UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Determination of SNP-induced Splicing Deregulation in the Human Genome

A thesis submitted in partial satisfaction of the

requirements for the degree Master of Science

in Integrative Biology and Physiology

by

Anna Konstorum

2010

The thesis of Anna Konstorum is approved.

_____

Emeran Mayer

_____

Arthur Arnold

_____

Xinshu (Grace) Xiao, Committee Chair

University of California, Los Angeles

2010

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABSTRACT OF THE THESIS

Computational Determination of SNP-induced Splicing

Deregulation in the Human Genome

by

Anna Konstorum

Master of Science in Integrative Biology and Physiology

University of California, Los Angeles, 2010

Professor Xinshu (Grace) Xiao, Chair

Improvement in sequencing and genotyping technologies has enabled identification of millions of single nucleotide polymorphisms (SNPs) and other types of genetic variants in the human population. Many genetic variants are postulated to affect disease predisposition, gene-environment interaction and treatment response. Recent genome-wide association studies (GWAS) have identified thousands of SNPs associated with various diseases and other human phenotypic traits. However, the mechanisms underlying the SNP and disease associations are largely unknown. Alternative pre-mRNA splicing is a process involved in over 70% of human genes as a mechanism for diversifying the proteome of a cell. The goal of this thesis was to create an algorithm that calculates the probability that an intragenic SNP disrupts splicing of the gene in which it resides in a tissue-specific manner. Data from high-throughput sequencing technology, along with knowledge of splice site strengths, *trans* factor binding motifs, and motif conservation were used to create a Bayesian network model of exon inclusion. The

model was tested using SNPs that have been shown *in vitro* to alter splicing. The model predicted that many GWAS SNPs associated with disease phenotypes may cause disruption of splicing in a tissue-specific manner, leading to putative predictions of the molecular mechanism of action of the disease-associated SNPs.

Introduction

One of the greatest promises offered by the completion of the human genome sequence is the development of tools to bring the concept of 'personalized genomics' to fruition (Mir K, 2009). It is now well accepted in the biomedical community that the susceptibility to, and heterogeneity in manifestation of, notable common diseases in any documented population can at least be partially explained by the variation found in the genomes of individuals (Biesecker et al., 2009). While environmental components certainly play a strong role in development of a disease, ultimately the response to environmental perturbations is also heterogeneous based on an individual's genetic profile. The original concept that genomic mutations that lead to disease do so by altering the codon 'signature' for a certain amino acid in that gene has certainly borne great discoveries in disease genetics (Anand et al. 2009). But with the advent of genome-wide association studies (GWASs, see below), a large quantity of common genomic variations are being found that are correlated with disease initiation and/or progression but that do not necessarily alter the amino acid composition of the protein, or that do not lie in protein-coding sequences. Due to these findings, the scientific community has begun to look for alternative explanations for the mutations/disease phenotype correlation.

In this thesis, I focused on identifying how Single Nucleotide Polymorphisms (SNPs), the most common type of variation observed in the human genome, may alter one mechanism of gene regulation, namely alternative splicing, using bioinformatic methods. I used high-throughput sequencing data to identify genomic and transcriptomic

'signatures' that are associated with high or low exon inclusion. Utilizing machine learning methods, I then built a model that could predict, based upon new genomic signatures as created by the SNPs, whether a given SNP could alter the alternative splicing of the corresponding exon to which it was nearest or within.

Background

I. Splicing

Unlike the genetic structure of prokaryotes, eukaryotic genes harbor both protein-coding exonic regions and non-coding intronic regions. After a gene is transcribed into pre-mRNA molecules, the introns must be removed, or spliced, from RNA molecule and adjacent exons ligated in order to produce the mature mRNA. Splicing is performed in the nucleus by a macromolecular complex termed the spliceosome. A discussion of the biochemical processing steps in pre-mRNA splicing must begin with characterization of conserved splice site sequences at exon / intron boundaries (Fig 1a). The sequence termed the 5' splice site is located at the 5' end of the intron. It includes a conserved 'GU' at the intron end and is surrounded by a less conserved concensus sequence. Similarly, the 3' splice site harbors a conserved 'AG' at the 3' end of the intron and is also surrounded by a concensus sequence (Black DL, 2003). In addition to the 5' and 3' splice sites, two other important sequence properties include a conserved adenosine within the intron (termed the 'branch point') and a polypirimidine tract found in metazoans 3' to the branch point. The splicesomal components, which include ribonucleoproteins composed of RNAs and associated proteins, catalyze the two trans-esterification reactions involved

in splicing (Fig 1b). The 2' hydroxyl group of the branch point adenosine nucleotide first attacks the 5' splice site and thus creates a 5' exon that is separated from the intron and an intermediate structure termed a 'lariat' that is composed of the 5' end of the intron attached in a phospho-diester bond to the adenosine. In the second reaction, the 3' splice site is attacked by the 5' exon and results in the two exons being ligated and the intron being removed from the molecule (Fig 1b) (Ritchie et al., 2009). These reactions are coordinated by the spliceosomal components and have been well characterized in plant and animal systems (Wachtel and Manley, 2009). Importantly, splice site choice by the spliceosome is not only dictated by the splice site sequences, but often by proteins that bind to non-splice site regulatory sequences. I shall refer to these regulatory sequences at length when discussing alternative splicing, but it is important to consider that they also exist around constitutive splice sites. Proteins that bind to these sequences are known as *trans* factors, and many have been shown as critical for splice site selection by the spliceosome (Ladd and Cooper, 2002).

While constitutive splicing is a very important component of the eukaryotic gene expression process, alternative splicing, or the joining of alternate exons to produce novel mRNA isoforms from the same genes, is now thought to be a major contributor to the complexity of higher organisms, and it is this process and the alterations possible therein that is the focus of this thesis.

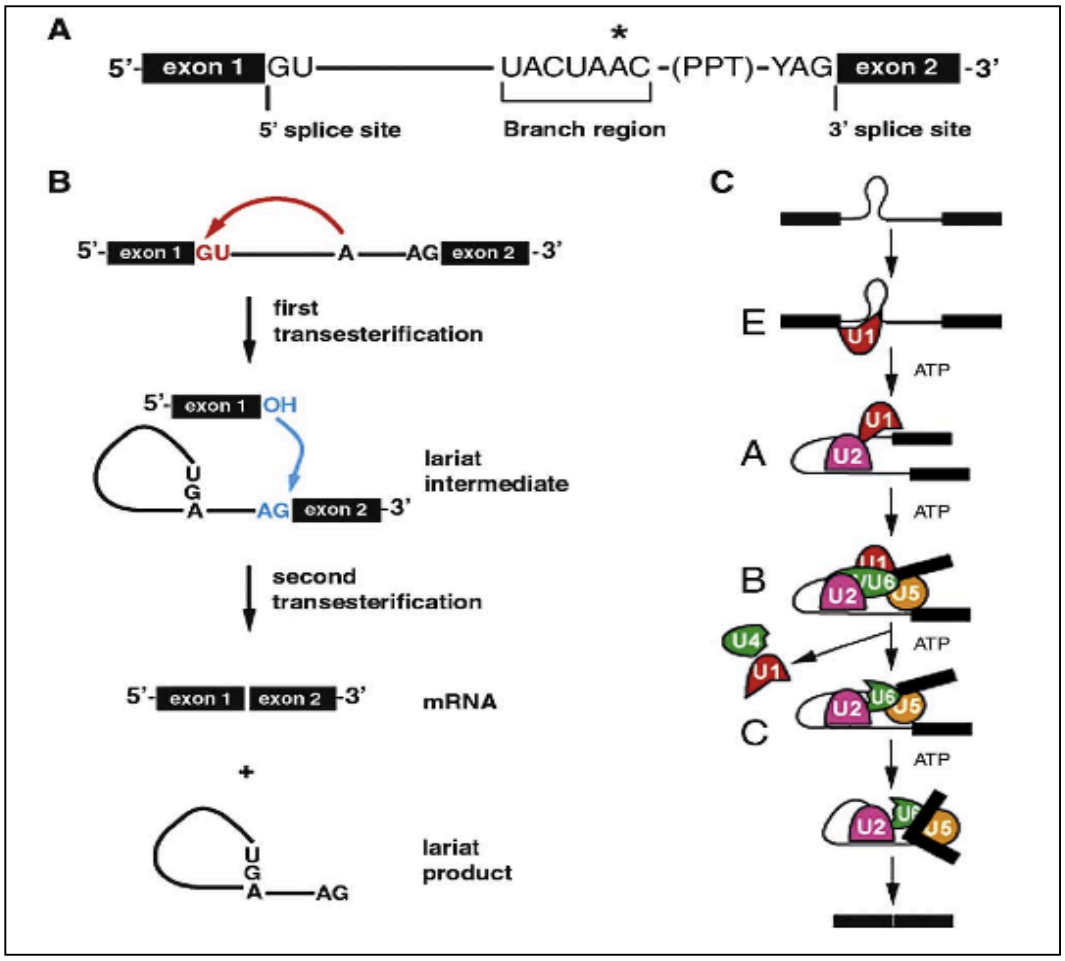


Fig 1: The splicing reaction
A.) Major sequence landmarks of eukaryotic introns: 5' splice site, branch region, polypyrimidine tract, and the 3' splice site; Diagram of the transesterification reactions required in splicing B.) without and C.) with additional diagrams of the splicing factors (adapted from Ritchie et al., 2009).

II. Alternative splicing and splicing regulation

After completion of the human genome, the number of genes within it was estimated to number approximately 24,000, only roughly double the number of genes found in fruitflies or worms. This discovery gave rise to questions of how so few genes could generate the complexity that we associate with *Homo Sapiens* and other higher mammals (Rubin 2001). Alternative splicing has been brought forth as a major generator

of this complexity, and with evidence that mammals have the greatest amount of alternative splicing compared to other organisms, there is certainly reason to consider alternative splicing as a major mechanism in the evolution of complexity (Keren et al., 2010). Moreover, it has recently been shown that over 90% of human genes undergo alternative splicing, which is a far greater percentage than earlier estimates of roughly 30% (Wang et al., 2008), evidence that strengthens the argument that alternative splicing plays an important role in diverse protein generation. Thus, it becomes critical to understand how splicing is coordinated in order to more fully understand the mechanisms of gene regulation and how it can be disrupted by mutations and other pathophysiological processes.

Alternative splicing (AS) may come about by three general variations in splicing: exon skipping or inclusion, intron retention, and alternative 5' or 3' splice site choice (Fig 2a). A gene may have one or more of these types AS. Additionally, AS of many genes is known to differ between different tissues in one organism, thus demonstrating how AS may increase diversity within an organism (Wang and Burge, 2008). *Cis*-acting elements that surround exon/intron boundaries are important in aiding the splicing machinery in distinguishing true exon/intron boundaries from pseudo-splice sites that are found along exons and introns and in directing the splicing reaction itself (Wang and Cooper, 2007). These *cis* elements can be classified by location (intronic or exonic) and whether they serve to enhance or suppress splicing, thus allowing for four categories of elements: ESEs, exonic slicing enhancers, ESSs, exonic splicing silencers, ISEs, intronic splicing enhancers, and ISSs, intronic splicing silencers (Fig 2b). Families of proteins,

termed *trans* factors, bind to these elements in order to carry out the above described functions. The best characterized binding factors include SR proteins, which bind ESEs, and promote splicing by recruiting splicing machinery to nearby splice sites. The hnRNP protein family, conversely, has been shown to bind to ESS and ISS motifs and to repress binding of splicing machinery (Black DL., 2003). While SR and hnRNP proteins are also important in constitutive splicing, many other families of proteins, including the FOX and CELF families, in addition to NOVA and TIA1 proteins, have been shown to only act as *trans* factors in an alternative splicing environment.

One must keep in mind that the *cis* elements serve to coordinate splicing in a combinatorial fashion. As an example, an HIV1 *tat* exon ESS binds hnRNPA1 and inhibits splicing, but splicing can be activated if an upstream ESE binds SF2/ASF (an SR protein). A similar example can be found in the *IgM* gene, where an ESE bound to a *trans* factor can antagonize an ESS-PTB complex adjacent to it (Matlin et al., 2005). Another mechanism of combinatorial action occurs when spliceosomal components are diverted by 'decoy' pseudo-splice sites from the true splice sites, causing the splicing of that element to be silenced (Cote et al., 2001). More generally, splicing regulator activity has been described as dependent on the genetic context in which regulatory elements reside as well as the cellular environment. With regards to the latter, differential expression of *trans* factors in various cell and tissue types can help explain differential splicing of the same gene in different tissues of one organism (Xiao and Lee, 2010). Thus, when considering how mutations may alter splicing, one must be cognizant of the complexity of the control mechanisms that are involved.

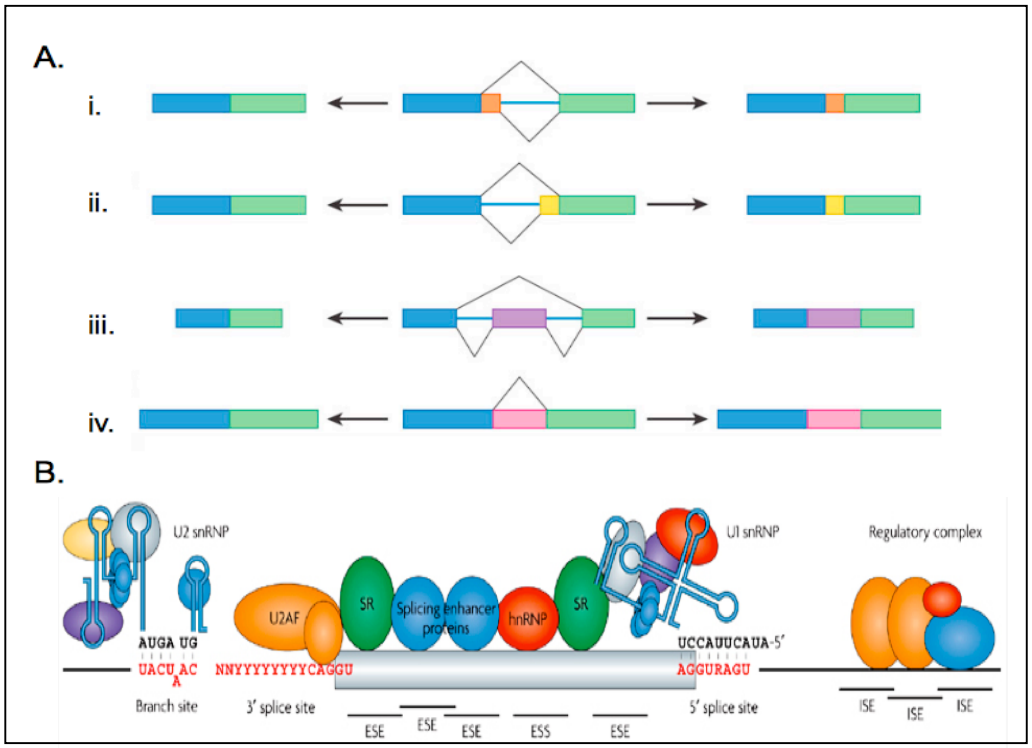


Fig 2: Alternative Splicing.
A.) Types of alternative splicing: i. alternative 5' splice site, ii. alternative 3' splice site, iii. cassette exon skipping, iv. intron retention (adapted from Nilsen and Graveley, 2010); B.) Diagram of *cis-elements* with associated *trans*-factors and spliceosomal components bound to an exon and surrounding intronic regions (adapted from Wang and Cooper, 2007).

III. Technology to probe splicing

As the cost for high-throughput transcriptomic technology has decreased, global studies of splicing have provided novel information on alternative splicing regulation and occurrence within and between tissue-types and organisms (Fox et al., 2009). The main technologies that have been critical in probing AS are cDNA and EST libraries, and next-generation sequencing technologies (NGS).

There are currently large public databases of ESTs and cDNAs that can be aligned

to the genome to search for AS events. cDNAs are reverse-transcribed nucleotide sequences of full-length mRNAs whereas ESTs are are incomplete sequencing reads from cDNA clones. The limitation inherent in using EST databases to search for AS events is that when ESTs are created, they are usually biased towards the 3' and 5' ends of transcripts due to the high rate of mRNA degradation *in vitro* (Blencowe, 2006). Thus, EST databases may provide incomplete representations of the transcriptome. In addition, EST databases can suffer from poor sequence quality and vector contamination, as well as bias towards a particular cell or tissue type (Kan et al., 2001). Development of AS-specific microarrays have been used to overcome these limitations, but because probes often exhibit high levels of cross-hybridization and can only be created for known exon/exon junctions, novel events cannot be discovered and non-specific binding noise needs to be taken into account when analyzing results (Fig 3). (Marioni et al., 2008; Wang and Cooper, 2007).

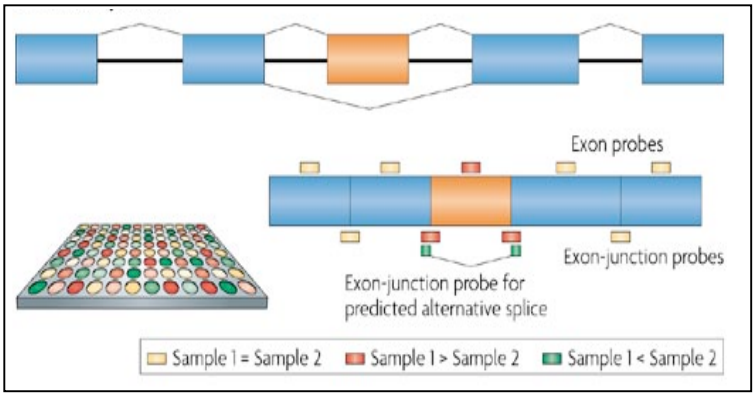


Fig 3: Microarrays to screen for alternative splicing.
The microarray will have olignucleotide probes complementary to exons or exon-exon junctions. cDNAs derived from the different samples to be labeled with Cy-3 (green) and Cy-5 (red) probes, respectively, in order to detect differential splicing between the two different samples. Note that because oligos must be constructed before performing the array experiment, novel splicing events cannot be discovered (adapted from Wang and Cooper, 2007).

Next generation sequencing technologies (NGS) have created an unprecedented set of transcriptomic data from which information on AS can be mined in given cell- and tissue-types. Unlike microarrays, NGS technologies have high sensitivity and low background noise, and are not dependent on previously annotated events (Xiao and Lee, 2010). There are a number of technologies that work in this manner, including the Roche/454 sequencer, the Applied Biosystems SOLiD system, and the SOLEXA/Illumina RNA-seq machine. All these technologies rely on a similar general method of sequencing: a population of RNA is converted to a library of cDNA fragments, and adapters are attached to one or both ends. Each molecule is then sequenced to obtain short sequencing reads from one end. After sequencing, the resulting reads are aligned to a reference genome (Fig 4; Wang Z., 2009). These technologies allow for direct sequencing of the cDNA, instead of relying on a hybridization signal as in microarrays. Recently, RNA-Seq technology has been directly compared to microarrays and has shown higher sensitivity to transcript expression (Marioni et al, 2008).

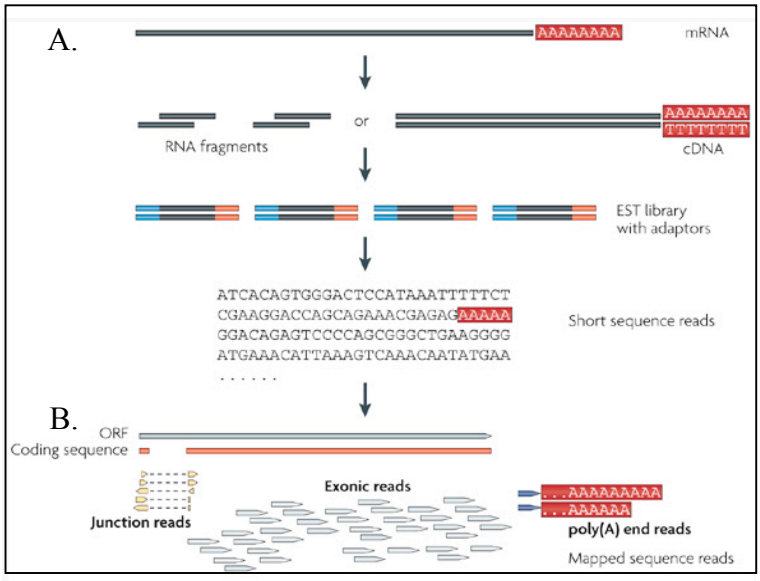


Fig 4 Scheme of high throughput RNA sequencing technology.
A.) mRNA strands from cells / tissues of interest are fragmented, converted to cDNAs, and capped with adapter sequences to facilitate sequencing.
B.) The short 'reads', as these are known, are then aligned to a reference genome (adapted from Wang Z et al., 2009).

Thus, NGS technologies are revolutionizing the understanding that scientists have of the transcriptome and of alternative splicing in reference to a given genome, which is important if one is to consider how AS changes with respect to specific polymorphisms in the genome.

IV. Single nucleotide polymorphisms (SNPs)

Humans share approximately 99.5% sequence identity between genomes, leaving 0.5% of variation to account for phenotypic differences and disease susceptibility (Levy et al., 2007). SNPs, or single nucleotide polymorphisms, are the most common types of variations observed across individuals. SNPs occur when signal nucleotides differ between members of a species. For a single base pair mutation to be considered a SNP, it must occur in >1% of the population (Yue and Moult, 2006). Aside from substituting one base pair for another, single nucleotides may also be removed or added in a given genome, such SNPs are known as 'ins/del' SNPs and while they are important contributors to human genetic variation (Vali et al., 2008), the thesis research presented here is only focused on substitution SNPs, although it can be improved to assay ins/del SNPs in the future. Of the substitution SNPs, roughly 4% are found in coding regions, and those can either create 'missense' or 'nonsense' mutations. Missense mutations do not alter the amino acid composition of the corresponding protein, while nonsense mutations do (Krawczak and Cooper., 1997).

While several studies have shown how nonsense mutations can effect phenotype by changing protein stability and / or degradation, the most recent SNP databases show

that most SNPs do not map to open reading frames, thus causing scientists to hypothesize that many deleterious mutations may occur via altering transcriptional or translational regulation of genes (Chorley et al., 2008, Hardy and Singleton, 2009). As the money and effort spent on cataloging SNPs increases, and databases of SNPs associated with all sorts of phenotypes increase, it becomes imperative to work on bioinformatic methods to identify and predict how the SNPs may affect phenotype, especially if they do not act via changing amino acid composition of associated proteins.

V. Genome-wide association studies (GWASs)

There have been several approaches used to map the genes that underlie common diseases. Two main categories of these approaches include candidate-gene studies and linkage mapping. In candidate-gene studies, the alleles of base pairs in a gene or genes that are hypothesized to be involved in the phenotype are sequenced in case and control groups. Drawbacks to this approach include the reliance on *a priori* hypotheses, as well as the questionable import of rare mutation associations in the greater population (Hirschhorn and Altshuler, 2002). In linkage mapping, disease genes are mapped by typing genetic markers that flank the disease gene in families and identifying regions that are associated with the disease. In this approach, family pedigrees are used to track the inheritance of the disease allele. Linkage analysis cannot identify gene variants that have low penetrance and cannot progress from a broad region of linkage without an impossibly large number of sibling-pairs (Altshuler et al., 2008). Genome-wide association studies, GWASs, survey the entire genome for causal variants and do not require assumptions of

causative genes or large family pedigrees. There are still many challenges and open questions in GWAS study design, including (but not limited to) appropriate selection of markers, sample size selection, and bias from population heterogeneity. Nevertheless, with the growth of SNP databases and lowered cost of genotyping technologies, GWASs have become the gold standard to identify associations between SNPs and diseases (Hirschhorn and Daly, 2005).

In GWASs, a group of individuals with and without the disease or phenotype of interest are selected while controlling for other variables, including age, ethnicity, environmental exposure, etc. Next, utilizing a library of common SNPs, researchers assay each individual for the allele frequencies at all the SNPs. If allele frequencies significantly differ between 'cases' and 'controls', the SNP is said to be associated with the phenotype or disease (Frazer et al., 2009).

From single labs to multi-national groups, such as the UK-based Welcome Trust Case Control Consortium, have come hundreds of GWASs which have implicated over 1000 of SNPs in common diseases and phenotypes (Samani et al., 2007; Hindorff et al., 2009). The main challenge now rests on understanding if the SNPs can actually cause molecular changes in the cellular environments that can contribute to the phenotype (Shen et al., 2009). It must also be noted that SNPs are often in linkage disequilibrium (LD), or non-random association due to lower recombination in local regions, with nearby SNPs, thus it is difficult to differentiate whether a causal SNP identified in a GWAS is the causal variant or in LD with it (Schmitt et al., 2010). Importantly, there has been criticism of the GWAS approach to disease mutation identification, including

lack of clear functional significance of identified SNPs, insufficient control for population stratification, and inability to find rare disease-causing alleles (McClellan and King, 2010; Manolio, 2010). The issue of population stratification may be dealt with by utilizing appropriate statistical methodologies (Hao et al., 2010) and research on rare-alleles can complement findings of GWASs. But if the first criticism is effectively met by the scientific community, ie. if SNPs identified by GWASs are shown to alter the phenotype of the organism harboring the mutation, then GWASs will remain a critical tool in uncovering the contribution of genetic alterations to disease.

VI. Experimental results of SNP effects on splicing

With the knowledge that while SNPs are the most prevalent polymorphism in the human genome, that most do not change protein-coding sequences, and that alternative splicing plays a much larger role in generating genomic diversity than previously considered, the question arises whether a SNP may cause changes in alternative splicing of a given gene, and thus generate its phenotypic effects by this method. Strikingly, it has been proposed that 60% of mutations that contribute to disease do so by altering splicing (Lopez-Bigas et al., 2005). Thus, it becomes of critical importance to better understand and predict how SNPs may alter splicing of a gene if we want to better understand the molecular nature of the SNP-disease associations that are now being discovered via GWASs.

Polymorphisms may change splicing by altering *cis* acting factors such as splice sites, *trans* factor motifs, and other regions in and around the gene that is being spliced.

SNPs may also change expression / stability of *trans* factors (Wang and Cooper, 2007). While mutations in the genes of *trans* factors are very likely to alter global splicing profiles, we shall focus on SNP effects on *cis* factors as we are interested in better understanding how structure of the gene sequences that are being spliced contributes to their splicing regulation.

Splice sites, as discussed previously, are crucial for exon recognition of the spliceosome. It has been estimated that 9-10% of point-mutation-mediated diseases fall within splice sites. SNPs may also either disrupt or create ESEs, ESSs, ISEs, and / or ISSs (Cooper et al., 2009). The effects mediated by alteration of these sequences are harder to predict due to the combinatorial nature of the splicing code. For example, a missense *cis*-acting mutation was discovered that inactivates an ESE in the MCAD (medium-chain acyl-CoA dehydrogenase) gene, causing increased exon skipping and loss of the resultant protein and MCAD deficiency. Yet, there exists a polymorphism upstream of the mutation that disrupts an ESS, and thus allows for higher exon retention (Nielsen et al.,2007). This example illustrates the complex manner by which SNPs may act on *cis* elements to alter splicing of a given gene. Currently, there are close to 100 *in vitro* studies that have shown SNP-mediated disruption of splicing. With the advent of high-throughput technologies, the field is primed to more thoroughly and systematically probe SNP-mediated effects on splicing.

VII. Comprehensive review of published in silico analyses of SNP effects on splicing

Due to the importance that SNPs may play in splicing deregulation, several groups

have used computational methods to predict how a given SNP may affect splicing. Recent studies have built upon statistical algorithms that were developed to predict the strength of splice sites and *cis* elements. I shall review these algorithms and then elaborate on the research that has utilized them to bioinformatically predict effects of mutations on splicing.

VIIa. Basic algorithms

Maximum entropy modeling was proposed by Yeo and Burge to predict the strength of splice site sequences (with higher strength referring to higher probability that the site is a true splice site and not a decoy). This model can incorporate base pair dependencies between adjacent and nonadjacent nucleotides (Yeo and Burge, 2004). Using this model, it has been shown that splice site sequences in constitutive exons have an average higher scores than alternatively spliced exons, and AS exons with higher levels of inclusion have higher splice site scores than AS exons with lower inclusion (Xiao et al., 2007). Thus this model is amenable to helping to predict changes in splicing when a SNP falls within the splice site regions. In addition, algorithms to interrogate whether certain sequences correspond to ESE or ESS motifs have also been developed. A widely used algorithm to identify and score ESEs is known as 'RESCUE-ESE', and was developed by Fairbrother and Burge in 2002. The group used their molecular knowledge of ESEs, such that the motifs function as splicing enhancers when found in exons but as splicing suppressors when found in introns, thus predicting that the motifs should be selected for in constitutively spliced exons (and avoided in intronic regions

near splice sites). Another molecular hypothesis they used was that ESEs would be more important in regulating splicing when splice sites of the exon had lower scores, as described above, thus requiring higher *cis*-regulation than strong splice sites. Thus, their method used a statistical algorithm which identified ESE hexamers as a.) much more prevalent in exons vs. introns (near splice sites) and b.) much more prevalent in exons with 'strong' splice sites vs. exons with 'weak' splice sites. The hexamers that most strongly displayed these properties were clustered by sequence similarity and candidate enhancer motifs from each cluster were tested and shown to enhance splicing of a reporter exon in a minigene assay. Another procedure to identify ESEs used the functional SELEX (Systematic Evolution of Ligands by Exponential enrichment) method, which allows researchers to identify the sequences of RNA motifs that are bound to specific proteins. Using this assay with splicing enhancer trans factor SR proteins SF2/ASF, SC35, SRp40, and SRp55, the authors were able to identify mRNA concensus sequence motifs that the proteins bound to (Cartegni et al, 2003). A procedure to search for ESSs was developed using an *in vitro* reporter assay to search for decanucleotides that, when inserted into a reporter exon, caused that exon to be spliced out of the reporter gene. 133 unique ESS decamers were discovered using this approach, termed FAS-ESS (Fluorescence-Activated Screen for Exonic Splicing Silencers) (Wang et al., 2004). The splicing motifs identified by these methods have been used by several groups in their respective algorithms to predict changes in splicing by SNPs.

VIIb. Research that combines the basic algorithms

Houdayer et al. used maximum entropy modeling and RESCUE-ESE, along with four similar tools (neural network-based splice site prediction, Splice-Site Finder, Automated Splice-Site Analysis, and Exonic Splicing Enhancer Finder) to examine if these methods could predict changes in splicing caused by mutations in the RB1 protein in select retinoblastoma patients. 17 variants showed mutations in canonical splice sites, and all were correctly predicted as deleterious by maximum entropy modeling (deleterious was defined as 20% score reduction by the mutation using any of the known methods). 84% of remaining variations that were within 60 bp of the splice sites were correctly predicted as deleterious by at least one tool (Houdayer et al., 2008).

A similar algorithm for predicting effects of SNPs derived from GWASs on general gene function is *SNPinfo (*Xu and Tayler, 2009). The authors developed the tool for researchers interested in formulating functional predictions for any published SNP. SNPs were only evaluated for efficacy in altering splicing if they were located within two base pairs of an intron/exon junction or were located in or created ESSs (as predicted by FAS-ESS) or ESEs (as predicted by RESCUE-ESE or ESEfinder) 70 bps from intron/exon junction. SNPs were classified to affect splicing activity if they were found to disrupt a splice site or disrupt or create an ESE or ESS, and if they were not located in or near constitutive exons.

ElSharawy et al. used a similar starting filter to screen for SNPs that could potentially modulate splicing (ElSharawy et al., 2009). Specifically, SNPs that were located at either canonical splice sites or ESEs within 30bp from the intron/exon junction

were experimentally tested for SNP-induced splicing changes by amplification of cDNAs from three tissues (brain, blood, or lymphoblastoid cell lines) that included samples that were homozygous and heterozygous for the SNPs and homozygous for the common alleles. PCR products were cloned to verify allele-dependent splicing. Experimental results showed that approximately 8% of tested SNPs altered splicing, with the authors ultimately concluding the *in silico* splicing predictions were not effective. It must be noted that the authors of the above-described studies did not examine the action of SNPs on splicing using knowledge of the complexity of of splicing, ie. since it is has now been established that that ESEs, ESSs, ISEs, ISSs, and splice sites all contribute to splicing in a combinatorial fashion, prediction of effects of SNPs on splicing must incorporate knowledge of the multiple factors that occur in and around the SNP location to coordinate splicing. It must also be noted that the authors did not assay changes in splicing in a tissue-specific manner, and as it has been shown that splicing often occurs in a tissue-dependent context (Wang et al., 2008), the SNPs that were tested may indeed cause changes in splicing in other tissues.

In a recently published study, sequences containing SNPs known to cause exon skipping were compared to sequences that harbored 'neutral' SNPs that were assumed to not change splicing (Woolfe et al., 2010). The authors found that, on average, SNPs that caused exon skipping were more likely to lose an ESE or gain an ESS than neutral SNPs. The 'splicing' SNPs were also more likely to occur near exon/intron borders and also to be in sequences that were more conserved and had higher frequences of ESEs, and lower frequency of ESSs, than the neutral SNPs. Importantly, while the authors characterized

general features of the two types of SNPs, they did not attempt to build an algorithm or model that could make probabilistic predictions of which of class of SNPs newly identified variants were members of. In addition, cellular-context specific factors (such as *trans* factor expression levels) were not assayed in the research.

VIII Purpose of current study

Altogether, these studies showed that prediction of splicing regulation by human polymorphisms must include more information than just location of SNPs within splice sites or *cis*-elements. The algorithms must at the very least contain information on the total makeup of *cis*-elements, splice site strengths, and trans factor levels in order to quantitatively predict changes in splicing that may be caused by the SNPs. This is precisely what the current thesis aims to accomplish. By utilizing knowledge derived from RNA-Seq of the genomic sequence profiles of highly included vs. lowly included exons in eight human tissues, together with expression of *trans* factors in each tissue, we have created an algorithm that aims to not only rank new SNPs in potential to alter splicing (specifically exon skipping) in a tissue-specific manner, but to additionally discover basic sequence properties of exons that are more or less highly included in human tissues.

Methods

I. Outline

The purpose of the current project, as stated in the background, is to use high-throughput gene expression data to better understand the genomic signatures of splicing and thus predict how SNPs can disrupt these genomic signatures. The general methodology is outlined in Figure 5 and below.

1. Selection of data to create statistical model.

2. Parameter selection and calculation from the data to train the model and calculation of *trans* factor expression scores.

3. Selection and testing of the model using cross-validation and positive (and negative) control SNPs.

4. Use of the model to predict changes in splicing caused by SNPs from GWASs.

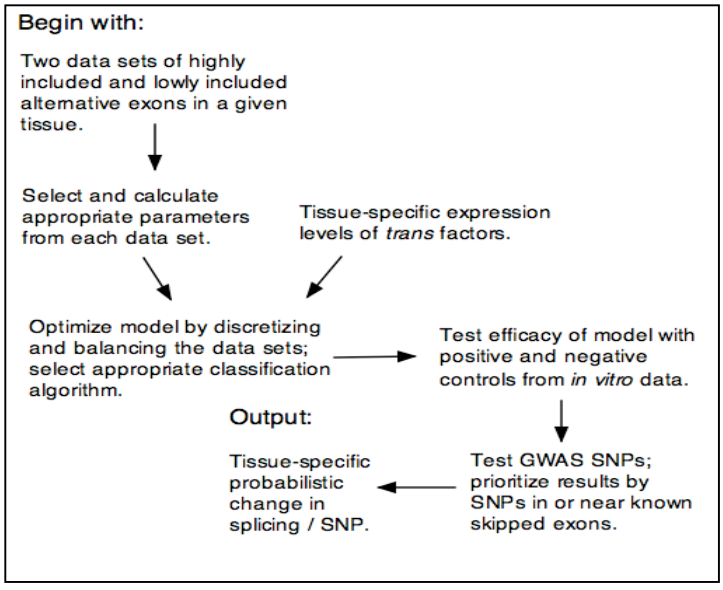


Fig 5 Diagram of research methods
Steps that were undertaken to create and use a functional statistical model of differential inclusion (high or low) of skipped exons in human tissues to predict changes in splicing caused by human SNPs. Each step is detailed in the *Methods* sections.

II. Selection of data

An ideal training data set would have contained isoform information for cells with genetic background with and without SNPs, such that we would be able to obtain a comprehensive understanding of the effects that various mutations, at various positions in the exons or introns, have on splicing efficiency. Since such a dataset is not available, we utilized RNA-Seq mRNA read data that allowed us to calculate the inclusion levels of exons in a tissue-specific manner. Our rational for this approach was that such a dataset would provide us with a feature set of parameters that would be specific to high and low inclusion exons. We then calculated the parameter values of the exons that SNPs were within or near and determined whether they were more similar to the high or low inclusion exons assayed earlier. If a SNP caused a change in parameter values such that the overall feature set of the exon became more similar to a different inclusion group than without the mutation, then we predicted the SNP would cause a change in splicing.

IIa. RNA-Seq data collection and genome mapping

Transcriptome data from human tissues was utilized that was originally obtained in Wang et al, 2008. The dataset included mRNA-Seq-generated sequences of up to 29 million 32 base-pair reads in eight human tissues. Tissues samples were obtained from both sexes of unrelated individuals. The reads had either been mapped directly onto the hg18 assembly of the human genome or onto known splice junctions using the Efficient Alignment of Nucleotide Databases software (ELAND), which allowed up to two mismatches. The splice junction database (http://genes.mit.edu/burgelab/mRNA-

Seq) had been created using known exons in UCSC knowngenes, Refseq, and Ensemble transcripts, as well as bioinformatically predicted exon junctions. The splice junction sequences were 56 bp in length and contained the first 28 bp of the downstream exon and the last 28 bp of the upstream exon. These length requirements were used in order to require a minimum of 4 bp matching of the read for each side of the junction.

IIb. Measuring EIL (exon inclusion levels) of skipped exons per tissue

Exon skipping has been demonstrated to be the most prevalent form of alternative splicing in the human transcriptome (Sultan et al., 2008). Thus, we chose to focus our analysis and predictions on alternative exon skipping events. Skipped exons were identified by parsing EST and cDNA data to identify alternatively spliced exons (C. Greer, personal communication). The level of skipping of a known skipped exon in a tissue can be quantified using a metric termed EIL: 'exon inclusion level'. Skipped exons that overlapped each other and reads that mapped to or near such events were excluded from the analysis. EIL was defined as the (inclusion density) / (inclusion density + exclusion density). Inclusion density (d_i) refers to the read density of the exon and both exon inclusion junctions. Exclusion density (d_e) refers to the read density of the exclusion junction. Read density, in turn, refers to the number of reads mapped to the region divided by the total number of mappable positions in that region (L_i for the inclusion region and L_e for the exclusion region). Thus, for L_i, the value corresponded to the length of the exon plus 28 to account for junction regions. Since the exclusion region only encompassed a junction, L_e equaled 28. 28 was used for junction length as the reads

were 32 bps long and for a read to map to a junction region, it had to have at least a 4 bp overlap with the other exon, thus effectively allowing for only 28 bps upon which it could map to. A single pseudocount was added to the total number of reads in order for very low read counts to not skew the results. The pseudocount (PS_i for inclusion and PS_e for exclusion) was divided proportionally to the number of mappable positions corresponding to exclusion and inclusion (Xiao et al., 2009). Thus, we have the following formula for EIL:

$$EIL = \frac{d_i}{d_i + d_e} \tag{1}$$

Inclusion and exclusion density, respectively, were calculated as follows, with L_i = (exon length + 28) and $L_e = 28$:

$$d_i = count_{body} + count_{incjunction} + \frac{PS_i}{L_i} \tag{2}$$

$$d_e = count_{excjunction} + \frac{PS_e}{L_e} \tag{3}$$

$$PS_i = \frac{L_i}{L_e + L_i} + \frac{L_i}{28 + L_i} \tag{4}$$

$$PS_e = \frac{L_e}{L_e + L_i} + \frac{28}{28 + L_i} \tag{5}$$

Events with less than 10 total reads ($count_{body}$+$count_{incjunction}$+$count_{excjunction}$) were excluded from the model and analysis.

III. Selection of parameters from data

For each alternative splicing event, parameters related to splicing regulation were calculated fin order to obtain a parameter profile which could be used to model splicing regulatory characteristics.

IIIa. Genomic regions

In addition to analyzing the exonic sequence, upstream and downstream intronic regions were also used in the analysis as *cis* acting factors in the intronic regions have been shown to play an important role in splicing (Cooper et al, 2007). For each exon, two upstream and two downstream regions were selected: one region nearest the skipped exon, and one region nearest the next upstream or downstream exon, respectively. If the total distance was less than 500 bp between two exons, the two intronic regions were divided in half whereas if the distance exceeded 500 bp, the 250 bp nearest the exon was termed the 'up' or 'down' stream region, whereas the 250 bp nearest the neighboring exon was identified as the 'up_up' or 'down_down' stream region (Fig 6). Motif and conservation scores (see below) were calculated within each genomic region, and 5' donor and 3' acceptor splice site scores were calculated for only the exon.

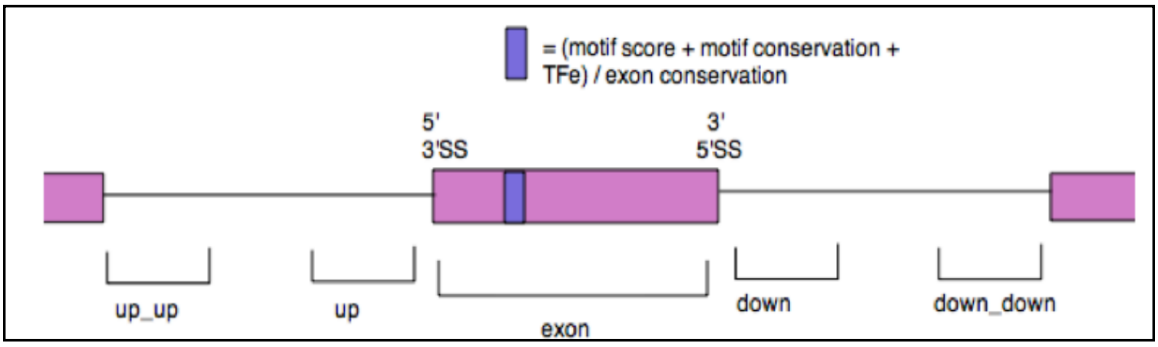


Fig 6 Schematic of skipped exon parameters
Motif scores (weighed by conservation and *trans* factor expression) were calculated for motifs in exons, up and up_up-stream introns, and down and down_down stream introns (see text for region definitions). Additional parameters included average region conservation, 5'SS and 3'SS scores, and region length.

IIIb. Splice site scores

As described in the background section, maximum entropy modeling can be used to score the information content of splice sites using available knowledge of true splice sites and can incorporate information from both adjacent and nonadjacent positions (Yeo and Burge, 2004). The {-20 to +3} positions were used to score 3' splice sites (with '-20' referring to the bases of the upstream intron and the '+3' referring to the first three exonic bases) and the {-3 to +6} positions were used to score 5' splice sites. Higher scores predict higher probability that the site will be used to splice in the exon. For predicting effects of SNPs on splicing, scores were calculated with and without presence of the SNP allele.

IIIc. Motif scores

Motif scores were calculated for *cis* elements in exons (ESEs and ESSs) and in introns (ISEs and ISSs). *Cis* elements were divided into two groups: known and

bioinformatically predicted. Known motifs were selected from the literature where *trans* factors were found to bind to the sequence elements *in vitro* (Matlin et al. 2005; Smith et al., 2006; Cooper et al., 2009). Bioinformatically predicted motifs that were used included those identifed by the FAS-ESS (Wang et al., 2004) and RESCUE-ESE (Fairbrother et al., 2002) procedures discussed earlier. Motifs were grouped either by *trans* factor to which they bound, for known motifs, or by sequence similarity, for bioinformatically predicted motifs. The motif score, M_m, was calculated as the log-transformed frequency of of the motif in constitutive exons (F_e) divided by the frequency of the motif in surrounding introns (F_i) :

$$M_m = \log_2(\frac{F_e}{F_i}) \qquad (6)$$

F_e was calculated as the frequency of the motif in constitutive exons divided by the combined length of all constitutive exons. F_i was similarly calculated as the sum of the frequency of the given motif in upstream and downstream introns of constitutive exons divided by the length of the upstream introns plus the length of the downstream introns. All calculations excluded splice site regions as defined above. This method of scoring motifs takes into account the fact that motifs that are more frequently found in constitutive exons have been shown to function as ESEs, and motifs that are less represented in exons vs. introns are more likey to function as ESSs. One can see that motifs that serve to repress splicing would be less included in constitutive exons since these exons are always spliced in, and motifs that enhance splicing would be more highly included for the same reason. This reasoning has been verified experimentally and it has

also been shown that ESEs (ie. motifs with high scores) found in introns actually function in splicing repression (Fairbrother and Burge, 2002). Thus, motifs found in introns were multiplied by '-1' in order to weight them such that they would be predicted to have opposite effects on splicing than they would have had had they been found in exons. Scores for motifs within a group were added together to obtain a total score for the group. Since there is overlap between known motif sequences and predicted sequences, if a known motif was scored, any predicted motifs that overlapped that specific region were excluded from score calculation for the predicted groups.

IIId. Conservation

Conservation was assessed for the main genomic regions in and around the exon (exon, up_intron, upup_intron, down_intron, downdown_intron) as well as for individual motifs that were identified within the regions. A conservation score was determined using the PhastCons program, which identifies evolutionarily conserved elements in a multiple alignment given a phylogenetic tree, based on a phylogenetic hidden Markov model (phylo-HMM, Siepel et al., 2005). The score used as a parameter was an average of the phastCons score (S_{pc}) derived over the region of interest (exon, intron, or motif), such that the conservation (C) score of a given element or motif over the length of the region (L_r) equaled the following:

$$C = \frac{\sum_{1}^{L_r} S_{pc}}{L_r} \qquad (7)$$

IIIe. Trans factor scores

The importance that a given *cis* element plays in the regulation of alternative splicing of a given exon will necessarily be dependent on whether the corresponding *trans* factor which it binds is expressed in the tissue / cell type of interest. Since many transcription factors have tissue-specific expression profiles, the weight, or importance, of a *cis* element in any tissue can be weighed by the expression of its *trans* factor in that tissue. Thus, *cis* element motif scores that were calculated for known *trans* factors were weighed by tissue specific *trans* factor expression levels (see below). Using the RNA-Seq data, *trans* factor expression levels were calculated and converted to RPKM (reads per kilobase of exon model per million mapped reads). Read counts were converted to RPKM since any given gene expression levels are a function of the molar concentration of RNA and of the transcript length. The equation to calculate the RPKM for a given gene (*trans* factor, tr) is the following:

$$T = RPKM_{tr} = 10^9 \frac{c}{NE} \qquad (8)$$

In this calculation, c represents the number of mappable reads in exons of the corresponding gene, N is the total number of mappable reads in the tissue, and E is the sum of exons in base pairs (Mortazavi et al., 2008).

IIIf. Combining motif, trans factor, PhastCons scores

Since a *cis*-acting SNP would not change conservation or *trans* factor expression in a given cell / tissue type, we used conservation and *trans* factor scores as 'weights' on motif scores, such that the final 'score' of a motif group, S_m, equaled the products of the

individual motif scores (M), the motif conservation scores (C), and the *trans* factor expression scores (T) summed across all the motifs in that group.

$$S_m = \sum MCT \qquad (9)$$

During our analysis, we had discovered that exon conservation was widely different between the high and low EIL groups. We surmised that this high conservation would mask conservation of motifs within exons, thus C_e=average phastCons score of the exon and for motif scores within exons:

$$S_m = \frac{\sum MCT}{C_e} \qquad (10)$$

IIIg. Element length

The lengths of the regions (exons, up_, upup_, down_, and downdown_introns) were also included as parameters as it has been shown that shorter exons and introns are more highly associated with skipped exons, possibly due to the kinetic binding properties of spliceosomal components (Hertel 2008). Recently, it has been shown that exons that harbor SNPs known to alter splicing are significantly shorter than the average exon length for all HapMap exons (Woolfe et al., 2010).

IV. Building the model

Machine learning broadly refers to the creation of algorithms that can computationally aide in classification of data by construction of statistical models (Tarca et al., 2007). In the bioinformatics field, these statistical approaches have had many

applications, including but not limited to disease classification based on gene expression microarray data, protein structure prediction from primary sequence data, and subcellular structure determination based on fluorescence microscopy images (Baskar et al, 2006). All potential statistical models were tested using Weka, an open-source software in Java that allows one to run various machine learning algorithms as well as data pre-processing, classification, and clustering (Hall et al., 2007). We tested several machine-learning approaches in order to arrive at an optimal model to classify the two types of events, including fast decision tree learner methods (Quinlan 1993)**,** sequential minimal optimization modeling (J. Platt 1998), and Bayesian network analysis (Bouckaert, 2008). There were usually 5-10 times more events in the high EIL category vs. the low EIL category. Thus, we chose a random subset of events from the high EIL group to match the number of low EIL events. Non-random techniques, such as Tomek links (Souto et al., 2006) to parse the data did not provide superior results to the random method. Important attributes were selected by computing the value of the chi-squared statistic with respect to the class.

V. Testing the model

SNPs that have been validated *in vitro* to either cause increased exon skipping or to have no effect on exon inclusion levels were recently collected into one database (Woolfe et al., 2010). We used these SNPs to test whether our model could predict the *in vitro* results when the reference and SNP genomic sequences were entered into the model. Since the database did not include SNPs that occur in splice sites and have been

shown to alter splicing, we performed a literature search to identify such SNPs and included these SNPs in our positive control dataset (Nalla and Rogan 2005, Pan et al., 2002, Kawase et al., 2007, Sivagnanam et al., 2008, Kubo et al., 2005).

VI. Running the model

The model was used to predict effects on splicing of alternative exons of SNPs identified via GWASs in a tissue-specific manner (see Hindorff et al., 2009 for catalog of published SNPs). Since SNPs occur in linkage disequilibrium (LD) with surrounding SNPs and as GWASs cannot functionally identify causal SNPs, we tested all SNPs in LD with the database SNPs and used the SNP with the highest probability to alter splicing within the haplotype as the predicted functional SNP. LD SNPs were identified using the UCSC human genome browser database and were phased with the GWAS SNP using phase data available through the International HapMap Consortium (Rhead et al., 2009). In addition, the model was used to predict if multiple SNPs in LD with each other (haplotypes) could alter splicing of a skipped exon and we examined if this prediction differed from the effects predicted by the model for individual SNPs.

Results and Discussion

I. Properties of RNA-Seq data

Read data from RNA-Seq experiments on eight human tissues was mapped onto known genes from the Ensemble database. Our results showed that, as previously

reported, the mismatch frequency was highest at the beginnings and ends of reads (Fig 7). This phenomenon occurs due to a higher rate of sequencing errors in the RNA-Seq procedure, but so far this mismatch imbalance has not been shown to cause increased mapping error (Wang et al., 2008). Over 50 million reads mapped to known Ensemble genes and junctions, with a ratio of junction / total reads averaging at 0.074 (Table 1).

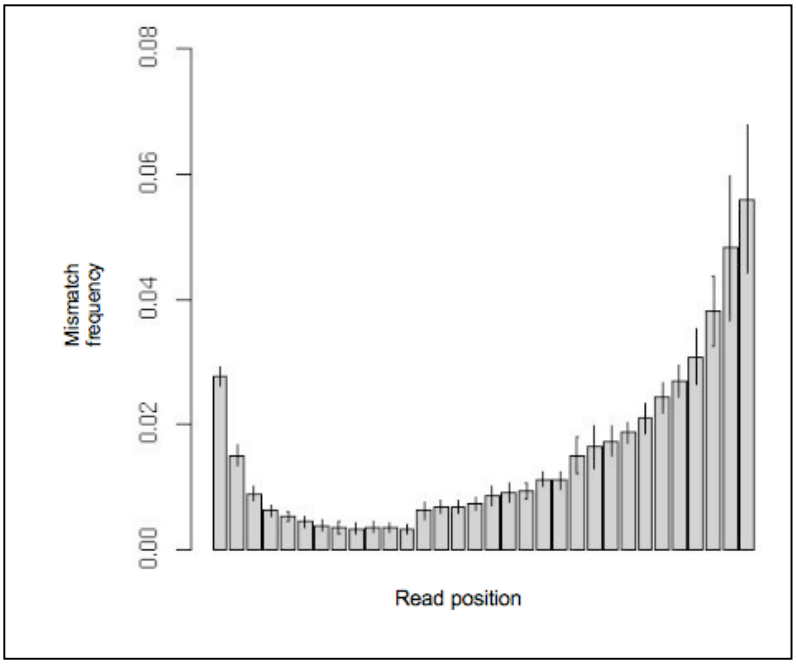


Figure 7: RNA-Seq read mismatch frequency
Mismatch frequency of reads based on position in read where mismatch occurs. Average mismatch frequency was 0.4 errors/read. Note that mismatch frequency is highest at tail ends of reads (where highest rate of sequencing error is known to occur in RNA-Seq).

Tissue	Body	Junction	Junction / Total
Adipose	8,320,106	714,156	0.079
Brain	4,398,884	394,180	0.082
Colon	8,155,643	610,751	0.070
Heart	4,981,990	363,296	0.068
Liver	5,977,296	366,093	0.058
Lymph node	6,663,504	564,893	0.078
Skeletal muscle	6,707,483	543,061	0.075
Testes	8,630,594	769,235	0.082
Total	53,835,500	4,325,665	0.074

Table 1: RNA-Seq read counts
Counts of reads in Ensemble genes of eight human tissues. 'Body' reads refer to reads that were within the bodies of exons, while 'junction' reads refer to reads that spanned known or predicted exon-exon junctions. The ratio of junction to total reads averaged 0.074 (SD +/- 0.008).

II. RNA-Seq data distribution in skipped exons in eight human tissues

After calculation of the EIL for known skipped exons in all eight human tissues, tissues were hierarchically clustered based on the EIL scores for the exons (Fig. 8). As skeletal muscle and heart are both composed of contractile muscle tissue, it is not surprising to note that exons in both tissues have similar EILs, signifying that similar isoforms of many genes are expressed in both tissues. It is also of interest to note that testes and lymph tissue clustered together as both tissues harbor a larger fraction of developing immature cells, germ cells and B cells respectively, than the other tissues and thus proteins that function in rapid cell maturation (possibly expressed in supporting cells) may have similar mRNA isoforms across the two tissues.

The EIL distribution per tissue was highly polarized such that most exons either had a very high (>0.80) or very low (<0.20) EIL (Fig. 9). Based on these observations, we decided to parse the skipped exons into tissue specific high EIL and low EIL groups and build a statistical model to determine the variables that contributed to high (and low) exon inclusion levels and thus ultimately predict how SNPs may alter these variables leading to alterations in inclusion levels of the exons near or in which they reside. Note that at a read count less than 10 ($< \log_{10}(10)$), the event distribution was discontinuous, and thus all events with a read count of less than 10 were not used in model development.

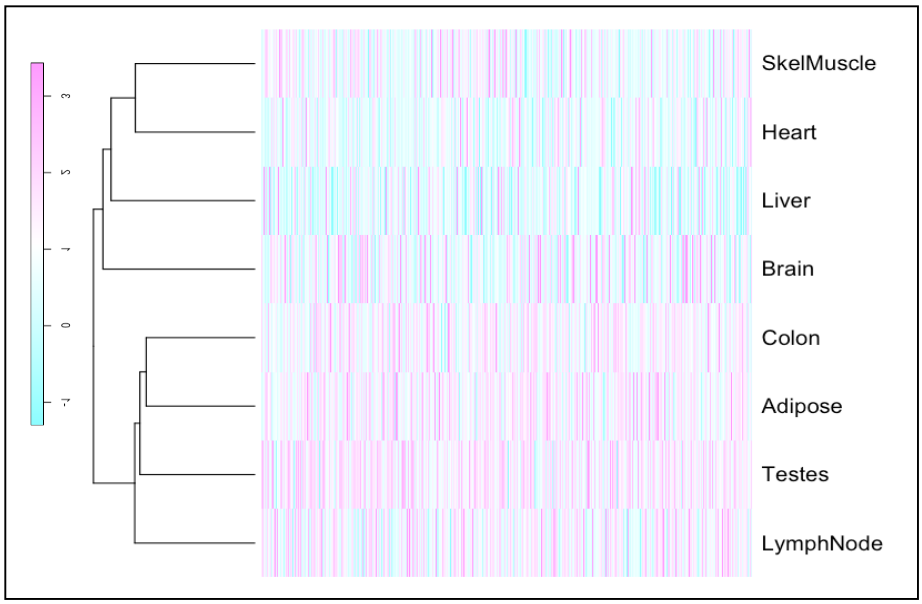


Figure 8: Hierarchical clustering of tissues based on EIL scores in skipped exons
A dendrogram was created by clustering tissues based on similarities of EIL scores. The heatmap shows EIL scores (purple = high; blue = low) for ~30,000 known skipped exons. Exons to which less than 10 reads mapped were not included in the analysis.

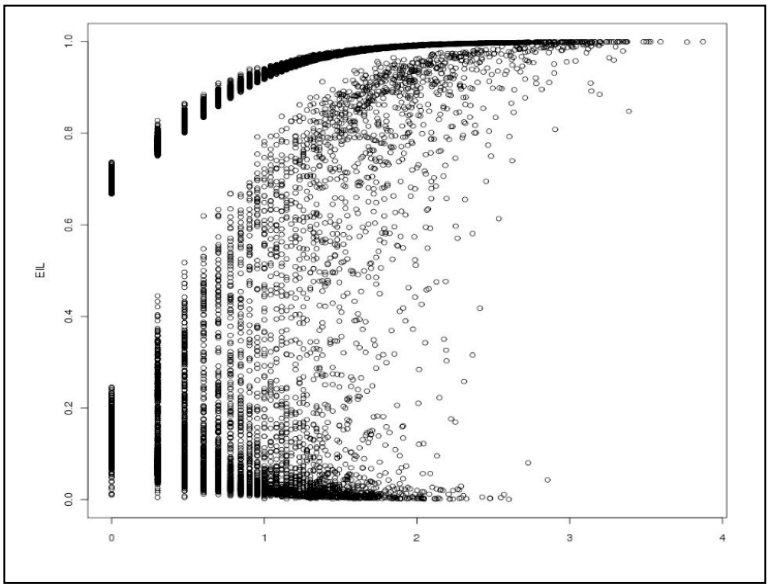
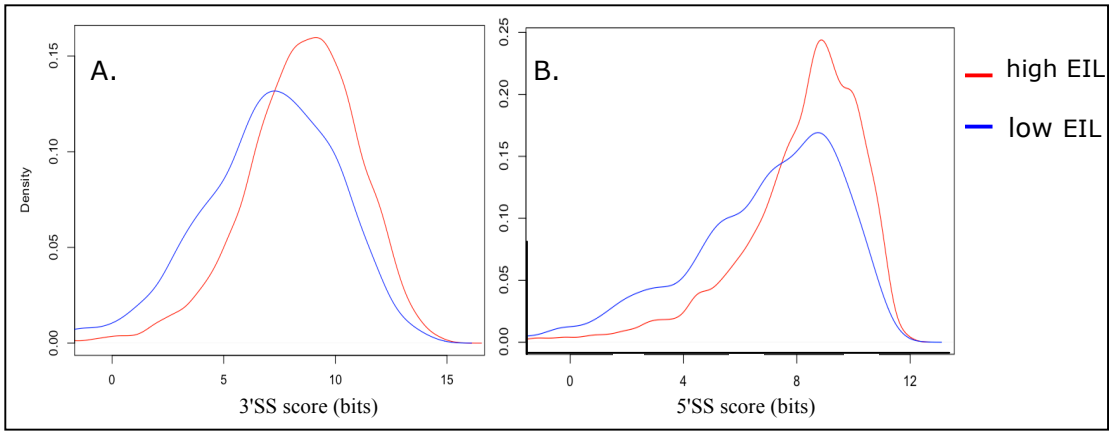


Figure 9: Distribution of EIL in skipped exons in adipose tissue.
EIL values as a function of log-transformed read counts in known skipped exons of adipose tissue. Note the clustering of values at high and low EIL and artifactual EIL distribution at read counts lower than 10.

III. Parameter calculation and results

A total of 144 parameters were calculated for each skipped exon event in either the high or low EIL group. To verify that the parameters represented biologically-plausible values for the events, we performed a number of analyses on specific parameter values. Splice site scores are defined such that a higher score represents that the exonic region in which the splice site occurs has a higher probability of being spliced in by the splicing machinary. Accordingly, both 5' donor and 3' acceptor splice site scores differed significantly between the high- and low-EIL datasets ($p<0.001$ for both splice sites), with high-EIL datasets having higher values for both scores (Fig. 10). It has been shown that alternative exons with stronger splice sites are more likely to be spliced in than those with weaker splice sites, which is corroborated by our data (Xiao et al., 2008) Generally, with respect to motif frequency and score calculation, we expected to see higher frequency of both known and bioinformatically predicted ESE motifs in exons vs. introns in constitutive exons, and the reverse for ESS motifs. The pattern of distribution of both types of motifs confirmed our predictions (Fig. 11). When known ESE and ESS motif frequency was assessed between high and low EIL exons, high EIL exons exhibited ESE and ESS frequency distributions similar to consititutive exons (Fig. 12). Low EIL exons, on the other hand, had a lower frequency of ESEs than high EIL exons (Fig. 12A). ESS frequencies were less distinct between the two groups, although the high EIL exons did exhibit a trend towards lower ESS frequency than low EIL exons (Fig. 12B). Similarly, ESE scores were consistently higher in the high EIL group, and ESS scores were

consistently lower (or stronger) in the low EIL group, with the exceptions of hnRNP H/F and ESS C and F groups (Table 2). These results mostly corroborated with the predictions that splicing enhancer motifs would be stronger and splicing silencing motifs would be weaker in highly spliced-in exons. The motifs in the hnRNP H/F and ESS C and F groups are mostly G-rich motifs likely recognized by hnRNP H/F. Interestingly, this splicing factor has been shown to enhance inclusion in certain contexts, and specifically it has been hypothesized that it competes with the silencing factor hnRNP A1 to bind motifs (Han et al., 2005). In addition, exon and intron length distributions between the two groups were statistically different in a similar manner to previous studies



(Table 2).

Figure 10: Splice site score distribution for high and low EIL events in adipose tissue
Exons with high EIL (>0.80; red) had higher 3'SS (A.) and 5'SS (B.) scores ($p<0.001$ for both splice sites) than exons with low EIL (<0.80, blue) in adipose tissue.

A general feature of conserved exons and surrounding introns is that constitutive exons generally have higher exon conservation but lower intron conservation than known skipped exons (Xing and Lee, 2006). We verified this presumption using our dataset of known constitutive and skipped exons (Fig. 13). With respect to the high- vs. low-EIL

groups we found a very large difference in exon conservation between the two with no difference in intron scores (using adipose tissue as an example, Fig. 14b). When we broke down the dataset into 10% EIL bins (such that the data was grouped into bins of 0-0.10 EIL, 0.10-0.20 EIL, etc.), we found a steady increase in exon conservation with increasing EIL, but the intron conservation, which was low at lowest EIL, rose to a peak at in the 0.20-0.40 EIL group and then dropped down again at highest EIL levels (Fig 14a). The observation that exonic conservation was directly correlated with EIL is not novel, and it has been effectively demonstrated that constitutive exons have higher average conservation than alternative cassette exons (which we also have shown, Fig 13), and that within alternative cassette exons, highly included events have a higher conservation than events with lower inclusion levels (Irimia et al., 2009).

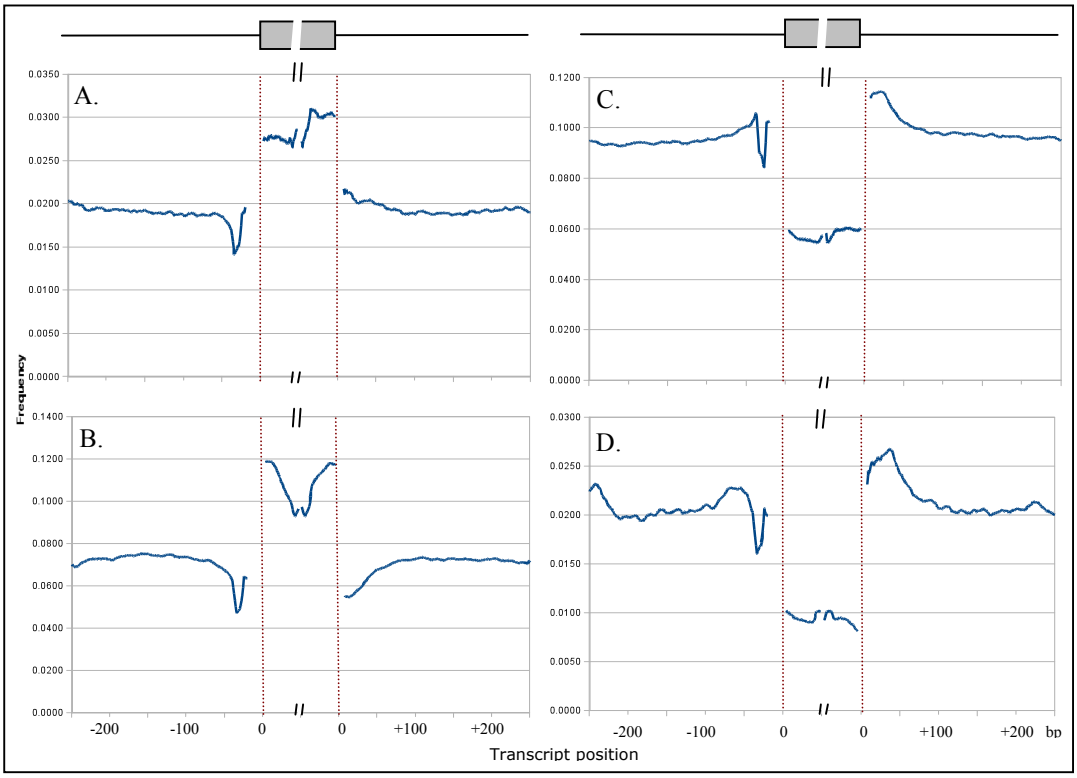


Figure 11: Frequency distribution of ESE and ESS motifs in constitutive exons and introns
Positional frequency was calculated by dividing the number of ESE motifs at each base pair by the total number of events at that position. The first and last 50 bases of exons and first and last 250 bases of introns are plotted. Excluded are splice site regions, specifically the {-20 to +3} and {-3 to +6} positions in 3' and 5'SS, respectively. Curves were smoothed by averaging the value at any base pair to the ten base pairs either directly upstream or downstream of it. A.) Frequency of known ESE motifs, B.) frequency of predicted ESE motifs, C.) frequency of known ESS motifs and D.) frequency of predicted ESS motifs.

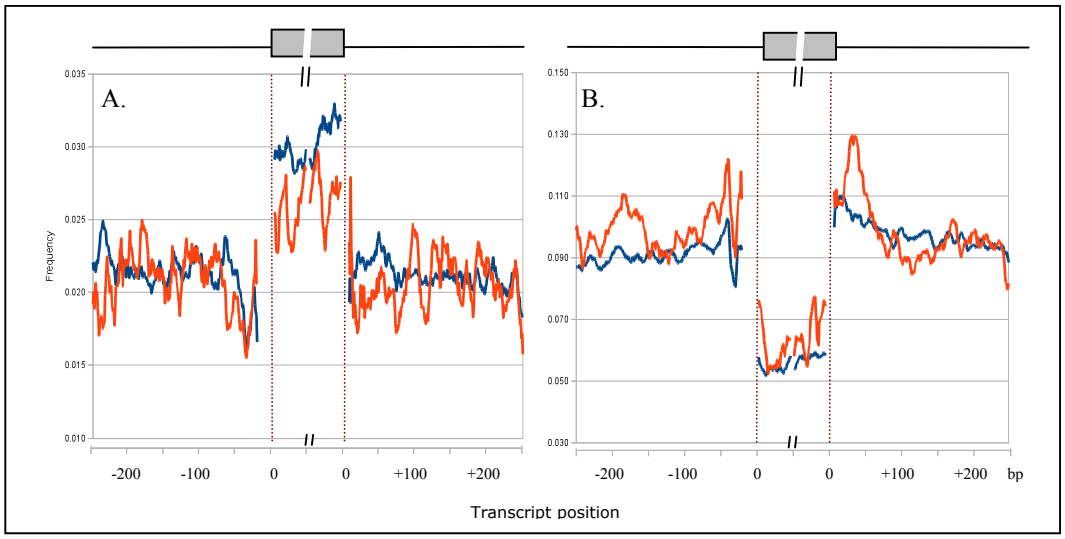


Figure 12: Frequency distribution of ESE and ESS motifs in high and low EIL skipped exons Positional frequency was calculated by dividing the number of ESE or ESS motifs at each base pair by the total number of events at that position. The first and last 50 bases of exons and first and last 250 bases of introns are plotted. Red dashed lines demarcate exon/intron boundaries. Excluded are splice site regions, specifically the {-20 to +3} and {-3 to +6} positions in 3' and 5'SS, respectively. Curves were smoothed by averaging the value at any base pair to the ten base pairs either directly upstream or downstream of it. A.) Frequency of known ESE motifs in high EIL (blue) and low EIL (red) events, C.) frequency of known ESS motifs in high EIL (blue) and low EIL (red) events.

Interestingly, we found higher conservation of introns in exons with intermediate EIL levels compared to both low- and high-EIL exons. While we did not calculate parameter values for these exons, it may be interesting to investigate these values in the future as these events may have higher reliance on intronic regulatory activity than low or high EIL events. While this is speculative, regulation of these events could be more complex than for the other two groups of events and thus may predispose them to a higher level of tissue specificity as the events would be highly dependent on circulating levels of trans factors and other tissue-specific environmental variables.

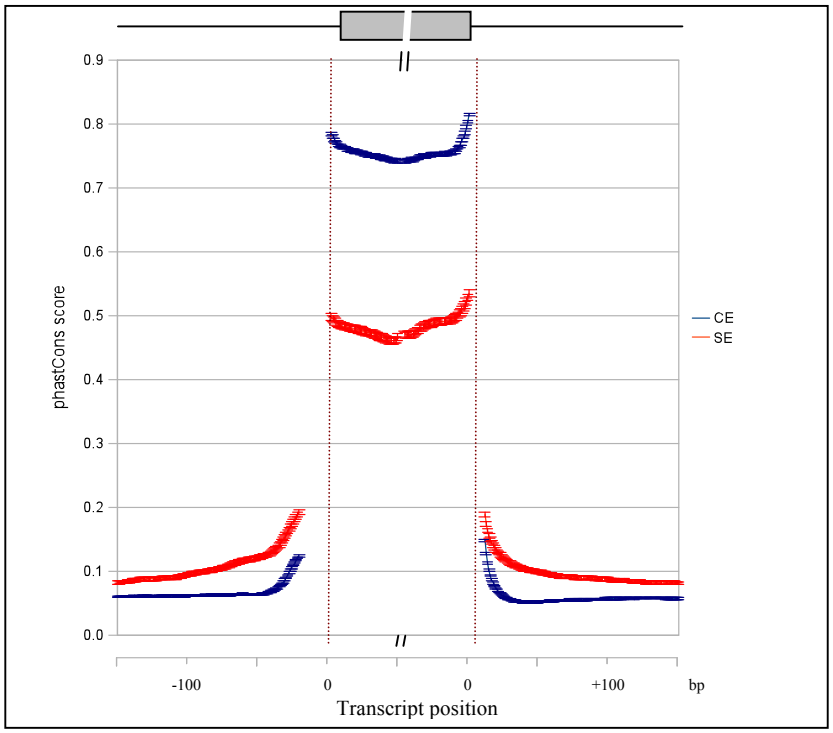


Figure 13: Conservation in exons and introns in skipped vs. constitutive exons and introns
PhastCons scores were calculated for all known skipped (SE) and consititutive (CE) exons and introns.The first and last 50 bases of exons and first and last 150 bases of introns are plotted. Red dashed lines demarcate exon/intron boundaries. Excluded are splice site regions, specifically the {-20 to +3} and {-3 to +6} positions in 3' and 5'SS, respectively.

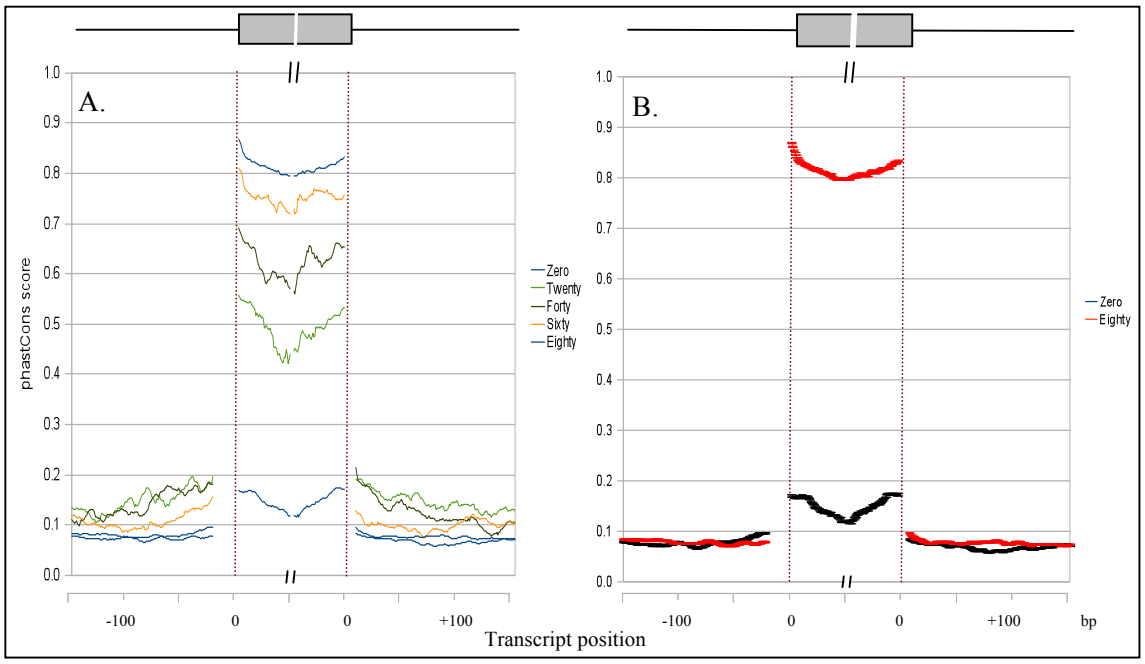


Figure 14: Conservation in exons and introns of skipped exons in adipose tissue classified by EIL
A.) Division of exons by EIL such that events with EIL 0.0-0.20 are grouped into group 'zero', etc. B.) Only events in zero and eighty groups with error bars (+/- 95% confidence intervals). The first and last 50 bases of exons and first and last 150 bases of introns are plotted. Red dashed lines demarcate exon/intron boundaries. Excluded are splice site regions, specifically the {-20 to +3} and {-3 to +6} positions in 3' and 5'SS, respectively.

Attribute	Low EIL (avg)	High EIL (avg)
Exon length	**112.024**	**122.561**
Up length	**203.098**	**193.684**
Upup length	**217.572**	**208.152**
Down length	**218.127**	**205.946**
Downdown length	**202.636**	**190.444**
Up conserve	0.085	0.088
Upup conserve	0.068	0.064
Down conserve	0.074	0.072
Downdown conserve	0.069	0.067
ESE Tra2B score	29.478	38.860
ESE 9G8 score	0.001	0.000
ESE SRp40 score	**170.688**	**281.310**
ESE SRp20 score	**1.473**	**5.576**
ESE SF2 ASF score	**2.580**	**5.456**
ESE SC35 score	2.999	4.917
ESE SRp55 score	**2.575**	**7.517**
ESE 3B score	**0.088**	**0.121**
ESE 3C score	**0.250**	**0.582**
ESE 3E score	**0.745**	**1.227**
ESE 3F score	0.374	0.423
ESE 3H score	0.158	0.189
ESE 5A3G score	**0.353**	**0.557**
ESE 5B3A score	**0.236**	**0.527**
ESE 5C3D score	**5.696**	**8.780**
ESE 5D score	**0.060**	**0.170**
ESE 5E score	**0.038**	**0.067**
ESS MBNL score	**2.510**	**6.072**
ESS CELF score	-17.856	-22.202
ESS PTB score	-119.987	-116.924
ESS_ OX score	-2.635	-2.092
ESS hnRNP A1 score	**-51.598**	**-30.318**
ESS hnRNP H/F score	**-17.794**	**-30.179**
ESS Nova score	**2.357**	**8.078**
ESS SF1 score	-41.268	-35.985
ESS TIA score	**-8.099**	**-6.231**
ESS A score	**0.015**	**0.097**
ESS B score	-0.212	-0.160
ESS C score	**-0.329**	**-0.496**
ESS D/E score	-0.129	-0.095
ESS F score	**-0.019**	**-0.036**
ESS G score	**-0.320**	**-0.137**

Table 2: Comparison of parameter values in adipose tissue between high (>0.80) and low (<0.20) EIL groups.
Bold values represent significant differences between the two groups. Significance was determined using a student's t-test with a p-value cutoff of 0.05. ESE and ESS motifs are first identified by known motif groups followed by bioinformatically predicted groups.

IV. Data classification via Bayesian network modeling

IVa. Model development

Several machine learning methods were tested on the data along with various cutoff points for the two classes. SMO, or sequential minimal optimization, trains support vector machines (SVMs) using polynomial kernels and was developed to allow fast SVM processing of large training sets (Platt 1998). The fast decision tree learner method (FTL) builds a decision tree (a set of rules that can determine which class the data falls into) by using information gain over the variance (Quinlan 1993). A Bayesian network, generally, is a directed, probabilistic cyclical graphical model that represents the conditional dependencies between random variables (Bouckaert, 2004). We tested two different Bayesian network models: one that allowed each node (or variable) to have a maximum of 10 parents (or variables it was dependent on) and one that allowed each node to have a maximum of one parent. We termed these models 'Bayes-10' and 'Bayes-1' , respectively.

The quality of the subsequent models devloped by the various techniques was determined by evaluating the true positive (TP) and false positive (FP) rates, precision, F-measure, ROC area, confusion matrix, and MCC correlation coefficient. A TP rate is the ratio of events which are classified as belonging to a given class, among all events which are truly in that class. It is equivalent to the 'recall' classification in statistics. The FP rate, conversely, is the ratio of events classified as in a given class, but belonging to a different class, among all events not of that class. The precision is the ratio of events which belong to a class over all events classified to that class. Finally, the F-measure

combines precision and recall into a single measure of model performance (2*recall*precision / (precision*recall)) and the confusion matrix combines information about the TP and FP rates for both classes. Namely, if the low EIL class is termed class 'a' and the high EIL class is termed class 'b', a confusion matrix will be a 2x2 matrix with four entries: aa, ab, bb, and ab with entries 'aa' and 'ab' in the first row and 'bb' and 'ba' in the second row. For our purposes, we classified events predicted to belong to the low EIL class that actually belonged to that class ('aa') as true positives and events that were predicted to belong to the high EIL class and were classified as members of that class ('bb') as true negatives. Events classified as belonging to the high EIL class ('b') but that were actually in the low EIL class ('a'), were classified as 'ab', or false negatives and conversely events classified as belonging to the low EIL class ('a') that actually belonged to the high EIL class ('b') were classified as 'ba', or false positives. The confusion matrix was used to calculate the Matthew's Correlation Coefficient (MCC), which measures the quality of two-class classifications and takes into account true and false positives and negatives (Baldi et al., 2000). The MCC can take on values between -'1 to +1' with +1 signifying the model provides perfect classification, a '0' signifiying that the model classifies events randomly, and a '-1' signifying that the model classifies an event in the opposite class to which it actually belongs. In addition, a receiver operating characteristic (ROC) curve is defined by all the possible TP rates on the y-axis and FP rates on the x-axis. In model development, one wishes to maximize the area under the ROC curve (AUC) as this signifies that many points on the curve lie on the upper left hand side of the ROC space, where the TP rate is the highest and the FP rate is the

lowest. Often many versions of a single model are tested and one model is chosen based on on its having the largest AUC among all the models considered (Lasko et al., 2005).

When all models were compared using the analytical parameters described above, the Bayes-10 and Bayes-1 models had generally higher TP rates, F-measures, ROC areas, and MCC scores than either SMO or FTL (Table 3, bottom panel). We chose to use the one-class Bayesian network model over the multi-class model because although the ROC and MCC coefficient was slightly higher in the multi-class model, the TP rate for the low EIL class was higher in the one-parent model and since this class was under-represented in the raw dataset, a higher TP rate gave evidence that the model was most robust at distinguishing between the two classes even before filters were applied. In addition, since events tended to cluster in either very high (>0.80) or very low (<0.20) EIL, we chose those two cutoffs for the two classes. More liberal constraints of (>0.60) or (<0.40) EIL (top panel in Table 3) resulted in lower scores in TP rates, ROC area, and the MCC (Table 3).

Parameter filtering techniques included discretization by applying the Minimum Description Length (MDL) algorithm to decide the partioning of intervals (Fayyad and Irani, 1993) and random balancing (Table 4). The first column represents the output of the model as presented in Table 3, where variable scores were discretized by an unsupervised method. After the parameter scores were discretized using the MDL algorithm, which incoroporates information about the distribution of the scores, the ROC area and MCC scores were improved (second column, Table 4). Next, due to the large discrepency in number of events in each class, events were under-sampled from the high

event dataset in either a non-random manner (by using Tomek links to determine which instances to remove, Souto et al., 2006) or a random manner such that the number of events in both classes would be approximately equal. Results of the former method are displayed in the third column of Table 4, and three different random under-sampling results are displayed in the final three columns. Since the random under-sampling technique resulted in a higher ROC and MCC score than the non-random method and additionally provided consistent results among different random samples, we chose this technique to balance the dataset.

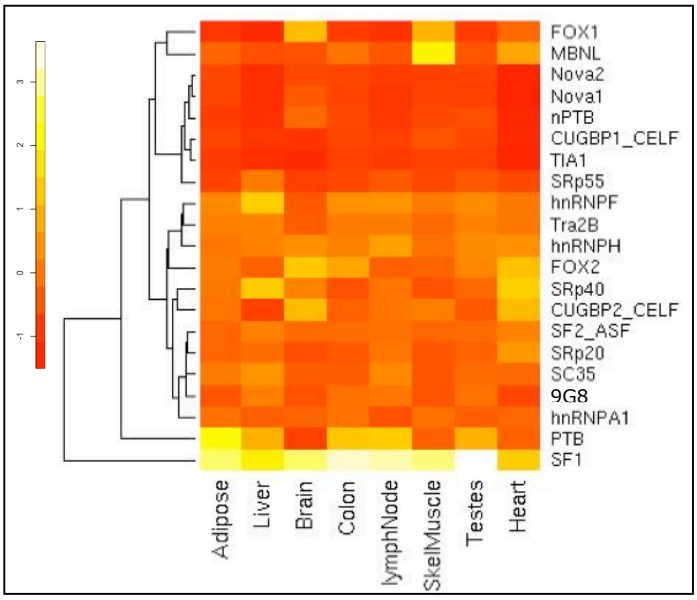
A one-class Bayesian network model was then built for all eight tissues (Table 5). The high ROC areas ranging from .889 to .906 and an average 82% correct classification rate that resulted from applying the model (after the data was balanced and discretized) showed that the model was very good at differentiating between high and low EIL events, even when exon conservation (which was the parameter with the highest difference between the two groups) was removed from the parameter values. This showed that given a certain set of parameter values, the model could be used to predict whether a given alternative exon will be highly or lowly included in a final transcript in a tissue specific manner with an accuracy of approximately 80%.

The tissue-specificity was derived from two sources: first, since EIL of exons varied per tissue, the exons that compromised the high- and low- EIL datasets were specific to each tissue; and second, motif scores were weighted by *trans* factor expression level scores. As evidenced by our analysis of *trans* factor RPKM values (Fig. 15) and previous work, *trans* factor expression levels are often highly tissue specific, for example

note the well-documented brain-specific expression of nPTB, expression of PTB in all tissues except brain, and the muscle-specific expression of MBNL (Rahman et al., 2002; Pascual et al., 2006). Also of note is the clustering of several SR proteins (SRp20, SC35, 9G8, SF2/ASF), all with relatively high RPKMs across all the tissues. As SR proteins are important in constitutive splicing in addition to alternative splicing, their ubiquitous expression is consistent with this function (Ladd and Cooper, 2002). Thus, we were able to create predictive models that were not only statistically robust but specific to the tissue of interest, which allowed for a more fine-grained understanding of alternative splicing and the pathological implications of tested SNPs.

Figure 15: Heat map of tissue-specific expression levels of known *trans* factors

RPKM values for *trans* factors were calculated and used to hierarchically cluster the factors by similarities in RPKM scores across tissues. The RPKM scores were used to weight the corresponding *cis* motif scores in a tissue-specific manner.



An analysis of the most important parameters in model development was performed by measuring the chi-square statistic with respect to the classes. Generally, ESEs, ESSs, 5' splice sites, and 3' splice sites were the attributes with the highest values in all the tissues (Table 6). A surprising finding was that the ESS Nova scores were consistently ranked as largely different between the two groups. As expression of Nova is very low in all tissues except for brain, and we did not expect Nova binding motifs to have such a consistent disparity in scores. One hypothesis for this is that the binding motifs that were in the Nova group have functions that are not related to binding Nova, but that are correlated with splicing of the exon, despite the fact that they have been shown to bind Nova *in vitro* (Matlin et al., 2006). This hypothesis is supported by the finding that average ESS Nova scores were positive for both groups of events, thus indicating that the motifs have functions that are not specific to silencing splicing activity.

IVb. Positive and negative control testing

Positive and negative controls (outside of splice site SNPs) were obtained from a dataset of *in vitro* studies of published SNPs on their ability to affect splicing (Woolfe et al., 2010). While positive controls had been shown to decrease inclusion of the exon (or in one case increase inclusion), negative controls were shown to have no effect *in vitro*. Unfortunately, information on the level of inclusion of the exon with and without the SNP, as well as the magnitude of change in splicing in the positive controls, was not available, nor was tissue-specific splicing information.

A total of 51 SNPs were evaluated by our model: 23 SNPs predicted to decrease exon inclusion (12 near or in alternative exons and 11 near or in constitutive exons), 9 predicted to have no effect on exon inclusion (6 near or in alternative exons and 3 near or in constitutive exons), 1 predicted to increase exon inclusion, and 8 predicted to decrease exon inclusion via disruption of splice sites. Our model was able to correctly predict a change toward lower exon inclusion in at least one tissue in 75% of positive controls (ie. SNPs that were shown to decrease exon inclusion *in vitro)* that were in or near alternative exons (Table 7, Fig. 16). The other 25% of the SNPs were predicted to cause no change. Importantly, none of these SNPs were predicted to cause an increase in inclusion. Thus, the model did not make any predictions opposite to what had been shown to occur *in vitro*. At worst, it failed to predict any change in splicing when it occurred. For positive control SNPs in or near constitutive exons, the model was far less accurate, predicting that less than 20% of the SNPs would decrease inclusion, and the rest would either increase inclusion or have no effect. This can be explained by the fact that our model was trained on alternative exons, and thus SNPs that cause exon skipping of constitutive exons may act via a different mechanism. It is for this reason that we only analyzed GWAS SNPs that were in or near alternative exons for our predictive study (Section V). In addition, 66% of negative controls were predicted to have no effect, and 33% were predicted to cause an increase in inclusion, irrespective of whether the SNPs were in or near alternative or constitutive exons, 88% of SNPs that disrupted splice sites were predicted to decreased inclusion levels, and the one SNP that had been shown to increase inclusion *in vitro* was shown to do so in our model. If we exclude the positive and

negative control SNPs that fell in or near constitutive exons, our model correctly predicts the effect of a SNP at a success rate of 78%. Furthermore, if we decide to narrow our goal to predicting only if SNPs can decrease inclusion of an exon (which one can surmise may be more likely to happen since it is more likely that a SNP will disrupt a *cis* element or splice site than create one) we can make binary the results of the controls as causing decreased inclusion or none to increased inclusion. In this case, the model has a success rate of 87%. In either case, the positive control results gave us confidence that our model would be able to identify novel effects on splicing of GWAS SNPs in or near alternative exons.

Model	SMO	FTL	Bayes - 10	Bayes - 1
Correctly classified instances (count / %)	5864 / 80.219	6056 / 82.845	6190 / 84.679	6017 / 82.312
Incorrectly classified instances (count / %)	1446 / 19.781	1254 / 17.155	1120 / 15.322	1293 /17.688
*Class 00 TP rate	0.021	0.307	0.583	0.663
*Class 00 FP rate	0.002	0.041	0.087	0.137
*Class 00 Precision	0.021	0.651	0.626	0.548
*Class 00 F-measure	0.040	0.417	0.603	0.600
Class 60 TP rate	0.998	0.959	0.913	0.863
Class 60 FP rate	0.979	0.693	0.417	0.337
Class 60 Precision	0.803	0.847	0.897	0.911
Class 60 F-measure	0.890	0.899	0.905	0.886
ROC area	0.509	0.730	0.856	0.853
Confusion matrix (a=00, b=60)				
aa	30	449	852	970
ab	1432	1013	610	492
bb	5834	5607	5338	5047
ab	14	241	510	801
MCC correlation coefficient	0.094	0.364	0.509	0.492
Correctly classified instances (count / %)	5544 / 81.565	5677 / 83.522	5844 / 85.979	5696 / 83.802
Incorrectly classified instances (count/ %)	1253 / 18.435	1120 / 16.478	953 / 14.021	1101 / 16.198
Class 00 TP rate	0.066	0.334	0.609	0.678
Class 00 FP rate	0.005	0.045	0.080	0.124
Class 00 Precision	0.761	0.640	0.645	0.567
Class 00 F-measure	0.121	0.439	0.626	0.618
Class 80 TP rate	0.995	0.955	0.920	0.876
Class 80 FP rate	0.934	0.666	0.391	0.322
Class 80 Precision	0.817	0.857	0.908	0.919
Class 80 F-measure	0.897	0.903	0.914	0.897
ROC area	0.530	0.746	0.878	0.871
Confusion matrix (a=00, b=80)				
aa	86	438	799	890
ab	1226	874	513	422
bb	5458	5239	5045	4806
ab	27	246	440	679
MCC correlation coefficient	0.187	0.379	0.541	0.519

Table 3: Performance of various statistical measures to classify skipped exon events as 'low' and high' EIL in adipose tissue.
(*) signifies that the 'Class 00' in the top panel encompassed a group of events between 0-0.40 EIL whereas in the bottom panel, 'Class 00' encompassed events between 0-0.20 EIL. 'Class 60' corresponds to events with EIL between 0.60-1 and 'Class 80' correspondes to events with EIL between 0.80-1.

	Adipose	Adipose *	Adipose **	Adipose ***	Adipose (b) ***	Adipose (c) ***
Bayes log-odds score	-176395.17	-181340.01	-60395.75	-60276.17	-59905.52	-63128.72
Correctly classified instances (count / %)	5696/83.802	5775/84.964	2127/81.090	2156/82.165	2141/81/561	2160/82.286
Incorrectly classified instances (count/ %)	1101/16.198	1022/15.036	496/18.909	468/17.835	484/18.438	465/17.7143
Class 00 TP rate	0.678	0.726	0.818	0.830	0.823	0.839
Class 00 FP rate	0.124	0.121	0.195	0.187	0.192	0.193
Class 00 Precision	0.567	0.590	0.798	0.816	0.811	0.813
Class 00 F-measure	0.618	0.651	0.807	0.823	0.817	0.826
Class 80 TP rate	0.876	0.879	0.805	0.813	0.808	0.807
Class 80 FP rate	0.322	0.274	0.182	0.170	0.177	0.161
Class 80 Precision	0.919	0.931	0.824	0.827	0.821	0.834
Class 80 F-measure	0.897	0.904	0.814	0.820	0.814	0.820
ROC area	0.871	0.893	0.891	0.902	0.899	0.897
Confusion matrix (a=00, b=80)						
aa	890	952	1040	1089	1080	1101
ab	422	360	232	223	232	211
bb	4806	4823	1087	1067	1061	1059
ab	679	662	264	245	252	254
MCC correlation coefficient	0.519	0.561	0.622	0.643	0.631	0.646

* = No balance / discretized
** = Non-random balance / discretized
*** = Random balance / discretized

Table 4: Refining the Bayesian network model with discretization and random sampling
After selection of the Bayesian network model, the data was discretized (*,**,***) and resampled due to the large discrepency in sample size between the two classes (**,***).

	Adipose	Liver	Testes	Skeletal muscle	Lymph node	Brain	Colon	Heart
Bayes log-odds score	-60276.17	-14059.75	-73908.76	-33324.10	-44221.03	-24683.88	-49691.58	-20085.89
Correctly classified instances (count / %)	2156/82.165	799/81.610	2463/81.529	1373/82.961	1737/80.528	1182/82.140	1839/80.840	1034/83.454
Incorrectly classified instances (count/ %)	468/17.835	180/18.392	558/18.470	282/17.029	420/19.471	257/17.851	436/19.162	205/16.545
Class 00 TP rate	0.830	0.824	0.826	0.837	0.822	0.829	0.829	0.832
Class 00 FP rate	0.187	0.192	0.195	0.178	0.211	0.186	0.213	0.163
Class 00 Precision	0.816	0.811	0.809	0.825	0.795	0.816	0.796	0.836
Class 00 F-measure	0.823	0.817	0.817	0.831	0.808	0.823	0.812	0.834
Class 80 TP rate	0.813	0.808	0.805	0.822	0.789	0.814	0.787	0.837
Class 80 FP rate	0.170	0.176	0.174	0.163	0.178	0.171	0.171	0.168
Class 80 Precision	0.827	0.822	0.822	0.835	0.816	0.827	0.822	0.833
Class 80 F-measure	0.820	0.815	0.813	0.828	0.802	0.820	0.804	0.835
ROC area	0.902	0.906	0.891	0.906	0.885	0.893	0.896	0.905
Confusion matrix (a=00, b=80)								
aa	1089	403	1247	692	886	596	943	515
ab	223	86	263	135	192	123	194	104
bb	1067	396	1216	681	851	586	896	519
ab	245	94	295	147	228	134	242	101
MCC correlation coefficient	0.643	0.632	0.631	0.659	0.611	0.643	0.617	0.669

Table 5: Bayesian network model for eight human tissues
Analytics for the bayesian network model that was constructed for the eight human tissues based on the parameter values of the high and low EIL exons that compromised the respective tissues.

Adipose		Liver		Colon		Brain	
259.12	ESS Nova	128.65	ESS PTB	211.30	5' SS	194.04	ESS Nova
250.00	ESS PTB	124.44	ESE 5C3D	208.37	ESS PTB	152.16	ESS TIA
239.00	ESS TIA	120.82	ESS TIA	207.62	ESS Nova	145.43	ESS PTB
234.81	ESE SRp40	110.91	ESS Nova	203.16	ESE SRp40	140.68	3' SS
222.68	ESE 5C3D	89.92	ESE 3E	193.09	3' SS	134.35	ESE 5C3D
202.29	3' SS	87.90	ESE SRp40	160.92	ESE 3C	126.72	ESE SRp40
187.13	5' SS	86.11	ESE 5B3A	152.69	ESE 5B3A	114.67	ESE 5B3A
185.86	ESE 5B3A	83.20	3' SS	152.00	ESS TIA	110.9	ESE 3E
183.90	ESS hnRNP A1	81.74	5' SS	149.28	ESS hnRNP A1	100.49	ESE 3C
179.79	ESE 3F	69.02	ESE 3C	148.60	ESE 3E	97.35	ESS MBNL
163.06	ESE 3C	65.70	ESS hnRNP H/F	136.00	ESE 5C3D	94.87	ESE 5A3G
156.56	ESS MBNL	58.96	ESS hnRNP A1	132.26	ESE 5A3G	86.91	ESS hnRNP H/F
152.14	ESE 5A3G	52.86	ESS CELF	131.88	ESS MBNL	84.99	5' SS
149.07	ESE 3E	47.72	ESE 5A3G	108.20	ESS hnRNP H/F	81.21	ESE 5D
119.80	ESS hnRNP H/F	45.07	up_conserve	106.58	ESS CELF	70.79	ESE 3B
Heart		**SkelMuscle**		**LymphNode**		**Testes**	
141.59	ESS PTB	186.58	ESS PTB	199.56	ESS Nova	337.49	ESS Nova
140.72	ESS TIA	180.07	ESS TIA	181.27	ESS TIA	294.42	ESS PTB
120.54	ESS Nova	162.66	ESS Nova	177.87	5' SS	293.18	ESE SRp40
96.15	ESE SRp40	155.67	ESE 5C3D	174.83	3' SS	232.11	ESS TIA
94.93	ESE 3E	146.34	ESE 3E	174.79	ESS PTB	231.65	5' SS
94.15	ESE 5C3D	138.63	ESE SRp40	170.57	ESE 5C3D	224.72	3' SS
92.17	5' SS	137.19	ESE 3C	169.54	ESE SRp40	200.29	ESS hnRNP A1
88.41	3' SS	135.54	3' SS	136.48	ESS hnRNP A1	197.78	ESE 5C3D
86.07	ESE 3C	135.02	ESE 5A3G	130.59	ESE 3E	191.81	ESE 3E
84.31	ESE 5B3A score	132.86	5' SS	129.23	ESE 5B3A	190.76	ESE 3C
79.94	ESE 5A3G score	125.24	ESS MBNL	129.02	ESE 5A3G	181.77	ESE 5B3A
79.03	ESS hnRNP A1	118.5	ESE 5B3A	111.79	ESE 5D	173.99	ESS hnRNP H/F
74.86	ESS MBNL	110.6	ESS hnRNP A1	108.89	ESE 3C	152.56	ESS MBNL
67.98	ESS hnRNP H/F	105.74	ESS hnRNP H/F	102.91	ESS hnRNP H/F	138.57	ESE 5A3G
66.52	ESE 3F	76.02	ESE Tra2B	96.93	ESS FOX	127.04	ESS CELF

Table 6: Chi-square calculation of important attributes per tissue
The 15 attributes with the highest chi-square scores per tissue are listed with the score on the left hand side and the attribute name on the right-hand side.

	Adipose	Liver	Testes	Skel muscle	Lymph node	Brain	Colon	Heart			
Pos_SE											
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_SRp40_score:3.742:-1.706		
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.145	ESS_Nova_score:-8.094:-7.753		
	-0.011	**0.000**	0.007	-0.251	-0.082	-0.052	-0.070	0.000	ESS_PTB_score:-35.629:-45.512	ESS_SF1_score:0.000:-9.684	
	0.170	-0.346	-0.373	-0.196	**0.204**	0.000	-0.257	-0.283	ESS_hnRNP_A1_score:0.000:-8.461		
	0.007	0.052	0.005	0.122	0.055	-0.015	0.064	**0.085**	ESE_SRp40_score:11.020:22.521	ESS_MBNL_score:18.619:0.000	ESS_C_score:0.000:-3.173
	0.011	0.219	**0.398**	0.075	0.000	0.244	0.234	0.204	ESE_3B_score:0.712:0.000		
	0.000	0.000	**0.221**	0.000	0.000	0.000	0.000	0.000	ESE_5B3A_score:0.965:1.608		
	0.003	0.000	0.016	**0.132**	0.000	0.021	-0.030	0.177	ESE_Tra2B_score:35.291:17.108	ESE_5C3D_score:14.547:2.665	
	0.013	0.006	0.000	0.001	**0.236**	0.071	-0.008	0.071	ESE_5A3G_score:1.035:0.000	ESS_CELF_score:-52.412:-50.041	ESS_B_score:0.000:-0.728
	0.207	0.084	0.005	**0.088**	-0.654	0.073	0.082	0.033	ESE_SRp40_score:39.785:19.278	ESE_5D_score:5.221:0.000	ESS_C_score:0.000:-2.380
	0.000	0.079	0.113	0.000	0.000	**0.156**	0.200	0.513	ESE_SRp40_score:13.080:5.445		
	0.220	0.163	0.209	0.186	**0.403**	0.349	0.420	0.000	ESE_3F_score:0.544:0.000	ESE_5A3G_score:1.176:0.000	ESE_5C3D_score:17.008:9.020
Pos_CE											
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_5C3D_score:21.605:21.024		
	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.005	ESE_3E_score:1.508:0.538	ESE_5A3G_score:1.352:0.000	
	0.119	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_C_score:-3.410:-6.538		
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_3F_score:1.371:1.012		
	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.005	ESE_3B_score:0.803:0.479	ESE_5C3D_score:49.198:50.281	ESS_hnRNPA1_score:-22.714:-34.508
	-0.402	-0.373	-0.349	-0.482	-0.463	**-0.198**	-0.258	-0.057	ESS_hnRNP_A1_score:0.000:-9.073		
	0.033	0.015	0.033	**0.184**	0.000	0.002	0.005	0.061	ESE_3C_score:2.249:0.000	ESE_5C3D_score:0.155:0.590	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_5A3G_score:1.556:0.000	ESE_5C3D_score:3.644:3.465	ESS_hnRNPA1_score:0.000:-9.234
	-0.052	**0.000**	-0.007	-0.208	-0.072	-0.201	-0.009	0.000	ESS_SF1_score:0.000:-9.428	ESS_TIA_score:-8.751:-14.247	
	-0.053	-0.039	**-0.008**	-0.298	-0.074	-0.188	-0.010	-0.029	ESS_PTB_score:0.000:-8.872	ESS_hnRNP_A1_score:-18.823:-27.890	
	0.000	0.000	-0.006	-0.205	0.000	0.000	0.000	0.000	ESS_B_score:0.000:-2.906		
Neg_SE											
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_SRp40_score:7.631:22.897	ESS_Nova_score:0.319:0.000	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_hnRNP_H_F_score:-0.038:-0.042		
	0.000	0.000	0.000	0.000	0.000	0.000	**-0.017**	0.000	ESE_3E_score:3.492:2.438	ESE_5C3D_score:54.159:50.816	
	0.000	0.000	0.000	0.000	**-0.197**	0.000	0.000	0.000	ESE_5C3D_score:54.159:38.435		
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_5C3D_score:54.159:43.076		
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_PTB_score:-26.562:-26.843		
Neg_CE											
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_SRp40_score:21.072:16.893	ESE_SF2_ASF_score:3.443:0.000	ESE_SRp55_score:3.725:1.282 / ESS_PTB_score:-1.447:-2.454
	-0.080	-0.028	-0.057	-0.021	-0.054	-0.044	-0.030	-0.054	ESS_PTB_score:-8.069:-16.943	ESS_A_score:0.000:0.666	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESE_5C3D_score:66.453:60.610		
Incre_EI											
	-0.320	-0.198	**-0.478**	-0.435	-0.343	-0.324	-0.308	-0.466	ESE_Tra2B_score:0.000:17.958	ESE_5C3D_score:1.941:14.659	
Splice site											
	0.006	0.000	0.006	0.002	0.006	0.000	**0.008**	0.002	3'SS:9.220:0.620		
	0.049	0.084	**0.104**	0.180	0.095	0.043	0.059	0.143	5'SS:10.240:1.730		
	0.023	0.054	0.023	0.055	0.031	0.010	**0.157**	0.003	3'SS:10.930:2.980		
	0.023	0.054	0.023	0.055	0.031	0.010	**0.157**	0.003	3'SS:10.930:2.980		
	0.201	0.247	0.180	0.123	0.171	0.243	0.113	**0.304**	5'SS:8.530:6.860		
	0.000	0.000	0.000	0.000	0.000	0.000	**0.035**	0.000	5'SS:10.650:7.790		
	0.247	0.216	0.107	0.002	0.047	0.240	0.130	**0.285**	5'SS:10.930:2.750		
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	3'SS:2.29:-5.670		

Table 7: Prediction of effects of SNPs known to alter levels of exon splicing *in vitro*

Positive control SNPs were grouped into 6 groups: SNPs near known skipped (Pos_SE) or constitutive (Pos_CE) exons that caused an increase in the probability that the exon would be skipped, or caused no change in this probability (Neg_SE, Neg_CE). In addition, one SNP had been shown to decrease this probability, or effectively increase the EI (Incr_EI). SNPs known to affect splice sites were grouped in the Splice site category. Shown are the values for each tissue that were computed by subtracting the probability that the mutant event would be in the low EI group from the probability that event with the reference genomic sequence would be in the low EI group. Positive values represent an increase in the probability that the exon would be skipped, while negative values represent a decrease in the probability that the exon would be skipped. The value changes of the parameters in the right-most column are for adipose tissue. Bolded values represent the value used for analysis in Fig 15.

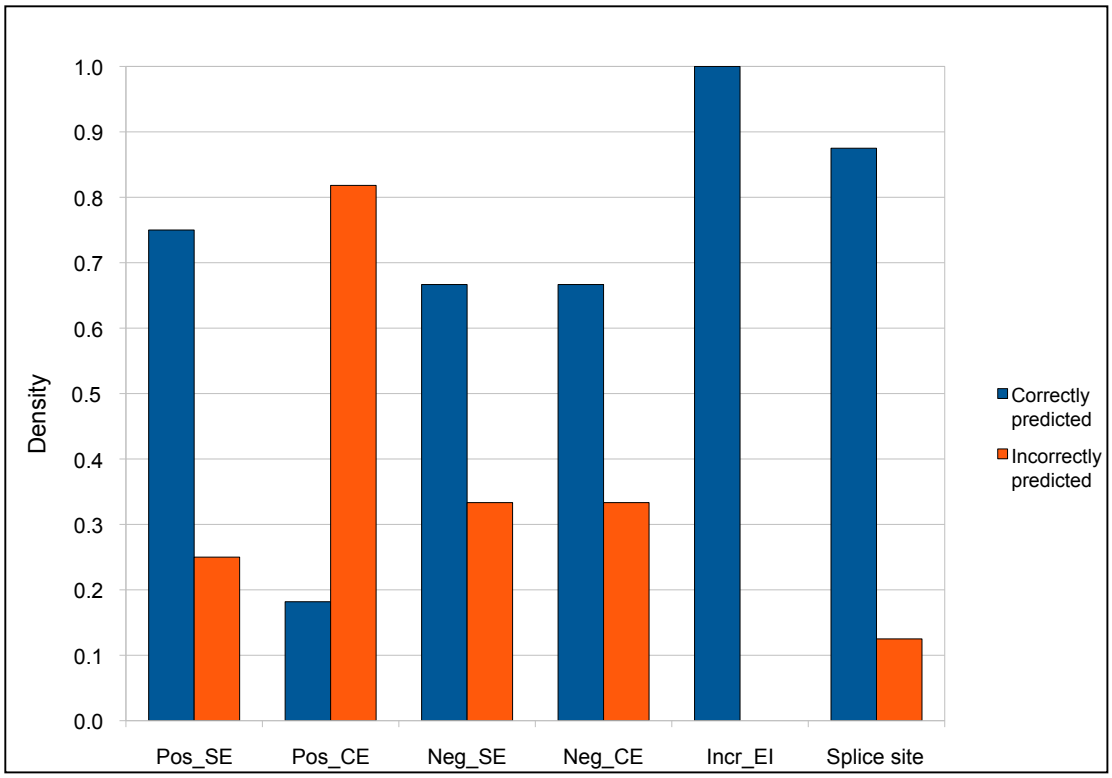


Figure 16: Summary of results of positive control test of the Bayesian network classifier
Results from Table 6 were summarized such that for every event in a SNP group, the prediction for effect of the SNP on splicing was classified as 'Correctly predicted' if it was consistent with the *in vitro* results and 'Incorrectly predicted' if it was not consistent with the results. The density was calculated as the fraction of SNPs that were correctly or incorrectly predicted, respectively, of the total amount of SNPs in that group.

V. GWAS SNPs

SNPs from GWASs were used to test whether they would affect splicing of exons near or in which they resided. SNPs in LD with the GWAS SNPs were also used in the model as molecular causality of the phenotype that GWAS SNPs are correlated with may be due to SNPs that they are in LD with. LD data was collected using the UCSC genome browser. Only SNPs in or near skipped exons were tested due to the high power of the model for predicting effects of SNPs near such events. Results were parsed by tissue.

The SNP with the highest change in splicing from a group of LD SNPs was chosen as the most likely causal candidate from that group. SNPs were ranked by the magnitude of the values that were calculated when the probability that the exon with the reference allele was in the low EIL category was subtracted from the probability that the exon with the SNP allele was in the low EIL category. Magnitude per SNP was used as the highest magnitude of change in any one tissue. Variant alleles that resulted in large changes in the probability that the exons which they were in or near would be spliced out were noted (Table 8). A table with effects of all GWAS SNPs is available upon request*. In addition, haplotype and phase data for SNPs was obtained using the UCSC human genome browser and the International HapMap Consortium databases. Effects on splicing of LD SNPs that corresponded to a causal GWAS SNP and that occured near and / or in the same skipped exon were calculated using the model, thus allowing for an assessment of cumulative effects of LD SNPs on the phenotype. As with the single SNPs, results were ranked by the haplotype that caused the largest magnitude change in any tissue (Table 9*).

Of the ~3000 thousand SNPs currently listed in the NIH GWAS database, we found ~500 that were in or near alternative exons, and of those, our model predicted any change in splicing due to ~200 of them. Table 8 shows the top 25 SNPs with the highest predicted change in splicing in any tissue. One interesting output of our model is the association of GWAS SNP rs2338104 and its LD SNP rs 2058804 with decreased HDL cholesterol. HDL is a lipid carrier in the blood and is generally thought to play a protective role from lipid accumulation in the vascular walls, and low concentrations of

HDL have been shown to be associated with increased risk for coronary heart disease (Holleboom et al., 2008). While the GWAS SNP is located in the locus of the MVK (murine mevalonate kinase) gene, which catalyzes an early step in the synthesis of cholesterol, the LD SNP is located in the UBE3B gene, a member of the ubiquitin ligase family (Fogarty et al., 2010). Our model predicts that only in liver tissue, an exon will be likely to be excluded from the UBE3B gene. We can hypothesize that this would cause decreased function of the corresponding protein (if the protein is produced at all), leading to weakening of it's ubiquitination activity. This could have the potential to affect proteins and processes involved in cholesterol biosynthesis and serum lipid levels, including MVK. Another interesting finding was that a SNP (rs3771202) in LD with a GWAS SNP (rs2310173), both in the IL1R2 gene, was predicted to cause exon inclusion only in lymph tissue, which was associated with increased risk for ankylosing spondylitis. This disease is a form of inflammatory arthritis that is characterized by inflammation in the spine and joints (TASC, 2010). IL1R2 encodes a decoy receptor that binds to the pro-inflammatory signaling molecule IL1 and thus prevents it from binding its native receptor, IL1R, and causing a pro-inflammatory response. If proper splicing is disrupted in the IL1R2 gene, one could hypothesize that response to IL1 would be amplified since IL1R2 would not be as capable in buffering the signal (Bellehumeur et al., 2009). This finding correlates with the profile of increased inflammation in patients with this disease.

Our haplotype analysis, where we assayed for multi-SNP action on any alternative exon, could hypothetically lead to two results: 1. A new model of splicing disruption that is different from the predictions made by the model for each single SNP,

or 2. A haplotype model that is identical to one of the SNPs. For all exons which had two or more surveyed SNPs, we were only able to identify a small subset that fit into the first category (Table 9). This result was primarily due to the fact that even most SNPs that caused a change in a parameter did not cause the model to change its prediction with respect to exon inclusion. To note in the haplotypes that did show a novel score that was different from any of the single SNPs, there exist 3 SNPs in LD in the TLE1 gene (rs2796465, rs911638) that are in LD with the GWAS SNP rs10867752 that is correlated with hippocampal atrophy. rs11792087 contributes the main prediction of the model with respect to this haplotype, namely a .423 increase in probability that an exon in the TLE1 gene will not be highly included in the final transcript in brain tissue. rs2796465 contributes a mild (0.074) increase in the probability for exon inclusion in adipose tissue while the last SNP has no effect. In this case, the changes in parameter values caused by each SNP are additive. TLE1 is a part of the transducin-like Enhancer of split family and has been shown to play an important role in inhibition of neuroprogenitor cell development into neurons during mammalian forebrain development (Buscarlet et al., 2009). A poorly spliced product in brain tissue may interfere with proper maintenance of the neural stem cell pool in hippocampal tissue and other mechanisms of neuronal maintenance and thus lead to a predisposition towards hippocampal atrophy.

Another example can be found in LD SNPs in the ERAP1 gene, which are correlated with predisposition to ankylosing spondilytis. ERAP1 (endoplasmic reticulum aminopeptidase 1) is involved in peptide trimming for MHC-1 antigen presentation in the endoplasmic reticulum (Haroon et al., 2010). In this case, multi-SNP effects are more

complex. In the adipose tissue, for example, rs26653 causes an increase in the cumulative ESE SRp40 score, and the model predicts that the probability that the exon will have a low EIL to be lowered by .118. Yet, rs3734016 causes a decrease in the ESS A motif group score (thus creating a stronger ESS signal), and leads to a prediction of an increase in the probability that the exon will have a low EIL by .302. In the final haplotype prediction, the two SNPs cause the model to yield a prediction that there will be only a very slight (0.008) increase in the probability that the exon will have a low EIL. Therefore, the model incorporates the multiple parameter changes to derive a new prediction that is different from the individual predictions.

Conclusions and future perspectives

The main purpose of this thesis was to derive a predictive model of splicing regulation in the human genome that incorporated the many parameters known to influence splicing at the tissue-specific level. The model would be used to better understand the differences in magnitudes of the parameters between exons that were more vs. less highly included in the splicing reaction and to predict how SNPs may alter the splicing level of exons near which they were located. Thus, the model and its application would provide a global splicing-centric SNP database that could be then utilized to experimentally validate the functional consequences of these SNPs.

I shall argue that the thesis was successful with respect to this purpose and that while the model did have several drawbacks, which I will discuss and which can be improved upon in future versions, overall the results were in support of previous studies

in this field and contributed to the knowledge of combinatorial mechanisms of splicing regulation alongside providing a biomedically useful set of functional predictions with respect to deleterious polymorphisms.

A drawback to our model that we have hitherto not mentioned is its inability to predict SNP-centric genomic changes. There are SNPs that have been shown to cause changes in splicing by creating premature stop codons or uncovering decoy splice sites, for example (Ward and Cooper, 2009). Our algorithm, since it was focused on testing genome-centric features, could not account for creation of specific factors by the SNPs. Thus, if a SNP created a specific event that altered splicing, our algorithm was not able to detect it. In future revisions, it will be optimal to have both SNP-centric and genome-centric features embedded in a predictive model.

Experimental validation of these results is critical to verifying the accuracy of our model. Thus, tissue-specific predictions of alteration in splicing by SNPs will need to be tested *in vitro* via assays in different tissue-specific cell lines to verify the strength of our conclusions. In addition, development of models that analyze alternative 3' and 5' splice site creation will be important in order to create a more global model of SNP-mediated effects on all types of alternative splicing.

Systems biology, as compared to traditional molecular biology, is the study of the interactions of the many physical components that make up the biological system under investigation (Lander, 2010). In this thesis, we examine such interactions with respect to the different *cis* and *trans* elements that regulate alternative splicing and then use this understanding to predict effects of the disruption of this machinery on the system as a

whole, thus we can classify this work as an addition to knowledge in 'systems biology of alternative splicing'. In this thesis, I have used machine learning combined with high-throughput RNA-Seq data and information from GWAS studies to identify the combinatorial molecular framework that is utilized to include or exclude exons during the alternative splicing reaction and we have used our knowledge of this framework to make predictions regarding effects of SNPs on splicing of individual exons, thus setting the stage for further studies and experimental validation of the biological and biomedical implications of our results.

* Full results of GWAS and / or haplotype analysis are available upon request to Dr. Grace Xiao (gxxiao@ucla.edu

Gene	Disease	GWAS SNP	LD SNP	Adipose	Liver	Testes	Sk Muscle	Lymph	Brain	Colon	Heart	
PECR	Alcohol dependence	rs7590720	rs1429148	-0.197	-0.045	-0.194	-0.181	-0.036	-0.796	-0.615	0.000	ESE_5C3D_score:4.162:1.046, ESS_FOX_score:-0.303:-3.006
LDLR	Myocardial infarction (early onset)	rs1122608	rs688	0.007	0.084	0.005	0.081	-0.654	0.068	0.088	0.033	ESE_SRp40_score:9466.136:4586.866, ESE_5D_score:5.221:0.000
IL10	Ulcerative colitis	rs3024493	rs1150258	-0.624	-0.335	-0.558	-0.019	0.034	-0.020	0.051	0.023	ESS_C_score:0.000:-2.380 ESE_5C3D_score:0.008:2.216, ESE_5E_score:2.166:0.000
ATL1	Cognitive test performance	rs17122693	rs2073348	0.000	0.608	0.000	0.000	0.000	0.000	0.000	0.000	ESS_Nova_score:-26.053:-50.681
ACADM	Serum metabolites	rs211718	rs3818855	0.259	0.261	0.605	0.009	0.245	0.113	0.541	0.207	ESS_MBNL_score:40.196:-50.514, ESS_TIA_score:-58.762:-116.411
CPEB1	Chronic lymphocytic leukemia	rs783540	rs9806591	-0.052	-0.567	-0.213	-0.027	-0.159	-0.096	-0.291	-0.370	ESS_hnRNP_H_F_score:0.000:-56.425, ESS_C_score:0.000:-0.878
FRMD6/GNG2	Hippocampal atrophy	rs11626056	rs3825600	0.041	0.000	0.050	0.000	0.120	0.557	0.116	0.552	ESS_PTB_score:-552.830:-990.482, ESS_Nova_score:-3.175:0.332 ESS_TIA_score:-296.356:-264.234
CTDSPL	Prostate cancer	rs9311171	rs156265	-0.320	-0.521	-0.211	-0.174	-0.233	-0.340	-0.252	-0.118	ESE_SF2_ASF_score:-6.659:27.588, ESE_5C3D_score:35.782:16.451
MAML2	Response to treatment for acute lymphoblastic leukemia	rs7115578	rs11549808	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.515	ESS_PTB_score:-677.145:-865.751
ACSM1/ERI2	Schizophrenia	rs433598	rs2301771	0.473	0.366	0.058	0.043	0.167	0.380	0.357	0.288	ESE_SRp40_score:3267.433:0.000, ESE_3F_score:1.797:0.115
MVK/UBE3B	HDL cholesterol	rs2338104	rs2058804	0.000	0.466	0.010	0.008	0.007	0.066	0.002	0.000	ESS_CELF_score:-17.544:-1268.250
CETP	HDL cholesterol	rs3764261	rs5883	-0.447	-0.148	0.000	-0.227	0.000	-0.402	-0.140	0.000	ESS_Nova_score:2.795:1.826, ESS_TIA_score:-2.198:-4.676
UST	Response to antidepressants	rs2500535	rs10782309	-0.189	-0.285	-0.310	-0.202	-0.118	-0.178	-0.441	-0.263	5Pmaxent:5.180:10.150
RP11/TLE1	Hippocampal atrophy	rs10867752	rs11792087	0.000	0.000	0.000	0.000	0.000	0.423	0.000	0.000	ESE_5B3A_score:12.924:12.873, ESS_FOX_score:-0.043:-0.058
FBX07	Hematological parameters	rs9609565	rs9726	0.215	0.404	0.124	0.380	0.235	0.029	0.183	-0.275	ESE_SRp40_score:486.309:0.000, ESS_PTB_score:-844.852:-663.542 ESS_TIA_score:-25.935:-46.008,ESS_B_score:0.000:-0.643
CD69/KLRF1	Type 1 diabetes	rs4763879	rs2232548	-0.275	-0.134	-0.399	-0.108	-0.244	-0.104	-0.048	-0.296	ESS_PTB_score:0.000:-412.997, ESS_TIA_score:-2.281:-23.062
IL1R2	Ankylosing spondylitis	rs2310173	rs3771202	0.000	0.000	0.000	0.000	-0.280	0.000	0.000	0.000	ESE_SRp40_score:0.000:60.785
CRP/SLAMF8	C-reactive protein	rs7553007	rs17365054	0.004	0.378	0.018	0.230	0.001	0.316	0.001	0.174	3pmaxent:6.580:-1.780
MTHFD1L	Coronary disease	rs6922269	rs572522	0.000	0.000	0.000	0.000	0.000	0.378	0.000	0.256	ESS_MBNL_score:0.000:3.714
MAN2A2/PRC1	Attention deficit hyperactivity disorder	rs2677744	rs12911192	0.192	0.265	0.361	0.031	0.099	0.375	0.313	0.306	ESS_Nova_score:13.210:2.133
PON1	Response to TNF antagonist treatment	rs854555	rs13306699	0.374	0.218	0.170	0.353	0.275	0.179	0.117	0.275	ESS_PTB_score:32.282:0.000, ESS_hnRNP_A1_score:-153.442:-230.578
PARK16/SLC2 6A1	Parkinsons disease	rs947211	rs16856462	-0.009	0.295	0.286	0.348	0.181	0.282	-0.016	0.222	ESE_3C_score:1.234:0.000, ESE_3H_score:2.251:1.045
PKP1/TNNT2	Panic disorder	rs860554	rs3729547	-0.077	-0.027	-0.165	-0.004	-0.022	-0.063	-0.346	-0.056	ESS_hnRNP_A1_score:0.000:-167.944, ESS_G_score:0.000:-2.010
CSK/COX5A	Diastolic blood pressure	rs6495122	rs4131311	0.282	0.001	0.012	0.342	0.095	0.038	0.073	0.188	5Pmaxent:8.680:1.930
FARP2	Chronic lymphocytic leukemia	rs757978	rs13013693	0.106	0.340	0.132	0.248	0.262	0.260	0.000	0.241	3pmaxent:6.900:5.080

Table 8: Predicted effects of GWAS SNPs on splicing

The top 25 (of approximately 500) predictions of GWAS SNPs to effect splicing. These SNPs showed the highest changes in probability that the splicing event would be in a low EIL group due to the SNP. Shown are the values that are calculated when the probability that the exon with the reference allele is in the low EIL category is subtracted from the probability that the exon with the SNP allele is in the low EIL category in each tissue. Positive values represent an increase in the probability that the exon would be spliced out, while negative values represent a decrease in the probability that the exon would be spliced out. The first gene name in the left column corresponds to the gene in which the GWAS SNP is located, the second gene name corresponds to the gene in which the LD SNP is located. If only one gene name is provided, then both SNPs are in that gene.

Gene	Disease	GWAS SNP	LD SNP	Adipose	Liver	Testes	Sk Muscle	Lymph	Brain	Colon	Heart	
			rs3825601, rs3825604,									
FRMD6/GNG2	Hippocampal atrophy	rs11626056	rs3825600, rs3825602	0.000	0.000	0.050	0.000	0.000	0.557	0.116	0.552	ESS_MBNL_score:0.584:-0.110, ESS_PTB_score:-552.830:-960.464 ESS_Nova_score:-3.175:0.332,ESS_TIA_score:-296.356:-227.425
			rs3825602	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_MBNL_score:0.584:0.251, ESS_PTB_score:-552.830:-537.711
			rs3825600	0.041	0.000	0.050	0.000	0.120	0.557	0.116	0.552	ESS_PTB_score:-552.830:-990.482, ESS_Nova_score:-3.175:0.332, ESS_TIA_score:-296.356:-264.234
			rs3825601	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_MBNL_score:0.584:0.118, ESS_TIA_score:-296.356:-241.905
			rs3825604	0.018	0.000	0.011	0.000	0.011	0.171	0.000	0.000	3pmaxent:9.680:7.420
	Hematological and biochemical		rs9263692, rs3095302,									
CDSN/PSORS1C1	traits	rs3094212	rs3094198	-0.498	-0.040	-0.018	-0.005	-0.139	0.000	-0.141	-0.006	ESE_5A3G_score:4.936:3.711, ISE_PTB_d_score:0.020:0.266 5Pmaxent:6.840:7.530
			rs9263692	-0.126	-0.040	0.000	-0.005	0.000	0.000	-0.141	-0.006	5Pmaxent:6.840:7.530
			rs3094198	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ISE_PTB_d_score:0.020:0.266
			rs3095302	-0.297	0.000	-0.018	0.000	-0.139	0.000	0.000	0.000	ESE_5A3G_score:4.936:3.711
			rs11792087,									
RP11/TLE1	Hippocampal atrophy	rs10867752	rs2796465, rs911638	0.074	0.000	0.000	0.000	0.000	0.423	0.000	0.000	ESE_5B3A_score:12.924:12.873, ESS_PTB_score:-4121.967:-4092.806 ESS_FOX_score:-0.043:-0.058,ESS_A_score:1.057:3.791
			rs2796465	0.074	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_PTB_score:-4121.967:-4104.439, ESS_A_score:1.057:3.791
			rs911638	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_PTB_score:-4121.967:-4110.730
			rs11792087	0.000	0.000	0.000	0.000	0.000	0.423	0.000	0.000	ESE_5B3A_score:12.924:12.873, ESS_FOX_score:-0.043:-0.058
			rs592595,									
PTBP2	Schizophrenia	rs7544736	rs190645	-0.140	-0.033	-0.011	0.038	-0.114	-0.071	-0.061	0.057	ESE_3H_score:0.000:0.746, ESS_CELF_score:0.000:-59.464 ESS_PTB_score:-155.028:0.000
			rs190645	-0.135	-0.031	-0.012	-0.157	-0.105	-0.104	-0.064	-0.120	ESE_3H_score:0.000:0.746
			rs592595	-0.015	-0.005	0.004	0.227	-0.019	0.057	0.013	0.282	ESS_CELF_score:0.000:-59.464, ESS_PTB_score:-155.028:0.000
	Ankylosing		rs3734016,									ESE_SRp40_score:10389.798:12117.143,
ERAP1	spondylitis	rs27434	rs26653	0.008	-0.001	-0.002	-0.130	0.000	0.000	-0.017	-0.002	ESE_SRp55_score:5.544:24.896 ESE_3B_score:0.912:0.506,ESS_A_score:1.180:1.721
			rs26653	-0.118	-0.001	-0.002	-0.130	0.000	0.000	-0.017	-0.002	ESE_SRp40_score:10389.798:12117.143, ESE_SRp55_score:5.544:24.896, ESE_3B_score:0.912:0.506
			rs3734016	0.302	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ESS_A_score:1.180:1.721
			rs1674777,									
FCGR2A	Ulcerative colitis	rs10800309	rs2045571	0.003	0.000	0.000	0.000	0.000	0.000	0.010	0.000	ESE_SRp40_score:127.057:79.526, ESE_5C3D_score:0.024:0.016 ESS_SF1_score:-597.454:-17.513
			rs1674777	0.003	0.000	0.000	0.000	0.000	-0.001	0.009	0.000	ESS_SF1_score:-597.454:-17.513
			rs2045571	0.002	0.000	0.000	0.000	0.000	0.000	0.007	0.000	ESE_SRp40_score:127.057:79.526, ESE_5C3D_score:0.024:0.016
			rs8187724, rs8187725,									ESE_SRp40_score:7093.098:10272.470,
SLC22A2	Serum creatinine	rs3127573	rs2292334	0.000	0.000	0.000	-0.002	0.000	0.000	0.000	-0.001	ESE_SRp55_score:38.296:0.000 ESS_Nova_score:1.737:-11.657,ESS_SF1_score:-611.770:-424.046 ISE_TIA_u_score:48.798:46.465
			rs2292334	0.000	0.000	0.002	0.009	0.001	0.000	0.000	0.006	ESE_SRp40_score:7093.098:10272.470, ESE_SRp55_score:38.296:0.000
			rs8187725	0.000	0.000	0.000	-0.002	0.000	0.000	0.000	-0.002	ESS_Nova_score:1.737:-11.657, ESS_SF1_score:-611.770:-424.046
			rs8187724	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	ISE_TIA_u_score:48.798:46.465

Table 9: Predictions of effects of SNP haplotypes on splicing

Haplotype effects are shown along with effects of SNPs individually below the haplotype predictions. Shown are the values that are calculated when the probability that the exon with the reference allele is in the low EIL category is subtracted from the probability that the exon with the SNP allele is in the low EIL category in each tissue. Positive values represent an increase in the probability that the exon would be spliced out, while negative values represent an decrease in the probability that the exon would be spliced out. Gene names as in Table 8

References

(2005) A haplotype map of the human genome. Nature 437: 1299-1320.

Anand VS, Braithwaite SP (2009) LRRK2 in Parkinson's disease: biochemical functions. FEBS J 276: 6428-6435.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 6:412-24.

Bellehumeur C, Blanchet J, Fontaine JY, Bourcier N, Akoum A (2009) Interleukin 1 regulates its own receptors in human endometrial cells via distinct mechanisms. Hum Reprod 24: 2193-2204.

Bhaskar H, Hoyle DC, Singh S (2006) Machine learning in bioinformatics: a brief survey and recommendations for practitioners. Comput Biol Med 36: 1104-1125.

Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, et al. (2009) The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res 19: 1665-1674.

Black D (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291-336.

Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126: 37-47.

Bouckaert, R "Bayesian Network Classifiers in Weka", Technical Report, Department

of Computer Science, Waikato University, Hamilton, NZ 2008.

Buscarlet M, Hermann R, Lo R, Tang Y, Joachim K, et al. (2009) Cofactor-activated phosphorylation is required for inhibition of cortical neuron differentiation by Groucho/TLE1. PLoS One 4: e8107.

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, et al. (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659: 147-157.

Cooper TA, Wan L, Dreyfuss G (2009) RNA and disease. Cell 136: 777-793.

Cote J, Dupuis S, Wu JY (2001) Polypyrimidine track-binding protein binding downstream of caspase-2 alternative exon 9 represses its inclusion. J Biol Chem 276: 8535-8543.

de Souto Marcilio, CP, Bittencourt, VG, Jose, A (2006) An empirical analysis of under-sampling techniques to balance a protein structural class dataset. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS 4234: 21–29.

Fairbrother W, Yeh R, Sharp P, Burge C (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007-1013.

Fogarty MP, Xiao R, Prokunina-Olsson L, Scott LJ, Mohlke KL (2010) Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK. Hum Mol Genet 19: 1921-1929.

Fox S, Filichkin S, Mockler TC (2009) Applications of ultra-high-throughput

sequencing. Methods Mol Biol 553: 79-108.

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479-2481.

Hao K, Chudin E, Greenawalt D, Schadt EE (2010) Magnitude of stratification in human populations and impacts on genome wide association studies. PLoS One 5: e8695.

Han K, Yeo G, An P, Burge CB, Grabowski PJ (2005) A combinatorial code for splicing silencing: UAGG and GGGG motifs. PLoS Biol 3:e158.

Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360: 1759-1768.

Hertel KJ (2008) Combinatorial control of exon recognition. J Biol Chem 283: 1211-1215.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87: 4438-4441.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.

Holleboom AG, Vergeer M, Hovingh GK, Kastelein JJ, Kuivenhoven JA (2008) The value of HDL genetics. Curr Opin Lipidol 19: 385-394.

Irimia M, Rukov JL, Roy SW, Vinther J, Garcia-Fernandez J (2009) Quantitative regulation of alternative splicing in evolution and development. Bioessays 31: 40-50.

J. Platt: Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

Kan Z, Rouchka EC, Gish WR, States DJ (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res 11: 889-900.

Kawase T, Akatsuka Y, Torikai H, Morishima S, Oka A, Tsujimura A, Miyazaki M, Tsujimura K, Miyamura K, Ogawa S, Inoko H, Morishima Y, Kodera Y, Kuzushima K, Takahashi T (2007) Alternative splicing due to an intronic SNP in HMSD generates a novel minor histocompatibility antigen. Blood :1055-63.

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11: 345-355.

Krawczak M, Cooper DN (1997) The human gene mutation database. Trends Genet 13: 121-122.

Kubo T, Kim SR, Sai K, Saito Y, Nakajima T, Matsumoto K, Saito H, Shirao K, Yamamoto N, Minami H, Ohtsu A, Yoshida T, Saijo N, Ohno Y, Ozawa S,

Sawada J (2005) Functional characterization of three naturally occurring single nucleotide polymorphisms in the CES2 gene encoding carboxylesterase 2 (HCE-2). Drug Metab Dispos 33:1482-7.

Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3: reviews.

Lander AD (2010) The edges of understanding. BMC Biol 8: 40.

Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 38:404-15.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.

Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010;363:166-76.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517.

Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6: 386-398.

McClellan J, King MC (2010) Genetic heterogeneity in human disease. Cell 141: 210-217.

Mir KU (2009) Sequencing genomes: from individuals to populations. Brief Funct Genomic Proteomic 8: 367-378.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628.

Nitesh V. Chawla et. al. (2002). Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16:321-357.

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, et al. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16: 929-941.

Pan SS, Han Y, Farabaugh P, Xia H (2002) Implication of alternative splicing for expression of a variant NAD(P)H:quinone oxidoreductase-1 with a single nucleotide polymorphism at 465C>T. Pharmacogenetics 12:479-88.

Pascual M, Vicente M, Monferrer L, Artero R (2006) The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing. Differentiation 74: 65-80.

Quinlan, J. R. (1993) C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Rahman L, Bliskovski V, Reinhold W, Zajac-Kaye M (2002) Alternative splicing of brain- specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. Genomics 80: 245-249.

Rhead B, Karolchik D, Kuhn R, Hinrichs A, Zweig A, et al. (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38: D613-619.

Ritchie DB, Schellenberg MJ, MacMillan AM (2009) Spliceosome structure: piece by piece. Biochim Biophys Acta\ 1789\: 624-633\.

Rubin GM (2001) The draft sequences. Comparing species. Nature 409: 820-821.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

Sivagnanam M, Mueller JL, Lee H, Chen Z, Nelson SF, Turner D, Zlotkin SH, Pencharz PB, Ngan BY, Libiger O, Schork NJ, Lavine JE, Taylor S, Newbury RO, Kolodner RD, Hoffman HM (2008) Identification of EpCAM as the gene for congenital tufting enteropathy. Gastroenterology 135:429-37.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, et al. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15: 2490-2508.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956-960.

Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol 3: e116.

Usama M. Fayyad, Keki B. Irani: Multi-interval discretization of continuous valued

attributes for classification learning. In: Thirteenth International Joint Conference on Articial Intelligence, 1022-1027, 1993.

Vali U, Brandstr^m M, Johansson M, Ellegren H (2008) Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet 9: 8.

Wachtel C, Manley JL (2009) Splicing of mRNA precursors: the role of RNAs and proteins in catalysis. Mol Biosyst 5: 311-316.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476.

Wang G, Cooper T (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8: 749-761.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831-845.

Xiao X, Wang Z, Jang M, Burge CB (2007) Coevolutionary networks of splicing cis-regulatory elements. Proc Natl Acad Sci U S A 104: 18583-18588.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, et al. (2009) Splice site strength-dependent activity and genetic buffering by poly-G runs. Nat Struct Mol Biol 16: 1094-1100.

Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11: 377-394.

Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356: 1263-1274.