Genetic variation of pre-mRNA alternative splicing in human populations

Zhi-Xiang Lu,1 Peng Jiang1 and Yi Xing1,2∗

The precise splicing outcome of a transcribed gene is controlled by complex interactions between cis regulatory splicing signals and trans-acting regulators. In higher eukaryotes, alternative splicing is a prevalent mechanism for generating transcriptome and proteome diversity. Alternative splicing can modulate gene function, affect organismal phenotype and cause disease. Common genetic variation that affects splicing regulation can lead to differences in alternative splicing between human individuals and consequently impact expression level or protein function. In several well-documented examples, such natural variation of alternative splicing has indeed been shown to influence disease susceptibility and drug response. With new microarray and sequencing-based genomic technologies that can analyze eukaryotic transcriptomes at the exon or nucleotide level, it has become possible to globally compare the alternative splicing profiles across human individuals in any tissue or cell type of interest. Recent large-scale transcriptome studies using high-density splicing-sensitive microarray and deep RNA sequencing (RNA-Seq) have revealed widespread genetic variation of alternative splicing in humans. In the future, an extensive catalog of alternative splicing variation in human populations will help elucidate the molecular underpinnings of complex traits and human diseases, and shed light on the mechanisms of splicing regulation in human cells. © 2011 John Wiley & Sons, Ltd.

INTRODUCTION

Splicing of precursor mRNA (pre-mRNA) is an essential step of eukaryotic gene expression. During splicing, introns are removed from the pre-mRNA and exons are ligated to produce the mature mRNA product. This process is tightly regulated by cis elements within exons and surrounding introns as well as trans-acting factors that bind to these cis elements.1 In alternative splicing (AS), this basic process is variable, so that multiple mRNA and protein isoforms can arise from a single gene.2 Alternative splicing of a single gene can generate a variety of mRNA and protein products, each with distinct properties in stability, subcellular localization and function.3 Alternative splicing can be modulated by variation both in the cis genomic splicing signals and in the cellular pathways that regulate splicing.2 In humans, the frequency of alternative splicing has been the subject of scrutiny. During the past 15 years, increasingly powerful technologies have been developed that can detect alternatively spliced transcripts at the global level. As the technology advanced so did estimates of the frequency of alternative splicing in humans.2,4–6 The most recent estimate by high-throughput RNA sequencing (RNA-Seq) is that more than 90% of multi-exon genes in the human genome are alternatively spliced,5,6 revealing the extent to which alternative splicing expands the regulatory and functional complexity of higher eukaryotes.

Alternative splicing is not only important for normal cellular functions but also frequently is
involved in disease pathogenesis. Disrupting the normal splicing pattern can cause disease; and sometimes even a modest shift in the relative proportions of mRNA isoforms from a single gene can be pathogenic. The majority of disease-causing splicing mutations affect critical splicing regulatory signals in cis (e.g., mutating consensus splice site (SS) sequences at exon–intron boundaries or splicing enhancer/silencer elements within exons or introns). Nevertheless, disease-causing splicing mutations can also act in trans, by disrupting the expression or RNA-binding activity of splicing regulators. While cis mutations affect splicing of a single transcript, trans mutations can compromise regulated splicing of many downstream gene targets simultaneously. As such, trans splicing mutations are known to cause a broad spectrum of diseases, such as neurodegenerative disease, muscular dystrophy, heart failure, and cancer. The classic example is myotonic dystrophy type 1 (DM1), in which the trinucleotide repeat CUG is expanded in the 3′ untranslated region (UTR) of the gene DMPK. This expanded trinucleotide repeat sequesters proteins in the muscleblind family of splicing regulators and affects many muscleblind-dependent splicing events.

A wealth of information has demonstrated that pathogenic mutations can disrupt splicing in patient tissues; however, less well explored is how complex traits or disease susceptibility can be influenced by normal genetic variation of alternative splicing. In recent years, thanks to new technologies for exon or nucleotide level profiling of eukaryotic transcriptomes, growing evidence reveals widespread natural variation of alternative splicing in humans. A number of studies have sought to identify splicing variation among people that correlates with single nucleotide polymorphisms (SNPs) in surrounding genomic regions. In this review, we describe molecular mechanisms that underlie genetic variation of alternative splicing. We illustrate genomic tools for global analysis of alternatively spliced transcripts, and summarize recent studies using these tools to survey alternative splicing differences among human individuals. We discuss the functional impact and disease association of alternative splicing variation. Finally, we discuss how a comprehensive catalog of splicing variation in human populations will provide important insights into the mechanisms of alternative splicing regulation in human cells.

**MECHANISMS OF ALTERNATIVE SPLICING REGULATION**

There are five basic types of alternative splicing (Figure 1): exon skipping, alternative 5′ SS (donor), alternative 3′ SS (acceptor), mutually exclusive exon usage, and intron retention. In addition, alternative initiation and alternative polyadenylation provide two other common mechanisms for generating various transcript isoforms. Different types of alternative splicing and transcript isoform variation can occur in a combinatorial manner, sometimes yielding an enormous number of distinct mRNA and protein isoforms from a single gene. A famous example is the *Drosophila* gene *Dscam*, which produces 38,016 distinct isoforms through alternative splicing of several exon clusters, each containing a large number of exons with mutually exclusive usage.

Alternative exon selection and splice site choice are controlled by an intricate regulatory network involving cis splicing elements and trans splicing regulators (Figure 2(a)). The most essential core splicing elements include the 5′ SS (donor site) and the 3′ SS (acceptor site), which define the exon–intron boundary; also important are the branch site and polypyrimidine tract, which lie upstream of the 3′ SS. These core cis splicing signals are recognized by the spliceosome, a large complex of protein and RNA subunits that are ubiquitously expressed and assemble on the pre-mRNA during splicing. In addition to core cis splicing signals, other auxiliary cis elements in exons and flanking introns can promote or inhibit exon splicing. The locations of these

**FIGURE 1** | Basic types of alternative splicing and transcript isoform variation.
elements and their effects on exon splicing categorize these elements as exonic splicing enhancer (ESE), intronic splicing enhancer (ISE), exonic splicing silencer (ESS), or intronic splicing silencer (ISS). Such auxiliary elements are recognized by trans-acting splicing regulators that positively or negatively regulate exon inclusion. Positive trans regulators are depicted in blue and negative trans regulators are depicted in brown. (b) A cis polymorphism disrupts an ISE element and abolishes the interaction between the cis element and the trans regulator, resulting in the switch from the exon inclusion isoform to the exon skipping isoform. (c) A trans polymorphism alters the RNA-binding activity of the trans splicing regulator and abolishes the interaction between the cis element and the trans regulator, resulting in the switch from the exon inclusion isoform to the exon skipping isoform.

Since the splicing of pre-mRNA is determined by interactions between cis elements and trans regulators, splicing outcomes can differ among human individuals due to genetic variation that alters such interactions in cis or trans. In a hypothetical example (illustrated in Figure 2(b) and (c)), an exon is only included when it contains an ISE element that is recognized by a splicing regulatory protein. Here, if the enhancer element is disrupted by a cis polymorphism, then it is not recognized by the splicing regulatory protein: without the protein bound to the enhancer, the exon is skipped (Figure 2(b)). A similar outcome could arise from a trans polymorphism in which the splicing regulatory protein is mutated in a way that disrupts RNA binding or expression (Figure 2(c)). Both scenarios could cause genetic variation to affect alternative splicing in human populations. It should be noted, however, that while the effect of a cis polymorphism (Figure 2(b)) is expected to be local (restricted to a specific adjacent exon), the effect of a trans polymorphism (Figure 2(c)) might be global, because each splicing regulator may control hundreds to thousands of alternative splicing events in the transcriptome.

GENOMIC TOOLS FOR ANALYSIS OF ALTERNATIVE SPlicing

Alternative splicing plays critical roles in development, tissue differentiation, and disease, and so has been a longstanding subject of study. A variety of molecular and genomic tools have been developed to analyze alternative splicing. One widely used molecular approach is reverse transcription polymerase chain reaction (RT-PCR). With this method, any alternatively spliced exon(s) of interest can be monitored using a pair of forward and reverse PCR primers that hybridize to flanking exons (Figure 3(a)). After the RT-PCR reaction, electrophoresis can separate the protein bound to the enhancer, the exon is skipped isoform.

| FIGURE 2 | Mechanisms of alternative splicing regulation and variation in humans. (a) Alternative splicing is controlled by an intricate regulatory network involving cis splicing elements and trans splicing regulators. The essential core splicing signals include the 5′ splice site (SS), 3′ SS, branch site (A), and polypyrimidine tract (Y(n)). Other auxiliary cis elements in exons and flanking introns include the exonic splicing enhancer (ESE), intronic splicing enhancer (ISE), exonic splicing silencer (ESS), and intronic splicing silencer (ISS). These auxiliary elements are recognized by trans-acting splicing regulators that positively or negatively regulate exon inclusion. Positive trans regulators are depicted in blue and negative trans regulators are depicted in brown. (b) A cis polymorphism disrupts an ISE element and abolishes the interaction between the cis element and the trans regulator, resulting in the switch from the exon inclusion isoform to the exon skipping isoform. (c) A trans polymorphism alters the RNA-binding activity of the trans splicing regulator and abolishes the interaction between the cis element and the trans regulator, resulting in the switch from the exon inclusion isoform to the exon skipping isoform. |
Alternative splicing was first studied at the genomic level when full-length complementary DNAs (cDNAs) and expressed sequence tags (ESTs) were sequenced on a large scale. In this first wave of genome-level investigation, exon–intron structures and alternative splicing events could be identified by aligning cDNA/EST sequences to the reference genome. Despite the utility of this approach, cDNA and EST sequencing have a limited throughput. Currently, there are less than seven million cDNA/EST sequences for human genes in the NCBI UniGene database. It is important to note that these sequences are aggregated over a wide range of tissues, developmental states, and diseases. For any given alternative splicing event, the number of EST sequences supporting distinct transcript isoforms in a particular tissue or cell type is usually extremely limited. Thus, cDNA/EST sequencing can depict only a crude picture of alternative splicing in the human transcriptome, but cannot quantify isoform abundance in individual samples. Thus, new genomic technologies are needed for analyzing alternative splicing in any RNA sample of interest in a global and quantitative manner. Over the past few years, two powerful genomic approaches were developed for this purpose: high-density splicing-sensitive microarray and ultra-deep RNA-Seq.

High-Density Splicing-Sensitive Microarray
Splicing-sensitive microarrays target specific exons or exon–exon junctions with oligonucleotide probes (Figure 3(b)). In the results of a microarray experiment, the fluorescent intensities of individual probes reflect the usage of alternatively spliced exons. This microarray-based approach was pioneered by Clark et al. for a genome-wide analysis of mRNA splicing in Saccharomyces cerevisiae. Other research groups and companies subsequently developed microarray platforms for analyzing alternative splicing in a variety of organisms including human, mouse and Drosophila. Some of these designs only contain probes targeting exon sequences, while others include probes for both exons and exon–exon junctions.

Ultra-Deep RNA-Seq
‘RNA-Seq’ has emerged as a powerful technology for transcriptome analysis. Currently, a single lane of RNA-Seq run on the Illumina HiSeq 2000 sequencer can produce up to 350 million sequence reads. When tens to hundreds of millions of RNA-Seq reads are mapped to the genome, exons, and exon–exon junctions, one can annotate exon–intron structures and estimate the relative isoform abundance of individual alternative splicing events (Figure 3(c)).

Currently, the strengths of high-density splicing-sensitive microarrays and ultra-deep RNA-Seq in
alternative splicing analysis complement each other. High-density splicing-sensitive microarrays are a cost-effective way to assay exons and alternative splicing events that are known. Coupled with appropriate analysis algorithms, microarray platforms can detect changes in alternative splicing patterns at a low false positive rate, as demonstrated in numerous studies. Unlike microarray analysis, RNA-Seq does not require prior knowledge of gene structures and splicing events and has a nucleotide level resolution. Once RNA-Seq reaches sufficient coverage and splicing events and has a nucleotide level resolution, it can accurately estimate the exon’s inclusion levels in mature mRNA transcripts. Nevertheless, to obtain quantitative exon-level measurements over the entire transcriptome, especially for less abundant transcripts, RNA-Seq must go very deep, which requires substantial experimental cost. We anticipate that, in the near term, microarray and sequencing-based approaches for alternative splicing analysis will coexist. However, because the cost of high-throughput RNA-Seq has been dropping precipitously, we expect that RNA-Seq will eventually supplant microarrays as the standard tool for transcriptome analysis. Indeed, since the advent of RNA-Seq a number of studies have adopted this technology to characterize multiple layers of regulatory variation in the human transcriptome (reviewed in Ref 46).

It should be noted that genetic polymorphisms could potentially confound the splicing analysis of microarray and sequencing-based transcriptome profiles. The oligonucleotide probes on splicing-sensitive microarrays are designed using the reference genome sequence. In microarray hybridization, the fluorescent intensity of any given probe depends on the abundance of its target mRNA transcript as well as the binding affinity of the probe to its target. In the absence of any effect on expression and splicing, a genetic polymorphism can decrease probe intensity simply by reducing the binding affinity of a microarray probe to its target, creating a spurious signal for altered expression or splicing. An analysis of Affymetrix exon array data of 57 individuals showed that this confounding issue is a serious source of false positives for detecting genetic variation of alternative splicing. A practical solution is to remove all microarray probes overlapping with polymorphic sites from analysis. This problem is less severe for RNA-Seq, because RNA-Seq mapping protocols can tolerate a small number of mismatches. However, genetic polymorphisms can still complicate the mapping of RNA-Seq reads, especially larger-scale polymorphisms such as insertions or deletions. In the future it will be useful to develop computational procedures that can reliably map RNA-Seq reads to polymorphic sites in the genome and transcriptome.

GENOME-WIDE SURVEY OF ALTERNATIVE SPlicing VARIATION IN HUMAN POPULATIONS

The key to identifying genetically controlled variation in alternative splicing is to associate alternative splicing differences in people with particular genetic polymorphisms. Differences in alternative splicing between individuals/alleles could be manifested either as all-or-none switches between competing mRNA isoforms or as shifts in the relative proportions of multiple mRNA isoforms of a single gene.

The first large-scale survey of SNP-associated alternative splicing was conducted by Nembaware et al. through the combined analysis of the dbSNP and dbEST databases. Using ESTs that map to both SNPs and isoform-specific exon–exon junctions of alternatively spliced transcripts, the authors estimated that 21% of the alternatively spliced isoforms detected from EST data arose from allele-specific splicing, with a conservative lower bound estimate of 6% if one considered only SNPs that caused all-or-none switches between isoforms. This study provides the first genomic evidence that an appreciable percentage of alternative splicing events reported in human genes might be attributed to genetic polymorphisms. In another study, using RT-PCR Hull et al. investigated the splicing patterns of 250 exons in 22 individuals of European ancestry. Of the 250 exons, 6 showed considerable splicing differences among individuals. Moreover, for each of the six exons, the splicing differences observed among individuals correlated well with the genotypes of a particular cis-SNP in the neighboring genomic region, suggesting alternative splicing was being genetically controlled.

Recently, genome-wide patterns of alternative splicing variation in human populations have been investigated by several groups using the HapMap lymphoblastoid cell lines (LCLs) as the model system. Across the entire genome, the SNPs in these HapMap LCLs were characterized extensively by the international HapMap project, making it convenient to test associations of splicing patterns with genetic polymorphisms. By combining Affymetrix exon 1.0 array analysis of two unrelated European individuals and RT-PCR analysis of splicing patterns within a three-generation family, Kwan et al. demonstrated the Mendelian inheritance of alternative splicing patterns of three genes (OAS1, CAST, and CRTAP).

Subsequently, an exon 1.0 array study by Kwan et al. extended the analysis to HapMap LCLs
from 57 unrelated individuals of European ancestry. They identified 177 genes whose relative transcript isoform proportions (owing to alternative splicing, alternative initiation, and alternative polyadenylation) correlated strongly with surrounding SNPs. The same exon 1.0 array approach was also used to characterize splicing variation among 176 HapMap LCLs of European and African ancestry.

Such genetic studies of alternative splicing are facilitated by the advent of the high-throughput RNA-Seq technology. For example, Pickrell et al. generated 1.2 billion single-end RNA-Seq reads on LCLs of 69 Nigerian individuals from the HapMap project. By treating the isoform proportions of alternatively spliced genes as quantitative traits, they identified 187 genes with splicing patterns that correlated with neighboring SNPs, revealing putative splicing quantitative trait loci (sQTLs). In a similar study, 60 HapMap LCLs of European ancestry were analyzed by Montgomery et al. through paired-end RNA-Seq for discovery of putative sQTLs. Both RNA-Seq studies come with a caveat, however, because the sequencing depth was a modest ~20 million reads per individual, likely generating a high rate of false positives and false negatives in splicing analysis. Clearly, a true set of sQTLs must be confirmed by experimental validation and further replication. In another ultra-deep RNA-Seq analysis of two HapMap European individuals, the authors observed a significant enrichment of detected sQTLs within high-confidence expression quantitative trait loci (eQTLs). This result suggests that alternative splicing variation among human individuals can regulate steady-state overall transcript levels, potentially through the effect on mRNA stability and degradation.

The genetic information on HapMap LCLs is readily available, making them the cell type of choice for most studies on transcriptome variation in human populations. Nevertheless, alternative splicing regulation can strongly depend on tissue and cell type; thus, a SNP that affects splicing in LCLs may not necessarily affect splicing in other tissues. Such tissue specificity of sQTLs was examined by Heinzen et al., who combined genome-wide SNP arrays with the Affymetrix exon 1.0 arrays to compare genetic control of alternative splicing in 93 cortical brain samples and 80 peripheral blood mononuclear cell (PBMC) samples. Indeed, although 80 high-confidence sQTLs were identified, only 49% were shared between the two tissue types. In a similar study, LCLs and osteoblasts were compared by Kwan et al., who estimated that 78% of sQTLs overlap in the two tissues. Both studies confirmed that genetic regulation of alternative splicing can be both tissue-specific and tissue-independent.

Although most studies to date have focused on cis-acting polymorphisms, multiple lines of evidence suggest that trans-acting polymorphic sites could broadly affect splicing regulation in human populations. A recent study on cis and trans-regulation of gene expression variation indicates that the majority of polymorphic sites with influence on gene expression act in trans to target genes. In a variety of human diseases, genetic mutations that disrupt trans-acting splicing regulators can trigger disease pathogenesis or modify disease severity. Thus, it is entirely conceivable that common genetic variation of splicing regulators can act as modifiers of complex traits or diseases, by altering the splicing of downstream gene targets. Importantly, variation in a brain-specific splicing regulator FOX1 has been recently implicated in autism. The discovery and functional characterization of trans-acting splicing polymorphisms is expected to accelerate in the future.

**FUNCTIONAL IMPACT AND DISEASE ASSOCIATION OF ALTERNATIVE SPLICING VARIATION IN HUMAN POPULATIONS**

High-throughput transcriptome studies are producing a fast-growing catalog of splicing variation in human populations, but so far information on the functional impact of such splicing variation is limited. Currently, our ability to detect alternative splicing events greatly exceeds our ability to characterize the functions of alternatively spliced gene products. Nonetheless, for several genes, genetic variation of alternative splicing has well-established functions (Table 1). For example, in HMSD, an intronic SNP (rs9945924) causes pronounced skipping of exon 2 by weakening the strength of the 5′ SS. This HMSD exon skipping isoform generates a novel minor histocompatibility antigen, which affects the immune response and could serve as a potential target for immunotherapy. In another example, an allele of ERAP2 (a gene involved in MHC Class I antigen presentation) harbors an A-to-G SNP (rs2248374) at the canonical 5′ SS of exon 10. This SNP activates a downstream cryptic SS, subjecting the resultant transcript to nonsense-mediated decay and significantly reducing the steady-state mRNA level of ERAP2. Consequently, primary lymphocytes homozygous for the G allele express less MHC Class I at the B cell surface, suggesting this alternative splicing event affects MHC antigen presentation. Interestingly, genetic analyses of this SNP in six human populations revealed a strong signature of balancing selection.
### Table 1: Examples of Alternative Splicing Variation in Human Populations and Their Functional or Pathological Consequences

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Effect of SNP on Splicing</th>
<th>Functional or Pathological Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMSD</td>
<td>An intronic SNP (rs9945924) causes pronounced skipping of exon 2 by weakening the 5′ splice site (SS)</td>
<td>Generates a novel minor histocompatibility antigen, which affects the immune response and could serve as a potential target for immunotherapy</td>
</tr>
<tr>
<td>ERAP2</td>
<td>A SNP (rs22483734) residing within the canonical 5′ SS of exon 10 triggers activation of a downstream cryptic SS, leading to mRNA nonsense-mediated decay and a significant reduction in the steady-state mRNA level of ERAP2</td>
<td>Primary lymphocytes homozygous for the G allele express less MHC Class I at the B cell surface, suggesting that this alternative splicing event affects MHC antigen presentation</td>
</tr>
<tr>
<td>OAS1</td>
<td>A SNP (rs10774671) at the 3′ SS of exon 7 abolishes SS activity, resulting in the usage of an internal 3′ SS</td>
<td>Produces a protein isoform with reduced enzymatic activity. Affects the response to interferon (IFN) therapy in hepatitis C patients, and influences susceptibility to multiple sclerosis</td>
</tr>
<tr>
<td>SCN1A</td>
<td>An intronic SNP (rs3812718) modulates the alternative splicing of exon 5</td>
<td>Influences the dose response to antiepileptic drugs</td>
</tr>
<tr>
<td>IRF5</td>
<td>An intronic SNP (rs2004640) generates a consensus GT 5′ SS of the alternative first exon 1B, producing an alternative transcript isoform initiated at exon 1B</td>
<td>Associated with increased risk of systemic lupus erythematosus</td>
</tr>
<tr>
<td>CD45 (PTPRC)</td>
<td>An exonic SNP (rs17612648) within exon 4 disrupts an exonic splicing silencer (ESS), leading to increased inclusion of exon 4</td>
<td>Associated with multiple sclerosis</td>
</tr>
<tr>
<td>LDLR</td>
<td>An exonic SNP (rs688) promotes skipping of its exon 12</td>
<td>Associated with total and LDL cholesterol levels in females especially in premenopausal women</td>
</tr>
</tbody>
</table>

suggesting this alternative splicing polymorphism confers an unknown adaptive benefit. A similar example was found in OAS1, a gene important for the innate immune response to virus infection. Here, a G-to-A SNP (rs10774671) at the 3′ SS of exon 7 abolishes SS activity, resulting in the usage of an internal 3′ SS and the production of a protein isoform with reduced enzymatic activity. This SNP also affects the response to interferon (IFN) therapy in hepatitis C virus (HCV) patients. HCV patients homozygous for the A allele do not respond to IFN and exhibit a higher degree of liver fibrosis. In SCN1A, a gene encoding a neuronal sodium-channel α subunit, an intronic SNP (rs3812718) was found that modulates the alternative splicing of exon 5 and influences the dose response to antiepileptic drugs.

An alternative to gene-specific functional experiments is to examine whether SNPs that cause splicing differences can be linked to phenotypic traits or diseases. Common genetic variants that affect the alternative splicing of IRF5, OAS1, CTLA4, and PTPRC (also known as CD45) were linked to several autoimmune diseases. Genetic polymorphisms that affect the alternative splicing of NPSR1 (also known as GPRA) were implicated in asthma. Another example was found in the low-density lipoprotein receptor (LDLR), in which a SNP (rs688) promotes skipping of exon 12 and is strongly associated with total and LDL cholesterol levels in females especially in premenopausal women. Similarly, in cortical brain and PBMC samples, the 80 identified associations between SNPs and splicing were systematically compared to 41 published genome-wide association studies (GWAS) for 50 different traits. Here, up to 13 SNP-splicing associations appeared to be responsible for previously reported GWAS signals of human traits (out of a total of 84 reported). Likewise, splicing QTL data of human osteoblasts were compared with a GWAS of bone mineral density, showing ~20% of GWAS signals may be attributed to alternative splicing variation. Such studies illustrate that intersecting sQTLs with GWAS signals provides an effective means to identify the causal regulatory effects of GWAS hits.

### Genetic Variation of Alternative Splicing and the Splicing Code

Despite the importance of alternative splicing, the rules that govern splicing regulation (i.e., the ‘splicing code’) in individual tissues and cell types remain poorly understood. Many cis splicing elements are degenerate, and their effects on splicing usually depend on the surrounding sequence context as well as the abundance and activity of trans-acting...
regulators.\textsuperscript{1} Although experimental and computational approaches have made considerable progress in elucidating the splicing code of mammalian cells,\textsuperscript{1,68,69} currently it is still challenging to predict the precise alternative splicing pattern of a gene from the primary genomic sequence.

Natural variation of alternative splicing provides a rich resource for dissecting the mechanisms of splicing regulation. When a given \textit{cis}-SNP can be linked to the splicing pattern of a particular exon across human populations, this suggests that a critical \textit{cis} splicing regulatory element is created or disrupted by that genetic variant. Intuitively, this is analogous to conventional minigene studies of splicing regulation, in which the role of a putative splicing regulatory element can be assessed by mutational analysis of the element in an artificial minigene splicing reporter construct.\textsuperscript{70,71} Here, the genetic variation of alternative splicing that exists in natural human populations can be viewed as the results of ‘whole-genome mutagenesis experiments’ during human evolution. By globally correlating alternative splicing variation in the human transcriptome to genetic polymorphisms in the human genome, we can gain significant insights into the genomic signals of splicing regulation and identify novel regulatory elements.

It must be emphasized, however, that specific SNPs found to be associated with alternative splicing are often not causal but rather are in linkage disequilibrium with the causal polymorphisms. Indeed, most of the significantly associated \textit{cis}-SNPs identified by sQTL analysis are far from the exons of interest\textsuperscript{13}; in contrast, \textit{cis} splicing elements are typically in close proximity to the exons they regulate.\textsuperscript{1,69} Therefore, after an sQTL is identified, the exon and flanking intronic regions must be resequenced to identify potential causal SNPs that might regulate splicing.\textsuperscript{12,18} The causal effect of SNPs identified by such resequencing can be confirmed by minigene experiments, in which the genomic sequences corresponding to distinct alleles are cloned into a splicing reporter construct and the splicing efficiencies are compared. This approach has been used in several studies to pinpoint the causative SNPs responsible for sQTL signals as well as the affected \textit{cis} splicing regulatory elements.\textsuperscript{12,18}

Genetic variation of alternative splicing can also reveal higher order interactions among distinct classes of \textit{cis} splicing regulatory elements. This was demonstrated by a recent study on \textit{5'} SS SNPs and their context-dependent splicing effects.\textsuperscript{22} In this work, Lu et al. analyzed the HapMap LCLs of 7 individuals using RT-PCR to examine the splicing patterns of 129 exons containing \textit{5'} SS SNPs. Surprisingly, despite the critical role of the \textit{5'} SS in exon recognition, only a small fraction of the tested \textit{5'} SS SNPs affected splicing patterns. By comparing exons affected by \textit{5'} SS SNPs to exons that were unaffected, Lu et al. discovered that the effects of \textit{5'} SS SNPs were buffered by adjacent sequence signals that promote exon recognition.

For example, the GGG motif was the most enriched trinucleotide sequence downstream of exons unaffected by \textit{5'} SS SNPs, consistent with previous reports on the GGG motif as an ISE that promotes the recognition of weak \textit{5'} SS.\textsuperscript{72} Using a similar approach, Fu et al. showed that the effect of single nucleotide mutations at the \textit{5'} most nucleotide of an exon can be buffered by a strong polypyrimidine tract upstream of the \textit{3'} SS.\textsuperscript{73} These results indicate that the effect of a given SNP on splicing can be buffered by surrounding sequence elements (see Figure 4(a) and (b) for a schematic illustration). Consequently, the splicing of exons lacking such compensatory mechanisms would be more susceptible to genomic mutations.

\section*{CONCLUSION}

Inheritable differences in splicing add an important layer to the molecular underpinnings of complex traits and human diseases. Although the current knowledge about genetic variation of alternative splicing in humans is strongly biased toward a single cell type (LCL), the rapid improvement in the capacity and cost structure of high-throughput sequencing technologies will enable genome-scale analyses of alternative splicing variation in many other tissues in the near future. For example, the recently launched NIH Genotype-Tissue Expression Program aims to analyze genotype–transcriptome correlations in 30–50 tissues collected from \textasciitilde 160 deceased donors (https://commonfund.nih.gov/GTEx/overview.aspx). A comprehensive catalog of expression and splicing QTLs across diverse tissues and cell types will facilitate a variety of important investigations. It will provide a powerful resource for human geneticists to infer the functional relevance of GWAS results, by mapping significant GWAS signals to eQTLs/sQTLs of disease-related tissues (e.g., GWAS signals of psychiatric diseases to eQTLs/sQTLs of brain tissues). Additionally, the tissue specificity of splicing QTLs will inform the mechanisms of splicing regulation and reveal the creation or loss of tissue-specific \textit{cis} splicing elements. For example, if a particular SNP decreases splicing of an adjacent exon in the brain but not in the muscle, then hypothetically, the SNP disrupts a splicing regulatory element that is brain specific (Figure 4(c) and (d)). Likewise, systematic analyses of \textit{trans} splicing QTLs across multiple tissues should identify common genetic variation in master regulators.
of tissue-specific splicing that globally contributes to splicing variation of downstream target exons.

Another important future direction is to develop new computational tools that reliably identify pathogenic mutations that influence splicing. Many exonic and intronic mutations can disrupt splicing.\textsuperscript{9} For example, a systematic analysis of synonymous mutations within exon 12 of \textit{CFTR}, the disease gene of cystic fibrosis, revealed that approximately one quarter of synonymous mutations would result in aberrant splicing and non-functional protein products.\textsuperscript{74} Such splicing mutations constitute a major class of human disease mutations. However, our current ability to predict whether a given genomic mutation would alter splicing is very limited. A number of computational tools are available for this purpose (reviewed by Ref \textsuperscript{75}); however, these tools generally are inaccurate or designed to analyze SNPs in specific types of splicing signals (e.g., splice sites). To accurately predict the splicing effect of genomic variants, we need a comprehensive understanding of \textit{cis} splicing regulatory elements and how these elements interact with each other as well as with \textit{trans} splicing regulators. Recently, several studies demonstrated that advanced machine learning techniques can computationally predict tissue-specific alternative splicing patterns by integrating hundreds of genomic and RNA sequence features.\textsuperscript{69,76} A major goal for the future is the development of computational approaches that accurately predict common and rare genetic variants that alter splicing. This will provide the tools human geneticists need to follow up on exome sequencing and whole-genome sequencing studies of human diseases, and facilitate using alternative splicing variants as disease markers for personalized medicine.

\section*{ACKNOWLEDGMENTS}

We thank Peter Stoilov and Keyan Zhao for helpful comments on this manuscript. This work was supported by National Institutes of Health grant R01GM088342 and a junior faculty grant from the Edward Mallinckrodt Jr Foundation.

\section*{REFERENCES}


Focus Article


60. Heinzen EL, Yoon W, Tate SK, Sen A, Wood NW, Sisodiya SM, Goldstein DB. Nova2 interacts with a cis-acting polymorphism to influence the proportions of


