The Quest for Splicing Polymorphisms under Positive Selective Pressure in the Human Genome

Dustin Griesemer
Biophysics 205
April 28, 2013

Introduction and Motivation

We are in the midst of a revolution in understanding and characterizing human genetic variation. With the ready availability of whole-genome sequencing data has come new insights into the mechanisms that drive human variation. For example, recent evidence has suggested that 88% of disease-associated polymorphisms lie in non-coding regions of the genome[1]. Of particular interest in the understanding of human genetic variation is the identification of variants that arise from adaptive forces in the environment (such as infection, changes in diet, or exposure to toxins), as these variants may explain the continued existence of, or variability in resistance to, disease. However, pinpointing the specific variant that is under selection is challenging, since many nearby alleles are likely to be co-inherited with the mutation of interest. In order to identify the causal variant driving selection, the Sabeti lab recently developed Composite of Multiple Signals (CMS), a test which combines multiple signals of selection to distinguish causal variants under recent positive selection with a resolution up to 100-fold higher than previously possible[2]. When CMS was applied to the entire human genome, the vast majority (90%) of selected variants were predicted to be regulatory. However, the functional implications of most of these selected regulatory variants remain unknown.

Alternative splicing is a post-transcriptional regulatory mechanism that allows for the creation of many distinct mRNA and protein products from our limited repertoire of approximately 25,000 genes, each with distinct properties in stability, subcellular localization, and function[3]. This capability is critical in multicellular organisms, in which distinct modes of regulation and subtle variations of function are often needed in order to allow diverse cell types to carry out their specialized biological roles. In testament, the proportion of genes which undergo alternative splicing rises in progression from lower- to higher-order organisms, reaching ~90% in humans[4]. Hence, in higher-level organisms such as humans, splicing is poised as a key substrate for positive evolutionary change. Indeed, a vast array of polymorphisms which affect the identity and relative proportions of splicing products have been noted in the human genome, many of which have been directly linked with disease[5]. Disappointingly, however, very few studies have attempted to identify those polymorphisms which have been subject to positive selection, and examples of splicing variants that have allowed humans to adapt to environmental or biological challenges are generally lacking.

I propose a comprehensive search of the human genome for polymorphisms which have been subject to positive selective pressure in the recent evolutionary past to alter splicing isoforms or their proportions. We will first ask if the splicing architecture within CMS-identified positively selected regions is similar to that of the rest of the genome. This may enlighten us to general splicing properties which make a gene more or less susceptible to positive selection. We will then begin our search by exploiting mechanistic insight into the evolution and regulation of splicing to predict polymorphisms which create novel isoforms or alter the proportions of existing isoforms. We will utilize high-throughput expression data coupled with deep genome sequencing to verify the splicing effects predicted *in silico*. Finally, we will validate high-

confidence splicing polymorphisms using a minigene assay and perform additional population genetic analyses as well as follow-up molecular and biochemical experiments to unveil the evolutionary history of these polymorphisms and their biological role in human adaptation.

Specific Aim 1: Compare the splicing architecture of positively-selected regions to the rest of the genome and identify candidate splicing polymorphisms responsible for positive selection signatures.

We will first analyze the splicing architecture of genes within CMS-identified positively-selected regions in order to identify properties of the splicing machinery that positive selection may potentially rely upon (such as the ability to selectively regulate a specific isoform or the ease of creating a new splice site). Of note, we are interested in all outcomes of positive selection at this stage and not just those that act directly through the splicing machinery. For example, if alternative exons are more likely than constitutive exons to undergo positive selection by alteration of the protein sequence, this is an interesting finding that provides useful insight into the role of splicing in positive selection in the human genome. Although many genes present in CMS regions will be neutral "hitchhikers" not actually subject to recent positive selection, the enrichment of positively selected genes within CMS regions relative to the rest of the genome may allow detection of differences in parameters of the splicing architecture at statistically significant levels (calculated via the Kolmogorov-Smirnov test for goodness of fit between the distribution of the parameter under investigation within CMS regions and its distribution within the whole genome). Parameters relevant to our study of splicing architecture will include intron-to-exon ratio (because longer introns and shorter exons have been observed over the course of higher organism evolution, and alternative exons are often flanked by longer introns than constitutive exons), number of fixed and polymorphic *Alu* transposable elements (because *Alu* elements are often involved in the creation of novel splicing isoforms), and the predicted strength of splice sites and Exonic Splicing Enhancers (ESEs) (because the weakening of splice sites and ESEs is associated with the transition from a constitutive exon to a regulated alternative exon)[6].

Turning to our search for polymorphisms which directly interact with the splicing machinery, we will begin by identifying within the approximately 100 polymorphisms of each of the 412 positively-selected regions identified by running CMS on the pilot phase of the 1000 Genomes Project those with the potential to create novel splicing isoforms. Insight into the evolution of novel splice isoforms has identified 3 predominant mechanisms[6]: exon shuffling, exonization, and transition. Each of these mechanisms involves characteristic intermediates which can be identified in the genome. Exon shuffling is the process in which a new exon is inserted into an existing gene or an exon is duplicated in the same gene. Variants which arise from exon shuffling manifest as Copy Number Variants (CNVs). A recent study which mapped CNVs using 1000 Genomes data identified 1,119 CNVs that overlap with exons[7]. We will determine if any of these CNVs result in polymorphic exon duplications relative to the ancestral state and are present

within CMS regions. Exonization is a process in which intronic sequences gain splice sites and become exons. *Alu* elements contain sequences similar to consensus splice sites and hence novel splice sites can be created when a single element inserted in the antisense orientation acquires characteristic mutations or when the hairpin formed by two complementary elements inserted adjacent to each other is modified by A-to-I RNA editing. A recent study identified 6,000 polymorphic *Alu* element insertions using 1000 Genomes data, in addition to the 1.1 million *Alu* elements which are fixed in the human lineage[8]. We will identify fixed *Alu* elements containing polymorphisms that create novel splice sites or polymorphic *Alu* elements which have acquired mutations that create novel splice sites. Finally, in transition, alternative cassette exons are derived from constitutive exons. This can occur when two complementary *Alu* elements adjacent to each other form a hairpin in the upstream intronic sequence which serves as an Intronic Splicing Silencer[9]. We will identify polymorphic *Alu* insertions resulting in the formation of a hairpin upstream of an exon. We will also identify polymorphisms which weaken the splice sites and ESEs of constitutive exons (as described below).

In addition to identifying the creation of novel isoforms, we will also identify polymorphisms which alter the proportions of existing isoforms by altering the affinity of trans-acting splicing factors to cis-acting (sequence) elements (CAEs). Currently, the best-characterized CAEs include the 5'donor, 3'acceptor and branchpoint consensus splice sites, as well as Exonic Splicing Enhancers (ESEs). We will combine results from multiple *in silico* CAE prediction tools, including MaxEntScan, Splice Site Finder, ESE Finder, and RESCUE-ESE, while imposing appropriate thresholds to predict polymorphisms which alter splice site or ESE strength with maximal sensitivity and specificity[10]. ESE prediction has particularly low specificity and we will take special care to verify predicted ESE polymorphisms with RNA-sequencing data.

Specific Aim 2: Identify splicing quantitative trait loci for the LCL cell line in order to verify the predicted effect of candidate splicing polymorphisms.

Once we have predicted candidate splicing polymorphisms *in silico*, we will look for evidence that the predicted novel splicing isoforms or changes in isoform proportions occur in a natural biological setting. We will thus investigate genotype-phenotype correlations in LCLs (EBV-immortalized B-lymphocytes derived from peripheral blood) for which data is already available through the 1000 Genomes Project. A caveat of this approach is that not all trans-acting splicing factors are expressed in LCLs, leading to false-negatives. However, previous studies that used a similar approach attained a suitable sensitivity[5]. Polymorphisms which result in differential expression of splicing isoforms, such as those we will predict in our *in silico* studies, are referred to as splicing quantitative trait loci (sQTLs). In order to identify sQTLs in the LCL cell line we will require high-density genome sequencing data (in order to genotype each polymorphism) as well as genome-wide expression data (to identify the splicing isoforms produced) across many individuals in multiple populations (to maximize heterozygosity across loci).

Genome data will be derived from participants in the pilot phase of the 1000 Genomes Project. This is a natural choice, as our CMS genome-wide scan was carried out on 1000 Genomes data. Illumina RNA-sequencing data for LCLs is available for 60 CEU (European) participants[11] as well as 69 YRI (African) participants[12]. Unfortunately, RNA-seq data is not currently available for CHB/JPT (Eastern Asian) participants. We will derive LCL expression data for CHB/JPT participants from a study in which LCL expression data was measured using the Illumina Human-6 v2 Expression BeadChip from 109 CEU, 108 YRI, and 162 CHB/JPT participants in phase 3 of the HapMap Project, many of whom also participated in the 1000 Genomes project[13]. We will derive genome sequencing data as well as microarray expression data from the subset of participants of the aforementioned study that also participated in the 1000 Genomes Project. A caveat of this approach is that the Illumina BeadChip only contains probes for well-characterized splicing isoforms. Nonetheless, including expression data for the CHB/JPT population will markedly improve our power to measure sQTLs affecting well-characterized splicing isoforms present at high minor allele frequencies (MAF) in this population (presumably due to positive selection). Another concern is that pooling data across platforms may produce false positive correlations due to platform bias. To minimize platform bias, we will perform linear regressions between RNA-seq counts and microarray isoform intensities for CEU and YRI participants for which both RNA-seq and microarray data is available. We will use the regression coefficients we derive to generate predicted RNA-seq counts for CHB/JPT participants. We will ensure that platform bias is minimized by performing 10-fold leave-one-out-cross-validation and assessing cross-platform correlation coefficients after linear regression. We will discard isoforms that repeatedly result in poor cross-platform correlation (for example, due to non-linear probe hybridization).

For individuals for which RNA-seq data is available, we will map reads which are not present in RefSeq to the genome in order to identify novel splicing isoforms. To identify sQTLs, the fraction of reads in a gene that falls in a given exon will be treated as a quantitative trait, and linear regression will be performed against number of derived alleles for each polymorphism within the exon under investigation and its flanking introns. To correct for multiple tests of association, we will carry out multiple permutations on expression data followed by FDR correction. The effects of population stratification will be controlled for by Principle Component Analysis.

Specific Aim 3: Validate high-priority splicing polymorphisms with minigene assays and investigate the potential selective role of validated candidates.

Finally, we will validate that our candidate splicing polymorphisms are causal for the changes in splicing observed. Because our means of validation is low-throughput, we will prioritize high-confidence polymorphisms for validation. *In silico* predictions will be sufficient evidence for polymorphisms which alter splice sites, as 83% specificity is attainable at appropriate

thresholds[10]. However, verification of the predicted splicing effect in RNA-seq data will be required for polymorphisms which create novel splice isoforms or alter isoform ratios by interfering with ESEs, as *in silico* prediction tools developed for these purposes have markedly lower specificity. Furthermore, we will limit validation to polymorphisms whose CMS score places them among the top 20 candidate causal variants in a given CMS region after repeating CMS with simulation parameters specific to the variant of interest (including population demographics, allele frequencies, and local recombination rate). Although each CMS region contains an average of 100 candidate variants, the causal variant is among the top 20 variants in 50% of random simulations and in a larger proportion of targeted simulations.

Bench-top validation will involve a minigene assay[14], in which the exon of interest (the "test" exon) is cloned along with its flanking introns between constitutively spliced GFP exons, such that GFP fluorescence is inversely correlated with inclusion of the test exon. Constructs will be generated which are identical except with respect to the candidate polymorphism in order to assess if the derived allele of the candidate polymorphism is sufficient to alter splicing of the test exon. The minigene assay will be performed in an LCL cell line in order to replicate the context in which the predicted effect was (potentially) verified.

Once a splicing polymorphism is validated, we will perform follow-up molecular and biochemical experiments to unveil its biological role in human adaptation. Oftentimes coding exons correspond to distinct protein domains[15]. Therefore, if a coding exon is affected we will analyze available protein structural data to determine if the splicing event affects the modular structure of the protein product. We will also determine if the splicing event affects subcellular localization. If a non-coding exon is affected, we will test for differential mRNA stability or translation rate between isoforms. Finally, if a particular trans-acting splicing factor is involved, we will investigate the tissue expression profile of the factor as well as its evolutionary origin (alternative splicing events often arise after the emergence of a novel splicing factor).

In order to investigate the evolutionary history of the polymorphism, we will perform spatially-explicit population genetic simulations to predict the Bayesian posterior probability distributions of various population genetic parameters, including date and location of origin as well as selection intensity. Our spatially explicit model will rely upon allele frequencies observed in the 52 global populations studied in the Human Genome Diversity Panel (HGDP), and will consider various demographic processes, including population growth, sporadic long-range migration, cultural diffusion of farming technology, gene flow between demes and between cultural groups, and the effects of the spread of farming on carrying capacities[16]. Coupled with biological information, this will allow us to gain insight into factors which shaped the recent evolution of humankind.

References

1. Hindorff, L. A. *et al.* 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 (2009).
2. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327,** 883–886 (2010).
3. Lu, Z.-X., Jiang, P. & Xing, Y. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdiscip Rev RNA* **3,** 581–592 (2012).
4. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
5. Coulombe-Huntington, J., Lam, K. C. L., Dias, C. & Majewski, J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* **5,** e1000766 (2009).
6. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11,** 345–355 (2010).
7. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470,** 59–65 (2011).
8. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7,** e1002236 (2011).
9. Lev-Maor, G. *et al.* Intronic Alus influence alternative splicing. *PLoS Genet.* **4,** e1000204 (2008).
10. Houdayer, C. In silico prediction of splice-affecting nucleotide variants. *Methods Mol. Biol.* **760,** 269–281 (2011).
11. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
12. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
13. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8,** e1002639 (2012).
14. Wang, Z. *et al.* Systematic identification and analysis of exonic splicing silencers. *Cell* **119,** 831–845 (2004).
15. Gelly, J.-C., Lin, H.-Y., de Brevern, A. G., Chuang, T.-J. & Chen, F.-C. Selective constraint on human pre-mRNA splicing by protein structural properties. *Genome Biol Evol* **4,** 966–975 (2012).
16. Kamberov, Y. G. *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152,** 691–702 (2013).