation of 1.0 for response to the BCL2 inhibitor venetoclax. For venetoclax response, the tan module was found to be the most correlated feature

followed by the PML-RARα fusion gene. Presence of PML-RARα fusion gene has been previously identified for association with venetoclax sensitivity (Kurtz et al. 2019; Zhang et al. 2018). Other somatic mutations previously reported as being associated with venetoclax response were identified with moderate correlations in our models as well, including TET2 and the SF3B1 splicing factor gene (Kurtz et al. 2019; Zhang et al. 2018). It is important to indicate that these results do not prove a direct causality from alternative splicing in developing resistance to AML treatment, but rather indicate potential regulatory mechanisms for which a resistant phenotype may be obtained from cancer cells. The exact mechanisms for which deregulated gene expression and alternative splicing occurs towards developing therapeutic resistance merits much further study. Further, comparison of results from LASSO modeling using co-splicing modules to results using gene co-expression modules indicated that gene co-expression modules have strong correlations with drug response for a greater amount of inhibitor types. These results are not suprising as most genes often have a primary or dominant isoform following transcription. As will be discussed in the following section, it is possible that low correlations between drug response and splicing may be the result of limitations when quantifying splicing from short read RNA-seq data as doing so requires the use of junction-spanning reads which may be limited in certain sample sets. Some inhibtors, however, did show higher response correlations with co-splicing modules than with co-expression modules and around a dozen inhibitors showed similar results for both splicing and gene modules. These results indicate a complimentary role for alternative splicing alongside gene expression as a transcriptional mechanism for promoting therapeutic resistance in AML.

Limitations Most computational methods developed for studying genomic data from high throughput sequencing data, even those frequently utilized, are not free of limitations. The co-splicing inference method demonstrated in this dissertation is of

no exception. It is important to note that results of any de novo network inference analysis must be taken with proper understanding and caution. CO-expression inference methods always have the potential to introduce false positive correlations (edge weights). These can arise through a variety of factors including both technical and biological confounders. Normalization procedures and batch correction methods similar to those used in this dissertation try and resolve such issues, but such approaches are far from perfect and come with their own set of biases and limitations. The issue of falsely inferred edges and/or edge weights is of particular concern and arriving to direct conclusions from the presence of individual network edges should be avoided. Our approach for co-splicing borrows from the WGCNA framework where we try and reduce the potential for spurious edges through a soft thresholding technique in order to better promote strong correlations and demote weak ones. Doing so, however, requires the use of a soft-thresholding value which is unknown and thus an assumption of approximate scale-free topology is made to select such a threshold. The use of a module-based analysis is also effective in that characterization of co-expression in regards to phenotypes of interest is less focused on individual edges, but rather generalizes groups of highly co-expression genes (or splicing variants). Although it is likely for co-expressed genes to be co-regulated by shared regulatory factors, direct claims of regulation should not be made. To this end, de novo inference methods for identifying network modules are better utilized and interpreted when the task at hand is to characterize co-expression rather than identify direct regulatory interactions.

Another limitation to the described approach is the use of splice junctions to quantify alternative splicing. We discuss and demonstrate the use of splice graphs to comprehensively annotate and quantify complex splicing variants, however, the ability to do so is highly dependent on the depth of sequencing for each sample in the study. Methods utilizing a localized approach for splicing quantification, including the

MAJIQ framework, depend on a sufficient coverage of junction spanning reads in order to accurately estimate the relative usage of splicing variants. Junction reads, however, represent a small ratio of the total read output for a given sample. The number of available junction reads can greatly vary on factors such as the depth of sequencing, read length, and whether a stranded or non-stranded library prep was utilized. We attempted to apply stringent filters as to which LSVs were usable in each of the two analyses performed and only included LSVs that met our minimum coverage filters prior to SVR formulation. The GTEx RNA-seq dataset in particular contained non-stranded reads of only 76 bps compared to the stranded 97 bp reads of the Beat AML dataset leading to a smaller number of LSVs having sufficient coverage for the human tissue analysis. Further, the number of samples included in the analysis was particularly large (N = 1,621). Although large sample sizes are ideal for genomic studies, using our stringent filters of having all or nothing in regards to samples and LSVs greatly increased the number of LSVs lost in the analysis. Again, these were stringent requirements and could possibly be mitigated by the use of imputation methods which have been utilized in co-expression network studies.

This limitation may have also contributed to the lower number of inhibitors that had drug response correlations with splicing modules than with gene co-expression modules. Although more junction reads were present across AML samples due to the longer read lengths, many LSVs were still lost due to inconsistent coverage across samples and were not included during SVR formulation. This may have led to splice variants that are potentially significant towards therapeutic resistance for various inhibitors, but were not included during de novo network inference. The purpose of utilizing LSVs (and formulation of SVRs) is to mitigate the issues of full transcript estimation while providing more comprehensive and less redundant splicing representations than that of an exon-based approach. If sensitivity for capturing phenotype specific splicing variation is of great concern, the former approaches may

be appropriate. These of course come with the limitations that have been addressed extensively in this dissertation.

Nonetheless, our demonstration of the GTEx data showed that the SVRs formulated from LSVs were able to capture a great deal of tissue-specific variation. Further, multiple co-splicing modules were found to have moderate to strong correlations with drug response for a variety of inhibitors. This in turn adds to the current state of alternative splicing in regards to its role in cellular regulation and its contribution towards various forms of disease such as cancer. It is clear that splicing changes play a significant role in driving cellular phenotypes, made clear yet again by our results. However, the extent to which alternative splicing contributes alongside gene expression in regulating cellular functions is still difficult to ascertain given the limitations of short read RNA-seq technology. As more samples with increased depth of sequencing and read lengths become available, the co-splicing inference framework demonstrated in this dissertation will be particularly useful in elucidating mechansisms of coordinated alternative splicing regulation across biological systems.

In summary, this dissertation presents a novel approach for studying coordinated alternative splicing variation across phenotypes in a comprehensive and interpretable manner. We successfully applied the described framework on two use cases: variation across tissues and drug response in AML. This approach can be further applied towards a variety of future studies involving transcriptome characterization using high-throughput RNA-seq data.

References

Alarcón, C. R., H. Goodarzi, H. Lee, X. Liu, S. Tavazoie, and S. F. Tavazoie. 2015. "Hnrnpa2b1 Is a Mediator of m(6)a-Dependent Nuclear RNA Processing Events." Journal Article. *Cell* 162 (6): 1299–1308. https://doi.org/10.1016/j.cell.2015.08.011.

Al-Yousef, A., and S. Samarasinghe. 2021. "A Novel Computational Approach for Biomarker Detection for Gene Expression-Based Computer-Aided Diagnostic Systems for Breast Cancer." Journal Article. *Methods Mol Biol* 2190: 195–208. https://doi.org/10.1007/978-1-0716-0826-5_9.

Anders, S., A. Reyes, and W. Huber. 2012. "Detecting Differential Usage of Exons from RNA-Seq Data." Journal Article. *Genome Res* 22 (10): 2008–17. https://doi.org/10.1101/gr.133744.111.

Botia, J. A., J. Vandrovcova, P. Forabosco, S. Guelfi, K. D'Sa, Consortium United Kingdom Brain Expression, J. Hardy, C. M. Lewis, M. Ryten, and M. E. Weale. 2017. "An Additional k-Means Clustering Step Improves the Biological Features of WGCNA Gene Co-Expression Networks." Journal Article. *BMC Syst Biol* 11 (1): 47. https://doi.org/10.1186/s12918-017-0420-6.

Bourdon, J. C., J. Renzing, P. L. Robertson, K. N. Fernandes, and D. P. Lane. 2002. "Scotin, a Novel P53-Inducible Proapoptotic Protein Located in the ER and the Nuclear Membrane." Journal Article. *J Cell Biol* 158 (2): 235–46. https://doi.org/10.1083/jcb.200203006.

Broido, A. D., and A. Clauset. 2019. "Scale-Free Networks Are Rare." Journal Article. *Nat Commun* 10 (1): 1017. https://doi.org/10.1038/s41467-019-08746-5.

Crews, L. A., L. Balaian, N. P. Delos Santos, H. S. Leu, A. C. Court, E. Lazzari, A. Sadarangani, et al. 2016. "RNA Splicing Modulation Selectively Impairs

Leukemia Stem Cell Maintenance in Secondary Human AML." Journal Article. *Cell Stem Cell* 19 (5): 599–612. https://doi.org/10.1016/j.stem.2016.08.003.

Dai, C., W. Li, J. Liu, and X. J. Zhou. 2012. "Integrating Many Co-Splicing Networks to Reconstruct Splicing Regulatory Modules." Journal Article. *BMC Syst Biol* 6 Suppl 1: S17. https://doi.org/10.1186/1752-0509-6-S1-S17.

Dam, S. van, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes. 2018. "Gene Co-Expression Analysis for Functional Classification and Gene-Disease Predictions." Journal Article. *Brief Bioinform* 19 (4): 575–92. https://doi.org/10.1093/bib/bbw139.

De Kouchkovsky, I., and M. Abdul-Hay. 2016. "'acute Myeloid Leukemia: A Comprehensive Review and 2016 Update'." Journal Article. *Blood Cancer J* 6 (7): e441. https://doi.org/10.1038/bcj.2016.50.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." Journal Article. *Bioinformatics (Oxford, England)* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Fagnani, M., Y. Barash, J. Y. Ip, C. Misquitta, Q. Pan, A. L. Saltzman, O. Shai, et al. 2007. "Functional Coordination of Alternative Splicing in the Mammalian Central Nervous System." Journal Article. *Genome Biol* 8 (6): R108. https://doi.org/10.1186/gb-2007-8-6-r108.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." Journal Article. *Biostatistics* 9 (3): 432–41. https://doi.org/10.1093/biostatistics/kxm045.

Gaiteri, C., Y. Ding, B. French, G. C. Tseng, and E. Sibille. 2014. "Beyond Modules and Hubs: The Potential of Gene Coexpression Networks for Investigating

Molecular Mechanisms of Complex Brain Disorders." Journal Article. *Genes Brain Behav* 13 (1): 13–24. https://doi.org/10.1111/gbb.12106.

Grange, Pierre de la, Lise Gratadou, Marc Delord, Martin Dutertre, and Didier Auboeuf. 2010. "Splicing Factor and Exon Profiling Across Human Tissues." Journal Article. *Nucleic Acids Research* 38 (9): 2825–38. https://doi.org/10.1093/nar/gkq008.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." Journal Article. *Bioinformatics* 32 (18): 2847–49. https://doi.org/10.1093/bioinformatics/btw313.

Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. "Circlize Implements and Enhances Circular Visualization in r." Journal Article. *Bioinformatics* 30 (19): 2811–12. https://doi.org/10.1093/bioinformatics/btu393.

Hackl, H., K. Astanina, and R. Wieser. 2017. "Molecular and Genetic Alterations Associated with Therapy Resistance and Relapse of Acute Myeloid Leukemia." Journal Article. *J Hematol Oncol* 10 (1): 51. https://doi.org/10.1186/s13045-017-0416-0.

Hanahan, R., D. Weinberg. 2000. "The Hallmarks of Cancer." Journal Article. *Cell* 100.

Holm, F., E. Hellqvist, C. N. Mason, S. A. Ali, N. Delos-Santos, C. L. Barrett, H. J. Chun, et al. 2015. "Reversion to an Embryonic Alternative Splicing Program Enhances Leukemia Stem Cell Self-Renewal." Journal Article. *Proc Natl Acad Sci U S A* 112 (50): 15444–49. https://doi.org/10.1073/pnas.1506943112.

Horvath, S. 2011. "Weighted Network Analysis." Journal Article. *Springer*.

Iancu, O. D., A. Colville, P. Darakjian, and R. Hitzemann. 2014. "Coexpression and

Cosplicing Network Approaches for the Study of Mammalian Brain Transcriptomes." Journal Article. *Int Rev Neurobiol* 116: 73–93. https://doi.org/10.1016/B978-0-12-801105-8.00004-7.

Iancu, O. D., A. Colville, D. Oberbeck, P. Darakjian, S. K. McWeeney, and R. Hitzemann. 2015. "Cosplicing Network Analysis of Mammalian Brain RNA-Seq Data Utilizing WGCNA and Mantel Correlations." Journal Article. *Front Genet* 6: 174. https://doi.org/10.3389/fgene.2015.00174.

Khwaja, A., M. Bjorkholm, R. E. Gale, R. L. Levine, C. T. Jordan, G. Ehninger, C. D. Bloomfield, et al. 2016. "Acute Myeloid Leukaemia." Journal Article. *Nat Rev Dis Primers* 2: 16010. https://doi.org/10.1038/nrdp.2016.10.

Kim, H. K., M. H. C. Pham, K. S. Ko, B. D. Rhee, and J. Han. 2018. "Alternative Splicing Isoforms in Health and Disease." Journal Article. *Pflugers Arch* 470 (7): 995–1016. https://doi.org/10.1007/s00424-018-2136-x.

Kurtz, Stephen E, Christopher A. Eide, Andy Kaempf, Nicola Long, Anupriya Agarwal, Cristina E. Tognon, Motomi Mori, Brian J. Druker, and Jeffrey W. Tyner. 2019. "Patterns of Sensitivity Exhibited by Venetoclax-Inclusive Drug Combinations in Acute Myeloid Leukemia." Journal Article. *Blood* 134: 878–78. https://doi.org/10.1182/blood-2019-126020.

Langfelder, P., and S. Horvath. 2007. "Eigengene Networks for Studying the Relationships Between Co-Expression Modules." Journal Article. *BMC Syst Biol* 1: 54. https://doi.org/10.1186/1752-0509-1-54.

———. 2008. "WGCNA: An r Package for Weighted Correlation Network Analysis." Journal Article. *BMC Bioinformatics* 9: 559. https://doi.org/10.1186/1471-2105-9-559.

Langfelder, P., R. Luo, M. C. Oldham, and S. Horvath. 2011. "Is My Network

Module Preserved and Reproducible?" Journal Article. *PLoS Comput Biol* 7 (1): e1001057. https://doi.org/10.1371/journal.pcbi.1001057.

Li, Y. I., D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, and J. K. Pritchard. 2018. "Annotation-Free Quantification of RNA Splicing Using LeafCutter." Journal Article. *Nat Genet* 50 (1): 151–58. https://doi.org/10.1038/s41588-017-0004-9.

Liu, R., A. E. Loraine, and J. A. Dickerson. 2014. "Comparisons of Computational Methods for Differential Alternative Splicing Detection Using RNA-Seq in Plant Systems." Journal Article. *BMC Bioinformatics* 15: 364. https://doi.org/10.1186/s12859-014-0364-4.

Necochea-Campion, R. de, G. P. Shouse, Q. Zhou, S. Mirshahidi, and C. S. Chen. 2016. "Aberrant Splicing and Drug Resistance in AML." Journal Article. *J Hematol Oncol* 9 (1): 85. https://doi.org/10.1186/s13045-016-0315-9.

Park, E., Z. Pan, Z. Zhang, L. Lin, and Y. Xing. 2018. "The Expanding Landscape of Alternative Splicing Variation in Human Populations." Journal Article. *Am J Hum Genet* 102 (1): 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002.

Parsana, P., C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek. 2019. "Addressing Confounding Artifacts in Reconstruction of Gene Co-Expression Networks." Journal Article. *Genome Biol* 20 (1): 94. https://doi.org/10.1186/s13059-019-1700-9.

Pierson, E., G. TEx Consortium, D. Koller, A. Battle, S. Mostafavi, K. G. Ardlie, G. Getz, et al. 2015. "Sharing and Specificity of Co-Expression Networks Across 35 Human Tissues." Journal Article. *PLoS Comput Biol* 11 (5): e1004220. https://doi.org/10.1371/journal.pcbi.1004220.

Saha, A., Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, G. TEx

Consortium, B. E. Engelhardt, and A. Battle. 2017. "Co-Expression Networks Reveal the Tissue-Specific Regulation of Transcription and Splicing." Journal Article. *Genome Res* 27 (11): 1843–58. https://doi.org/10.1101/gr.216721.116.

Sanati, N., O. D. Iancu, G. Wu, J. E. Jacobs, and S. K. McWeeney. 2018. "Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma." Journal Article. *Front Genet* 9: 183. https://doi.org/10.3389/fgene.2018.00183.

Shen, S., J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. 2014. "rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data." Journal Article. *Proc Natl Acad Sci U S A* 111 (51): E5593–601. https://doi.org/10.1073/pnas.1419161111.

Sterne-Weiler, T., R. J. Weatheritt, A. J. Best, K. C. H. Ha, and B. J. Blencowe. 2018. "Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop." Journal Article. *Mol Cell* 72 (1): 187–200 e6. https://doi.org/10.1016/j.molcel.2018.08.018.

The GTEx Consortium. 2015. "The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." Journal Article 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation." Journal Article. *Nat Biotechnol* 28 (5): 511–15. https://doi.org/10.1038/nbt.1621.

Tyner, J. W., C. E. Tognon, D. Bottomly, B. Wilmot, S. E. Kurtz, S. L. Savage, N. Long, et al. 2018. "Functional Genomic Landscape of Acute Myeloid Leukaemia."

Journal Article. *Nature* 562 (7728): 526–31. https://doi.org/10.1038/s41586-018-0623-z.

Vaquero-Garcia, J., A. Barrera, M. R. Gazzara, J. Gonzalez-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch, and Y. Barash. 2016. "A New View of Transcriptome Complexity and Regulation Through the Lens of Local Splicing Variations." Journal Article. *Elife* 5: e11752. https://doi.org/10.7554/eLife.11752.

Wang, B. D., and N. H. Lee. 2018. "Aberrant RNA Splicing in Cancer and Drug Resistance." Journal Article. *Cancers (Basel)* 10 (11). https://doi.org/10.3390/cancers10110458.

Wang, X., S. Huang, and J. L. Chen. 2017. "Understanding of Leukemic Stem Cells and Their Clinical Implications." Journal Article. *Mol Cancer* 16 (1): 2. https://doi.org/10.1186/s12943-016-0574-7.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler: An r Package for Comparing Biological Themes Among Gene Clusters." Journal Article. *Omics : A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Zhang, Haijiao, Beth Wilmot, Daniel Bottomly, Stephen E Kurtz, Christopher A. Eide, Alisa Damnernsawad, Kyle Romine, et al. 2018. "Biomarkers Predicting Venetoclax Sensitivity and Strategies for Venetoclax Combination Treatment." Journal Article. *Blood* 132 (Supplement 1): 175–75. https://doi.org/10.1182/blood-2018-175.

Zhou, J., and W. J. Chng. 2017. "Aberrant RNA Splicing and Mutations in Spliceosome Complex in Acute Myeloid Leukemia." Journal Article. *Stem Cell Investig* 4: 6. https://doi.org/10.21037/sci.2017.01.06.